Karlsruher Institut für Technologie

Towards Reliability and Quality for ReRAM-based Content-Addressable Memory in Computation-in-Memory Paradigms

Zur Erlangung des akademischen Grades einer

Doktorin der Ingenieurwissenschaften

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

genehmigte

Dissertation

von

Haneen Seyam

aus Gizeh (Ägypten)

Tag der mündlichen Prüfung: 16. Juli 2026

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

Acknowledgments

Completing this dissertation has been one of the most formative journeys of my life, and I am profoundly grateful to everyone who offered guidance, friendship, and support along the way.

My deepest gratitude goes to my advisor, *Prof. Dr. Mehdi Tahoori*, for his trust, patience, and the intellectual freedom he afforded me. His constructive feedback and steady encouragement were invaluable, and his influence extends well beyond the pages of this work. I also sincerely thank *Prof. Lorena Anghel* for her support and for agreeing to serve on my doctoral committee.

I am deeply thankful to my colleagues at the Chair of Dependable Nano Computing (CDNC) at KIT for an environment that was both intellectually stimulating and genuinely supportive. A particularly warm thank you goes to *Dr. Mahta Mayahinia*, whose mentorship, generosity, and wisdom from my very first days at KIT meant more than I can say.

I am likewise grateful to *Soyed Tuhin Ahmed* for his kindness during my first year; *Dina Moussa*, whose quiet strength has always been inspiring; *Maha Shatta*, for a friendship that will last long beyond our time at KIT; *Mahboobe Sadeghipourrudsari*, for her relentless encouragement; *Shanmukha Mangadahalli Siddaramu*, for many memorable conversations over the years; *Christopher* and *Polykarpos*, who made my last days in KIT joyful; and *Vincent Meyers*, for his support and for proofreading our German documents. I further thank *Seyedehmaryam Ghasemi*, *Tara Gheshlaghi*, *Priyanjana Pal*, *Brojogopal Sapui*, and all other CDNC colleagues for their friendship and for making these years memorable. A special acknowledgement goes to *Iris Schröder-Piepka*, whose warmth and dedication made KIT feel like a home away from home.

To my friends in Egypt: *Badiaa*, my closest friend for over thirteen years, whose loyalty has been a constant source of strength; and *Hend* and *Yasmine*, for their friendship and for making every return to Egypt feel like a celebration. I also thank *Dr. Ayman Ismail*, who believed in my potential and encouraged me to pursue doctoral studies abroad, a trust that played a meaningful role in shaping the path that led here.

To my friends in Germany: *Dr. Mohamed Ali*, my mentor and guiding light, who first taught me in Egypt and was the first to welcome me upon my arrival in Germany. I am deeply grateful for his knowledge and for everything he taught me about life here. *Meshaiel*, my first friend in Germany and a sister in all but name, whose friendship has become one of the most enduring gifts of this journey. *Hadeer*, my roommate and constant source of motivation. *Nagwa*, who shared every stage of this journey with me, the laughter, the celebrations, and the difficult days alike. And *Nihal*, for her steadfast encouragement and belief that we could do it, despite every obstacle.

To *Hossam*, *Bilel*, and *Dr. Hassan*, for being like real brothers to me in Karlsruhe. To the wonderful people of *Casal Legends*, *Sarah*, *Yasmine*, *Radwa*, *Amira*, *Farida*, *Ziad*, *Monem*, *Yahya*, and *AbdelSalam*, for becoming my family away from home. And to *Abdelrahman*, *Lamis*, and little *Yahya*, for opening your home to me and for reminding me that family can be found far from where you were born.

None of this would have been possible without my family. To my parents, *Gamal* and *Mona*: everything I have achieved rests on the foundation of love and values you gave me. To my brothers, *Ahmed*, *Mohamed*, and *Tarek*, and to *Samar* and my niece *Dalida*, thank you for your love and for the warmth you bring every day. To my sister, *Leena*: you have always been my backbone, my greatest source of strength, and my safe place. I carry my gratitude for you in everything I do.

I would like to close with my deepest admiration for all those around the world who continue to stand for freedom, peace, dignity, and justice. This dissertation is dedicated to everyone who strives for a more just and peaceful world, and to all who hold onto hope even in the darkest of times.

Peace, freedom, and humanity for all.

Haneen Seyam
Karlsruhe

Hiermit erkläre ich an Eides statt, dass ich die von mir vorgelegte Arbeit selbstständig verfasst habe, dass ich die verwendeten Quellen, Internet-Quellen und Hilfsmittel vollständig angegeben haben und dass ich die Stellen der Arbeit – 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, 16. Juli 2026
Haneen Seyam

Abstract

Similarity search is a fundamental primitive in modern computing systems, enabling applications such as associative memory retrieval, Hyper-Dimensional Computing (HDC), Deoxyribonucleic Acid analysis (DNA) analysis, Bloom filters, and emerging Artificial Intelligence (AI) workloads. In conventional von Neumann systems, these operations incur significant data movement overhead, leading to severe latency and energy bottlenecks known as the *memory wall*. Computation-in-Memory (CiM) addresses this challenge by performing computation directly within memory arrays. Among Non-volatile Resistive Memories (NVM)-based CiM paradigms, emerging technologies such as Redox-based RAM (ReRAM), Phase Change Memory (PCM), Spin-Transfer Torque Magnetic RAM (STT-MRAM), and Ferroelectric Field-Effect Transistor (FeFET) offer significant performance and energy benefits through analog in-memory computation. Among these, ReRAM is particularly promising due to its large resistance window between High Resistive State (HRS) and Low Resistive State (LRS), high scalability, and Back-End-of-the-Line (BEoL) integration capability, making it well-suited for dense and energy-efficient CiM architectures.

In particular, ReRAM-based content-addressable memory (ReCAM) enables massively parallel similarity computation with high energy efficiency. However, these benefits cannot be fully realized without ensuring computational quality. Unlike conventional memories, ReCAM functionality depends on analog sensing behavior, in which Process Variation (PV), manufacturing defects, and circuit non-idealities distort similarity signatures, leading to decision failures and yield loss. Consequently, reliability in ReCAM is fundamentally a challenge of computational correctness, while conventional digital memory test methods are insufficient to capture analog fault manifestations and variation-induced errors. This thesis addresses these challenges by developing a comprehensive cross-level reliability framework for ReCAM architectures. The framework spans fault modeling, March-like testing, PV-aware testing, diagnosis, post-manufacturing calibration, and alternative sensing and monitoring of aging defects, collectively ensuring quality-similarity computation while preserving the performance and energy benefits of analog CiM paradigms.

Zusammenfassung

Ähnlichkeitssuche ist eine fundamentale Operation moderner Rechensysteme und bildet die Grundlage für Anwendungen wie assoziative Speicherabfrage, HDC, DNA-Analyse, Bloom-Filter sowie neuartige AI-Workloads. In konventionellen von-Neumann-Architekturen verursachen diese Operationen erheblichen Datenbewegungsaufwand, was zu gravierenden Latenz- und Energieengpässen führt, die als *Memory Wall* bezeichnet werden. CiM adressiert dieses Problem, indem Berechnungen direkt innerhalb von Speicherarrays durchgeführt werden. Unter den NVM-basierten CiM-Paradigmen bieten aufkommende Technologien wie ReRAM, PCM, STT-MRAM und FeFET durch analoge In-Memory-Berechnung erhebliche Vorteile hinsichtlich Leistung und Energieeffizienz. Unter diesen ist ReRAM besonders vielversprechend aufgrund des großen Widerstandsfensters zwischen HRS und LRS, seiner hohen Skalierbarkeit sowie seiner BEoL-Kompatibilität, was es für dichte und energieeffiziente CiM-Architekturen prädestiniert. Insbesondere ermöglichen ReCAM-Architekturen massiv parallele Ähnlichkeitsberechnungen mit hoher Energieeffizienz. Diese Vorteile können jedoch nur dann vollständig ausgeschöpft werden, wenn die Berechnungsqualität sichergestellt ist. Im Gegensatz zu konventionellen Speichern hängt die Funktionalität von ReCAM wesentlich vom analogen Ausleseverhalten ab, wobei PV, Fertigungsdefekte sowie Schaltungsnichtidealitäten Ähnlichkeitssignaturen verfälschen und zu Entscheidungsfehlern sowie Ausbeuteverlusten führen. Folglich ist Zuverlässigkeit in ReCAM grundlegend eine Herausforderung der Berechnungskorrektheit, während klassische digitale Speichertestverfahren nicht ausreichen, um analoge Fehlereffekte und variationsbedingte Fehler abzubilden. Diese Dissertation adressiert diese Herausforderungen durch die Entwicklung eines umfassenden ebenenübergreifenden Zuverlässigkeitsframeworks für ReCAM-Architekturen. Das Framework umfasst Fehlermodellierung, Marchähnliche Testverfahren, PV-bewusstes Testen, Diagnose, Post-Manufacturing-Kalibrierung sowie alternative Sensorkonzepte und Alterungsüberwachung — und gewährleistet damit qualitativ hochwertige Ähnlichkeitsberechnung bei gleichzeitiger Bewahrung der Leistungs- und Energieeffizienzvorteile analoger CiM-Paradigmen.

Contents

List of Publications

The following publications were authored or co-authored by Haneen Seyam *(H. G. Hezayyin)* during the course of her doctoral research. The first group constitutes the primary scientific contributions on which this dissertation is based, while the second group comprises supporting publications that provide additional context and complementary insights.

Major Publications

1. M. Mayahinia, H. G. Hezayyin, and M. B. Tahoori, "Reliable Analog Computation-in-Memory: From Technology to Application", in *Proceedings of the IEEE VLSI Test Symposium (VTS)*, 2024

2. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Fault Modeling and Testing of ReRAM-based CAM Array", in *Proceedings of the IEEE VLSI Test Symposium (VTS)*, 2025

3. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "PV-ReCAM: Process Variation–Aware Testing for ReRAM-based CAM", in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2025

4. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Design-for-Testability for ReRAM-based CAM", in *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS)*, 2025

5. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Tdsrecam: Time-domain sensing for reliable reram-based content-addressable memory", in *Proceedings of the IEEE European Test Symposium (ETS)*, accepted, 2026

6. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "PVTVAD: PVT-Variation–Aware Fault Diagnosis in ReRAM-based CAM Arrays", *IEEE Transactions on Device and Materials Reliability (TDMR)*, 2026, submitted

7. H. G. Hezayyin and M. B. Tahoori, "Runtime BIST for ReRAM-based CAM using Frequency-Domain Monitoring", in *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS)*, accepted for publication, 2026

Supporting Publications

8. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Testing ReRAM-based TCAM for Computation-in-Memory Applications", in *Proceedings of the IEEE International Conference on Design and Technology of Integrated Systems in Nanoscale Era (DTIS)*, 2024

9. H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "A Post-Manufacturing Calibration Method for ReRAM-based CAM", in *Proceedings of the IEEE VLSI Test Symposium (VTS)*, accepted for publication, 2026

Part I.

Preliminaries

1. Introduction

Conventional computing systems are primarily composed of processor cores and memory subsystems. In such architectures, data and instructions are continuously transferred between memory and processing units for fetching, execution, and storage operations. This frequent data movement introduces considerable latency and energy overhead, commonly referred to as the *memory wall* [10]. With the rapid growth of data-intensive applications, this bottleneck has become a major limitation of modern computing systems [11].

Many emerging applications, such as DNA analysis, Bloom filters (BF), AI, and HDC, heavily rely on similarity-search operations over large datasets [12], [13], [14]. Executing these operations on conventional architectures is inefficient due to the excessive data movement between memory and processors. To address this challenge, CiM has emerged as a promising computing paradigm in which selected computations are performed directly inside the memory array [15]. By minimizing data transfers, CiM significantly improves both performance and energy efficiency [16].

Among various CiM architectures, Content Addressable Memory (CAM) provides an efficient hardware platform for similarity search applications [17]. Unlike conventional memory, CAM is queried using input data rather than memory addresses, enabling row-parallel comparison between stored and input patterns for exact or partial matching. However, conventional CAM implementations are typically based on volatile charge-based technologies such as Static Random Access Memory (SRAM) and Dynamic RAM (DRAM), which suffer from high static power consumption, limited scalability, and lack of non-volatility [17].

To overcome these limitations, several emerging NVM technologies have been explored for CAM realization, including ReRAM, PCM, STT-MRAM, and FeFET [16], [18]. These technologies eliminate static leakage power while naturally enabling analog computation through their resistive behavior. Among them, ReRAM is particularly attractive due to its distinct HRS and LRS states, high density, and compatibility with analog similarity computation [19], [20]. As a result, ReCAM has emerged as a promising candidate for scalable and energy-efficient analog CiM architectures [21].

In ReCAM, similarity computation is performed in the analog domain by precharging the output voltage and conditionally discharging it through activated memory cells [22]. The resulting match-line (ML) voltage reflects the similarity between stored and input patterns and is typically digitized using an Analog-to-Digital Converter (ADC). Although this analog sensing mechanism enables highly efficient in-memory similarity computation, it also introduces significant reliability and quality challenges [23].

At the device level, ReRAM exhibits non-ideal behaviors such as PV, resistance fluctuations, stochastic switching, drift, and aging effects [20], [23]. At the circuit level, the integration of ReRAM with CMOS introduces additional challenges, including sensing offsets, interconnect parasitics, and peripheral circuit variations [24]. These non-idealities

distort the voltage output distributions and reduce the separation between adjacent similarity levels. Due to the analog nature of ReCAM, even small deviations can lead to incorrect similarity decisions.

Nevertheless, these effects do not always directly result in functional failure. In many cases, variations shift the overall system response rather than completely corrupting it. For instance, global (die-to-die) variations may shift the ADC outputs of an entire chip, causing it to fail conventional pass/fail tests despite remaining functionally correct. In contrast, local (within-die) variations may create overlap between similarity levels inside the same chip, making recovery considerably more difficult. In addition, runtime phenomena such as aging and drift can progressively degrade sensing margins [25].

Ensuring the *quality* of ReCAM architectures therefore requires a comprehensive framework that extends beyond conventional reliability analysis. It demands systematic solutions across multiple stages, including manufacturing testing, post-manufacturing calibration, fault diagnosis, robust sensing, and runtime monitoring. Traditional testing techniques, such as March tests originally developed for SRAM-based CAMs, are insufficient for ReCAM because they fail to capture analog behavior and ReRAM-specific fault mechanisms [26]. Furthermore, conventional go/no-go testing may unnecessarily reduce yield by discarding chips that could otherwise operate correctly after calibration.

1.1. Research Problem and Gap

Although ReCAM provides an attractive platform for energy-efficient similarity search, ensuring its reliable operation remains challenging. The main difficulty is that ReCAM does not produce purely digital outputs; instead, the computational result is encoded in analog ML voltage or timing behavior. As a result, the correctness of the search operation depends not only on the stored data but also on the sensing margin, device variability, peripheral circuitry, and operating conditions.

Conventional memory test methodologies are not sufficient to address these challenges. They are mainly designed to detect digital storage faults and typically assume that the read output is either correct or faulty. In contrast, faults and variations in ReCAM may appear as small analog deviations, shifted voltage distributions, similarity-level overlap, or gradual runtime degradation. These effects can either cause a correct chip to fail a fixed pass/fail test or allow defective behavior to escape detection.

Therefore, ReCAM requires a quality assurance framework that jointly considers defect modeling, process variation, sensing accuracy, calibration, diagnosis, and runtime monitoring. Such a framework must bridge the gap between device-level non-idealities, circuit-level sensing behavior, and architecture-level similarity computation.

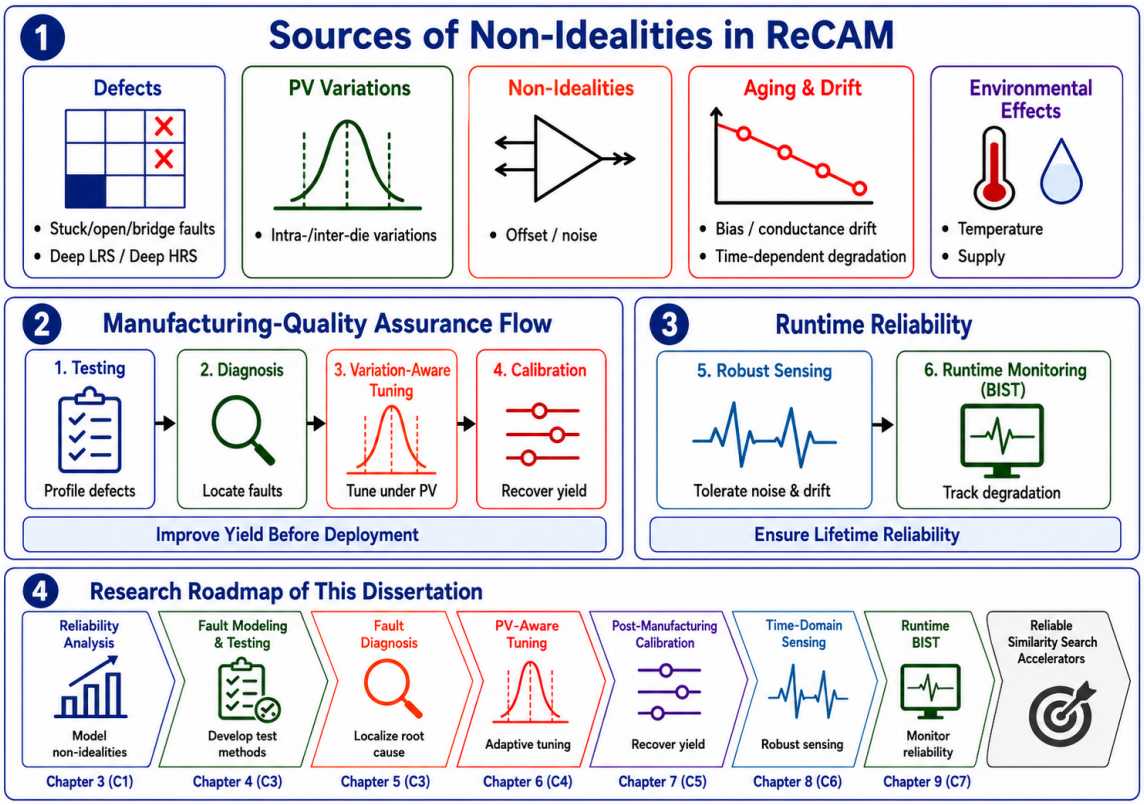


Figure 1.1.: Reliability and quality assurance roadmap for ReRAM-based CAM. The figure summarizes the main sources of non-idealities, the manufacturing-time quality assurance flow, the runtime reliability enhancement strategy, and the dissertation contributions from Chapter 3 (C1) to Chapter 9 (C7).

1.2. Thesis Scope and Objective

The objective of this dissertation is to develop a comprehensive quality assurance framework for ReCAM architectures used in analog CiM systems. The scope of the thesis covers both manufacturing-time and runtime reliability challenges. At manufacturing time, the thesis addresses defect-oriented fault modeling, March-like testing, process-variation-aware testing, fault diagnosis, and post-manufacturing calibration. At runtime, it investigates alternative sensing and monitoring techniques to improve robustness against degradation mechanisms such as aging and drift.

The overall reliability and quality assurance framework of this dissertation is summarized in Figure 1.1. The figure highlights the major sources of non-idealities in ReCAM, the manufacturing-time quality assurance flow, the runtime reliability enhancement strategy, and the mapping of the thesis contributions from Chapter 3 (C1) to Chapter 9 (C7).

Rather than treating testing, sensing, calibration, and monitoring as independent problems, this dissertation studies their interaction in the context of analog similarity computation. The central goal is to improve the quality of ReCAM while preserving the main advantages of analog CiM, namely high parallelism, low data movement, and energy-efficient search operation.

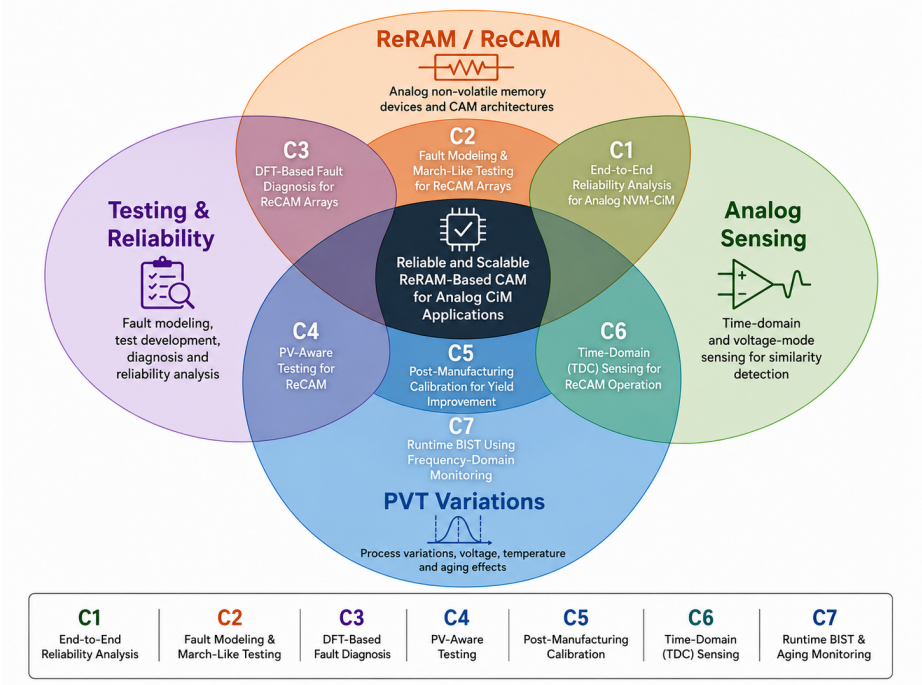


Figure 1.2.: Intersection of the main research domains and contributions of this dissertation. The figure maps the seven contributions C1–C7 across ReRAM/ReCAM technology, analog sensing, testing and reliability, and PVT variations, highlighting their combined role in enabling reliable and scalable ReRAM-based CAM for analog CiM applications.

1.3. Contribution and Research Directions

This dissertation focuses on the quality assurance of ReCAM architectures under manufacturing defects, process variations, sensing limitations, and runtime degradation mechanisms. Figure 1.2 illustrates the intersection of the main research domains addressed in this dissertation. The contributions span ReRAM/ReCAM technology, analog sensing, testing, reliability, and Process, Voltage, and Temperature Variations (PVT) variations. This view complements the roadmap in Figure 1.1 by showing how the individual contributions C1–C7 relate to the broader research landscape.

The highlighted contributions, denoted as C1–C7 in Figure 1.2, are summarized as follows:

C1 End-to-end reliability analysis for analog Computation in Non-volatile Resistive Memories (NVM-CiM): An end-to-end reliability analysis framework for analog NVM-CiM systems is developed that spans the technology, circuit, architecture, and application levels. The framework models the impact of NVM device variability, sensing offsets, interconnect parasitics, crossbar size, activated operands, and operand locations on the correctness of computation. The resulting hardware-level errors are further propagated to higher abstraction levels using a CiM-enabled full-system simulation framework. Although this contribution is broader than ReCAM, it establishes the central reliability mechanism addressed throughout this dissertation: analog computation errors caused by distorted sensing signatures and device/circuit

non-idealities. The remaining chapters specialize this insight to ReCAM, which is the main architectural target of the thesis. This work is presented in Chapter 3 and published in [1].

C2 Fault modeling and March-like testing for ReCAM arrays: A defect-oriented fault modeling methodology for ReCAM arrays is proposed by analyzing both intra-cell and inter-cell defects from the cell level to the array level. Based on the resulting faulty behaviors, equivalent fault groups are identified, and a symmetric March-like test algorithm is developed to achieve full coverage of ReCAM-specific faults. This work is presented in Chapter 4 and published in [2], [8].

C3 Design-for-Testability (DfT)-based fault diagnosis for ReCAM arrays: A DfT-based diagnosis methodology is proposed for accurate fault localization in ReCAM arrays. The approach exploits small ML voltage differences caused by fault locations and enhances their observability with a tunable Operational Amplifier (OpAmp) that adjusts gain and sampling time. A multi-step decision-tree diagnosis flow is further developed to reduce diagnosis latency while maintaining low area and power overhead. This work is presented in Chapter 5 and published in [4], [6].

C4 Process Variation (PV)-aware testing for ReCAM: The impact of process variation on ReCAM functionality is analyzed, demonstrating that conventional March tests may fail in the presence of PV-induced soft faults or when PV masks hard defects. To address this challenge, a size-dependent PV-aware March-like testing methodology is proposed. The approach targets adjacent similarity levels with the smallest ML voltage separation, achieving full coverage for both hard defects and PV-induced soft faults. This work is presented in Chapter 6 and published in [3].

C5 Post-manufacturing calibration for yield improvement: A variation-aware post-manufacturing calibration methodology for ReCAM arrays is proposed to improve yield under process variations. The method distinguishes between local variations, which may cause unrecoverable overlap between similarity levels, and global variations, which mainly shift ADC outputs across the chip. A per-chip Look-Up Table (LUT)-based mapping mechanism is developed to recover chips primarily affected by global variations. This work is presented in Chapter 7 and published in [9].

C6 Time-to-Digital Converter (TDC)-based time-domain sensing for reliable ReCAM operation: A time-domain sensing methodology is investigated as an alternative to conventional ADC-based voltage-domain sensing for ReCAM. The proposed TDC-augmented architecture converts ML voltage decay into a time interval that is digitized using a low-overhead TDC. The proposed approach reduces sensing area and power while improving robustness against PV and lowering the bit-error rate. This work is presented in Chapter 8 and published in [5].

C7 Runtime Built-In Self-Test (BIST) using frequency-domain monitoring: A lightweight runtime BIST framework based on voltage-controlled ring-oscillator (VCRO) sensing is proposed to detect aging, drift, and runtime degradation. The

framework converts ML behavior into frequency-domain signatures that can be monitored during runtime. In addition, a multi-tier k-of-n voting strategy is employed to improve robustness against transient variations, while fault localization and X-mask-based tolerance mechanisms help preserve array functionality. This work is presented in Chapter 9 and published in [7].

This dissertation is organized as follows:

- Chapter 2 presents the background on analog CiM, emerging NVM technologies, and ReCAM architectures, along with their associated reliability challenges, providing the foundation for the contributions presented in **C1–C7**.
- Chapter 3 introduces an end-to-end reliability analysis and mitigation framework for analog NVM-CiM systems across the technology, circuit, architecture, and application levels (**C1**).
- Chapter 4 presents defect-oriented fault modeling and March-like testing methodologies for ReCAM arrays (**C2**).
- Chapter 5 describes the proposed DfT-based diagnosis framework for accurate fault localization in ReCAM arrays (**C3**).
- Chapter 6 introduces a PV-aware testing methodology that improves fault coverage under variability conditions (**C4**).
- Chapter 7 presents a post-manufacturing calibration framework for compensating global variations and improving yield (**C5**).
- Chapter 8 proposes a time-domain sensing approach for ReCAM to reduce sensing overhead and improve robustness (**C6**).
- Chapter 9 introduces a runtime BIST framework based on frequency-domain monitoring for detecting aging and degradation effects (**C7**).
- Chapter 10 recaps the contributions **C1–C7** and suggests directions for future research.

Although Chapter 3 addresses analog NVM-CiM in a broader context, it establishes the fundamental reliability problem that motivates the ReCAM-specific contributions of this dissertation. In particular, it shows that the correctness of analog computation is strongly affected by device variability, sensing inaccuracies, interconnect parasitics, and peripheral circuit non-idealities. These effects manifest as decision failures, where the analog computation result is incorrectly digitized or interpreted. ReCAM represents a concrete and highly relevant instance of this broader analog-CiM reliability problem, since its similarity result is encoded in the analog match-line response. Therefore, the subsequent chapters specialize the general decision-failure analysis of Chapter 3 to ReCAM architectures and develop targeted solutions for defect modeling, March-like testing, fault diagnosis, process-variation-aware testing, post-manufacturing calibration, robust sensing, and runtime monitoring.

2. Background

This chapter presents the background and fundamental concepts required for the remainder of this dissertation. First, emerging NVM technologies are introduced with particular emphasis on ReRAM devices, their physical switching mechanisms, and inherent non-idealities. Next, the principles of CiM are outlined, followed by a discussion of conventional and NVM-based CAM (NVM-CAM) architectures and their operation. The chapter then presents ReCAM architectures in detail, covering their analog sensing mechanisms and the limitations of conventional voltage-domain sensing. Finally, the major reliability and quality-assurance challenges arising from device variability, circuit imperfections, fabrication defects, and runtime degradation are highlighted, thereby motivating the testing, diagnosis, calibration, and runtime-monitoring methodologies developed throughout this dissertation.

2.1. Emerging Non-Volatile Memory Technologies (NVMs)

Emerging NVMs have attracted significant attention as alternatives to conventional charge-based memories due to their non-volatility, high density, scalability, and compatibility with CiM architectures [15], [16], [18]. Unlike conventional volatile memories such as SRAM and DRAM, which require continuous power to preserve stored data, NVMs retain information even when the supply voltage is removed. This non-volatile behavior significantly reduces static leakage power and makes NVMs highly attractive for energy-efficient memory arrays and in-memory computing paradigms.

Several NVM technologies have been explored for storage and computation, including ReRAM, PCM, STT-MRAM, and FeFET. Although these technologies rely on different physical switching mechanisms, they commonly provide multiple electrically distinguishable resistance states that can be utilized for data storage and analog computation. Table 2.1 summarizes and compares these technologies alongside conventional memories (SRAM, DRAM, and Flash) across key metrics, including non-volatility, read speed, density, endurance, and analog-CiM suitability. As shown in Table 2.1, while technologies such as PCM, STT-MRAM, and FeFET offer moderate analog-CiM suitability, ReRAM stands out with excellent suitability, very high density, and competitive endurance, making it the primary focus of this dissertation.

Among these technologies, ReRAM has emerged as one of the most promising candidates for analog CiM, particularly for similarity-search applications, which constitute the primary focus of this dissertation. ReRAM offers several attractive properties, including compact device structure, high integration density, low leakage power, compatibility with high-density crossbar architectures, and a large resistance window between the HRS and

Table 2.1.: Comparison of Emerging Non-Volatile Memory (NVM) Technologies

Technology	Storage Mechanism	Non-Volatile	Read Speed	Density / Scalability	Endurance (Cycles)	Analog-CiM Suitability
SRAM [27], [28]	Charge in latch	**No**	Very High (~1 ns)	**Low**	$> 10^{15}$	**Limited**
DRAM [27], [28]	Charge in capacitor	**No**	High (~10 ns)	**Medium**	$> 10^{12}$	**Limited**
Flash (NAND) [29], [30]	Charge in floating gate	**Yes**	Low (~10–100 μs)	**High**	10^3–10^5	**Limited**
PCM [18], [31]	Phase change (amorphous / crystalline)	**Yes**	Medium (~10–100 ns)	**Medium**	10^6–10^8	**Moderate**
STT-MRAM [32], [33]	Magnetic tunneling junction	**Yes**	High (~5–20 ns)	**Medium**	10^{12}–10^{15}	**Moderate**
FeFET [34], [35]	Ferroelectric polarization	**Yes**	High (~10–50 ns)	**Medium**	10^8–10^{10}	**High**
ReRAM [16], [36], [37], [38]	Resistive switching (oxygen vacancy)	**Yes**	High (~10–50 ns)	**Very High**	10^8–10^{10}	**Excellent**

LRS [19], [20]. These characteristics enable efficient analog computation via resistive behavior and facilitate scalable CiM implementations.

2.1.1. Resistive-based RAM (ReRAM) Device and Switching Mechanism

A typical ReRAM device consists of an oxide-based switching layer sandwiched between two metallic electrodes, commonly referred to as the top electrode (TE) and bottom electrode (BE). As illustrated in Figure 2.1, the resistance state of the device is controlled by the formation and rupture of conductive filaments within the oxide layer, a process primarily governed by the redistribution of oxygen vacancies [20], [23].

Initially, the oxide layer exhibits a highly insulating behavior and may require an electroforming step to establish a conductive path. After forming, the device can be reversibly switched between two distinct resistance states. During the *SET* operation, the applied voltage promotes oxygen-vacancy migration and the formation of conductive filaments, thereby driving the device into the LRS. Conversely, during the *RESET* operation, the filament is partially ruptured or weakened, returning the device to the HRS. The resulting resistance contrast between HRS and LRS enables binary information storage and resistance-based sensing.

As shown in Figure 2.1(a), the geometry of the conductive filament strongly influences the device resistance. A thicker, more continuous filament generally results in lower resistance, whereas a thinner or partially ruptured filament results in higher resistance.

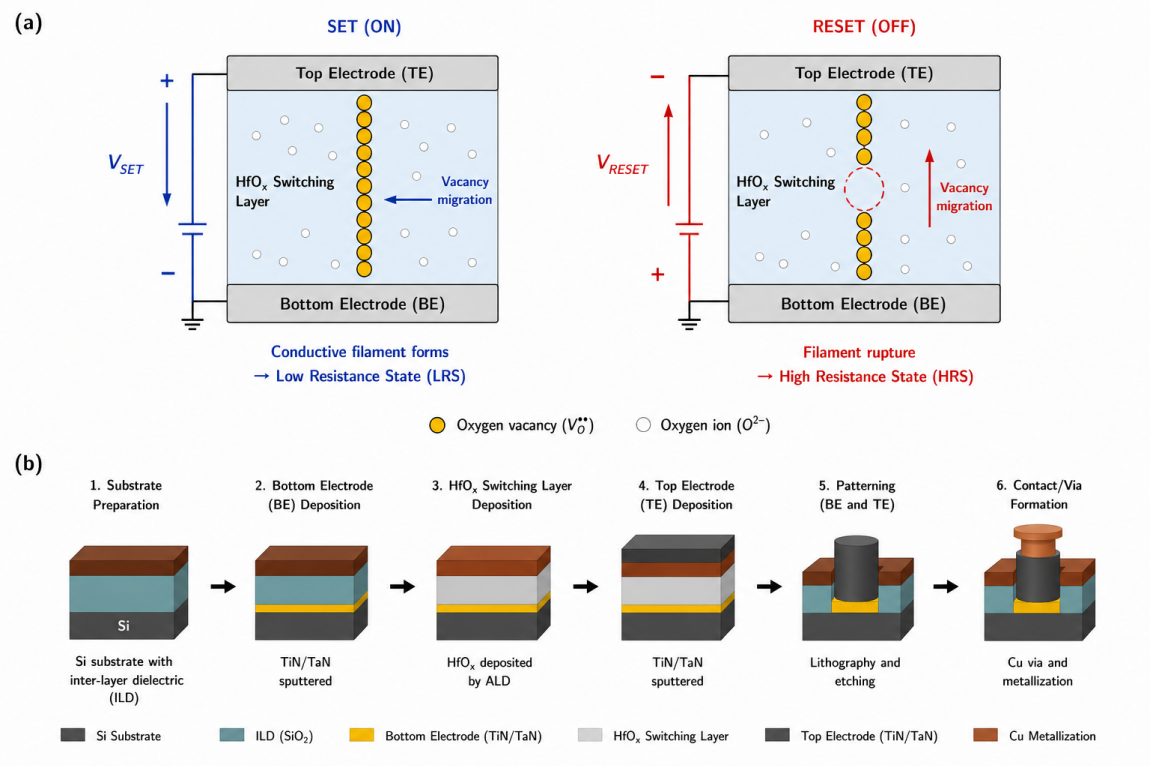


Figure 2.1.: Overview of the ReRAM device structure and fabrication process. (a) Filamentary switching mechanism in an HfO_x-based ReRAM device, where conductive filament formation (SET) and rupture (RESET) enable transitions between the low-resistance state (LRS) and high-resistance state (HRS). (b) Representative back-end-of-line (BEOL) fabrication flow of a ReRAM device, including bottom-electrode deposition, switching-layer formation, top-electrode deposition, patterning, and contact/via formation

Consequently, even small variations in filament shape, size, or stability can significantly affect the programmed resistance value and the reliability of read operations [39].

Beyond the switching mechanism, Figure 2.1(b) summarizes the typical back-end-of-line (BEOL) fabrication flow of HfO_x-based ReRAM technology. The process begins with substrate preparation and proceeds through bottom-electrode deposition, switching-layer formation, top-electrode deposition, patterning, and contact/via formation. This fabrication flow demonstrates the compatibility of ReRAM integration with advanced CMOS manufacturing processes and highlights the practical feasibility of ReRAM-based computing.

Industrial demonstrations and foundry-qualified ReRAM platforms, including TSMC's ReRAM technology [40], further emphasize the maturity and practical viability of ReRAM for large-scale CiM.

2.1.2. ReRAM Non-Idealities and Variations

Despite its promising characteristics, ReRAM suffers from several intrinsic non-idealities that are particularly critical in analog CiM architectures, where computation directly depends on resistance-dependent current or voltage behavior. The stochastic nature of filament formation and rupture causes the programmed HRS and LRS values to vary

across devices and across switching cycles, such that resistance states cannot be treated as deterministic values but rather as statistical distributions that broaden with continued use [20], [23].

Specifically, two primary sources of variation can be identified. Device-to-device (D2D) variation arises from fabrication imperfections, material non-uniformity, oxide thickness fluctuations, and local differences in filament formation. Cycle-to-cycle (C2C) variation originates from the stochastic redistribution of oxygen vacancies during repeated SET and RESET operations, leading to inconsistent resistance values even for the same target state.

In addition to these spatial and cycle-dependent variations, ReRAM devices are also subject to temporal degradation mechanisms. Resistance drift can gradually shift the programmed resistance over time, even without additional switching activity, while temperature, bias stress, and aging can further accelerate these shifts and reduce the separation between HRS and LRS distributions [20], [23]. These combined non-idealities pose significant challenges for reliable analog sensing in large ReCAM arrays, as even small resistance deviations can distort similarity signatures and lead to incorrect search results.

2.2. Computation-in-Memory (CiM)

In conventional computing systems, data and instructions are repeatedly transferred between memory and processor cores during program execution. This frequent data movement introduces considerable latency and energy overhead, commonly known as the *memory wall* [10]. The impact of this bottleneck has become increasingly severe as data-intensive applications such as AI, similarity search, and HDC have grown [11], [12].

CiM addresses this limitation by performing selected computations directly inside or near the memory array, as illustrated in Figure 2.2. In analog CiM, input operands are applied as voltages, while stored operands are encoded as device conductances. The resulting current or voltage reflects the computation according to Ohm's and Kirchhoff's laws, enabling massively parallel operations such as vector–matrix multiplication, Boolean computation, and similarity search. By eliminating repeated data transfers between memory and logic, CiM significantly improves both energy efficiency and computational throughput. As shown in Figure 2.2, both charge-based memories (SRAM, DRAM, and Flash) and resistance-based memories (ReRAM, PCM, and STT-MRAM) can serve as computational memory elements [15].

While CiM can be implemented using conventional charge-based memories, NVMs are especially attractive because their resistive behavior naturally enables analog computation [15], [16]. However, the fabrication process of NVM devices is not fully mature and is prone to process variations. As a result, rather than exhibiting fixed resistance values, the LRS and HRS states follow statistical distributions characterized by a mean μ_R and standard deviation σ_R. These distributions are further affected by temporal degradation mechanisms: after programming, the resistance state can drift over time, potentially causing data loss and retention failures [41]. Temperature sensitivity adds another source of instability, as thermal variations can shift resistance distributions and reduce the separation between HRS and LRS [42], [43]. Consequently, analog CiM relies on accurate sensing of small resistance-dependent signals, making the system inherently sensitive to the device vari-

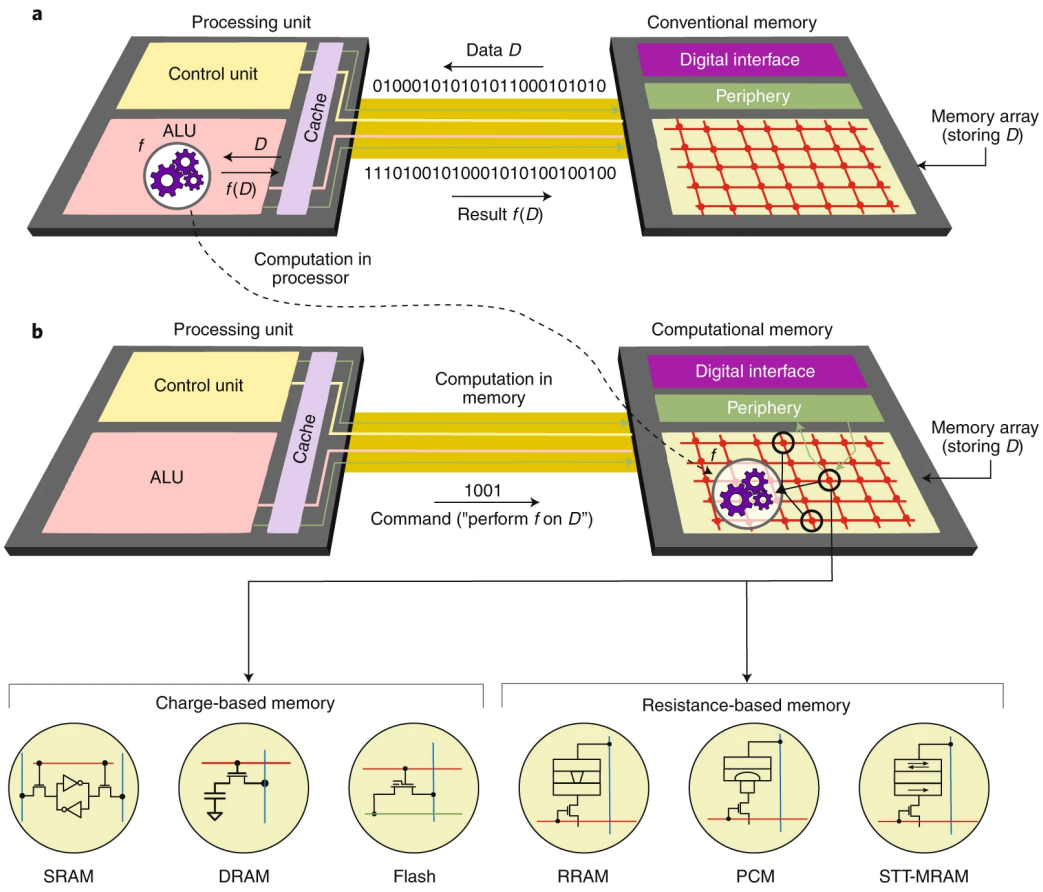


Figure 2.2.: (a) In conventional computing, executing an operation f on data D requires transferring data between memory and the processor, leading to latency and energy overhead known as the memory bottleneck. In contrast, in-memory computing performs $f(D)$ directly within the memory array by exploiting the physical properties of memory devices, reducing data movement and improving energy efficiency and performance. Computation is carried out within the memory array and peripheral circuitry without explicitly reading individual memory cells. Both charge-based memories (SRAM, DRAM, Flash) and resistance-based memories (ReRAM, PCM, STT-MRAM) can serve as computational memory elements [15]

ability, interconnect parasitics, and circuit imperfections — challenges that are addressed throughout this dissertation.

2.2.1. Analog NVM-CiM Principle

In analog NVM-CiM, input operands are typically applied as voltages, while stored operands are encoded as device conductances. The computation is performed inside the memory array by exploiting Ohm's law and Kirchhoff's current law. For example, in analog Multiply-Accumulate (MAC) operations, the current through each memory device is proportional to the product of the applied voltage and the stored conductance. The total current collected at the output line corresponds to the accumulation of these products, enabling massively parallel analog computation [16], [44].

Analog NVM-CiM can also support Boolean operations and similarity-search applications. In these architectures, the final computation result is obtained by sensing an analog quantity such as current, voltage, or discharge time. Therefore, the correctness of the

computation depends not only on the stored memory states but also on the quality of the sensing mechanism.

Further, works presented in [45], [46], [47], [48] considered the MAC-based NVM-CiM, and their reliability target is the IR drop. Although these works comprehensively analyzed the technology, circuit, and application levels, they did not consider 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 [49], [50].

2.2.2. Decision Failure in Analog CiM

A major reliability challenge in analog CiM is *decision failure*. Decision failure occurs when an analog signal generated by the memory array is incorrectly interpreted by the sensing circuitry, resulting in an erroneous digital output. This can be caused by device variability, resistance drift, sensing offsets, quantization errors, interconnect parasitics, or temperature-dependent behavior.

Unlike purely digital memory operations, analog CiM often requires distinguishing between multiple analog output levels. If the distributions corresponding to different computational results overlap, the sensing circuitry may assign the output to the wrong state. This problem becomes increasingly severe as the array size increases because the separation between neighboring output levels decreases while device and circuit non-idealities accumulate.

Among different CiM architectures, CAM-based architectures are particularly attractive for similarity-search applications because they inherently support massively parallel search operations.

2.3. Content-Addressable Memory (CAM)

CAM is a memory architecture designed for parallel search operations. Unlike conventional memories, which are accessed using addresses, CAM arrays are queried using data patterns, enabling fast lookup operations that are essential in networking, security, pattern matching, and similarity search applications [17]. The input query is compared in parallel against all stored entries simultaneously, enabling fast exact or partial matching.

Conventional CAM implementations are based on SRAM, which offers high performance, but suffers from large area overhead and high static power consumption due to its volatile, charge-based nature [17].

2.3.1. Conventional SRAM-based CAM (SRAM-CAM)

Conventional CAM implementations are typically based on SRAM. Although SRAM-based CAM (SRAM-CAM) provides high-speed search capability, they suffer from large cell area, high static power consumption, and limited scalability. Moreover, SRAM is volatile and requires continuous power to retain stored data. These limitations motivate the exploration

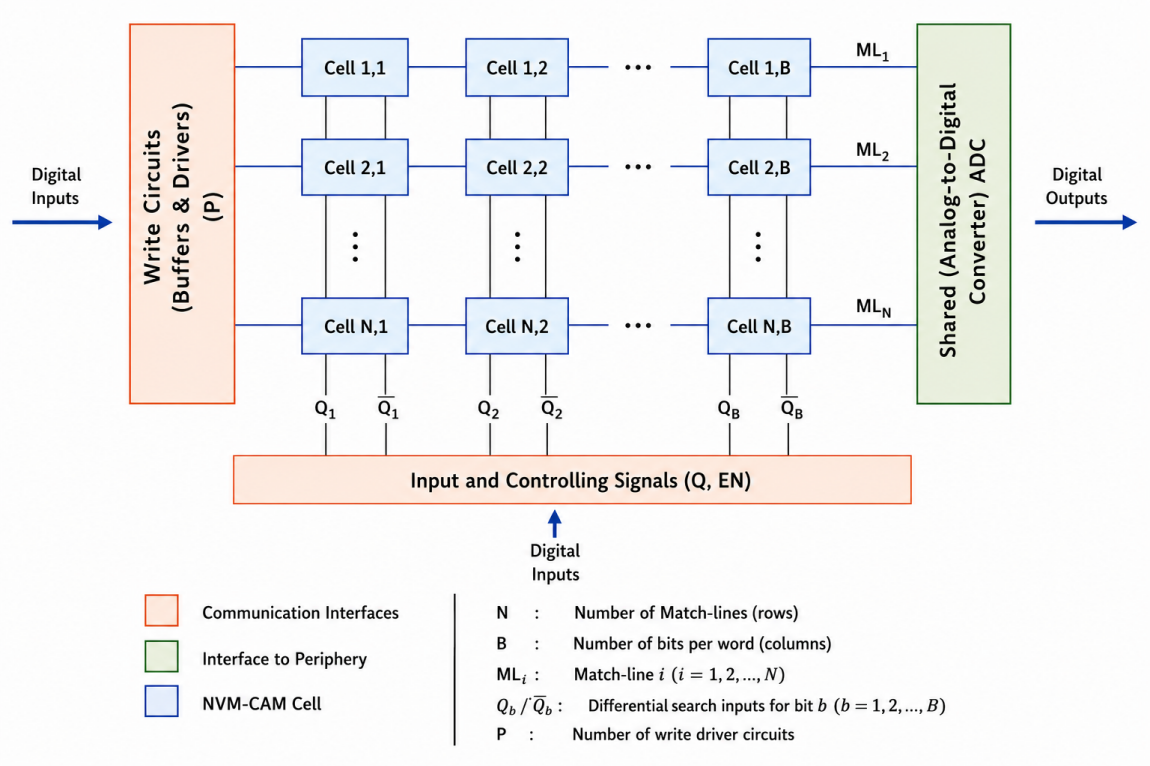


Figure 2.3.: NVM-based CAM (NVM-CAM)

of NVM-based CAM architectures, which eliminates static leakage power, provides denser storage, and naturally supports analog computation through their resistive behavior.

2.3.2. NVM-based CAM (NVM-CAM)

In NVM-CAM architectures, stored patterns are represented by resistive memory devices rather than volatile SRAM cells. The non-volatile nature of NVMs eliminates static leakage power for data retention, while their resistive states enable analog comparison inside the memory array. This makes NVM-CAM attractive for applications where stored prototypes are written infrequently but searched repeatedly.

A typical NVM-CAM compares two binary patterns: a stored prototype pattern P and an input query pattern Q. The comparison result determines the similarities between the patterns. Since the rest of the architecture is digital, an ADC is necessary to encode the similarity information, as shown in Figure 2.3.

2.3.2.1. ReRAM-based CAM (ReCAM)

Among different NVM-CAM implementations, ReCAM is particularly promising because it combines the high density and non-volatility of ReRAM with the parallel search capability of CAM [51]. A standard ReCAM bit-cell is implemented using a 2T2R structure, consisting of two ReRAM devices and two access transistors. The two ReRAM devices store a prototype bit P and its complement, while the two access transistors steer the comparison path based on the input query bit Q and its complement.

During a search operation, the ML is first precharged to V_{DD}. Depending on the relation between P and Q, the ML conditionally discharges through either a high-resistance or low-resistance path. In the case of a *match*, the discharge path includes an HRS device, resulting in slower ML voltage decay. In the case of a *mismatch*, the discharge path includes an LRS device, resulting in faster ML voltage decay.

At the word level, multiple ReCAM cells are connected to the same ML, and all bit-cells contribute to a combined discharge behavior that reflects the overall degree of similarity between the stored prototype and the query pattern. A larger number of matching bits results in a larger equivalent discharge resistance, slower ML discharge, and a higher residual ML voltage at the sampling time. The ML voltage can be approximated by the exponential discharge model:

$$V_{ML}(t_s) = V_{DD} \cdot e^{-\frac{t_s}{R_{eq} C_{ML}}}, \tag{2.1}$$

where V_{DD} is the supply voltage, t_s is the sampling time, R_{eq} is the equivalent discharge resistance, and C_{ML} is the ML capacitance. Since R_{eq} depends on the number of matching bits, the sampled ML voltage provides an analog representation of the Hamming similarity between the stored prototype and the query pattern, as illustrated in Figure 2.4(a). The resulting ML voltage decay and its corresponding ADC output distributions under process variation are shown in Figure 2.4(b) and (c), highlighting how variation-induced overlap between adjacent similarity levels can lead to incorrect search results. Therefore, the reliability of ReCAM strongly depends on the accuracy of the sensing mechanism used to interpret this analog ML behavior, as discussed in detail in Sections 2.4 and 2.5.

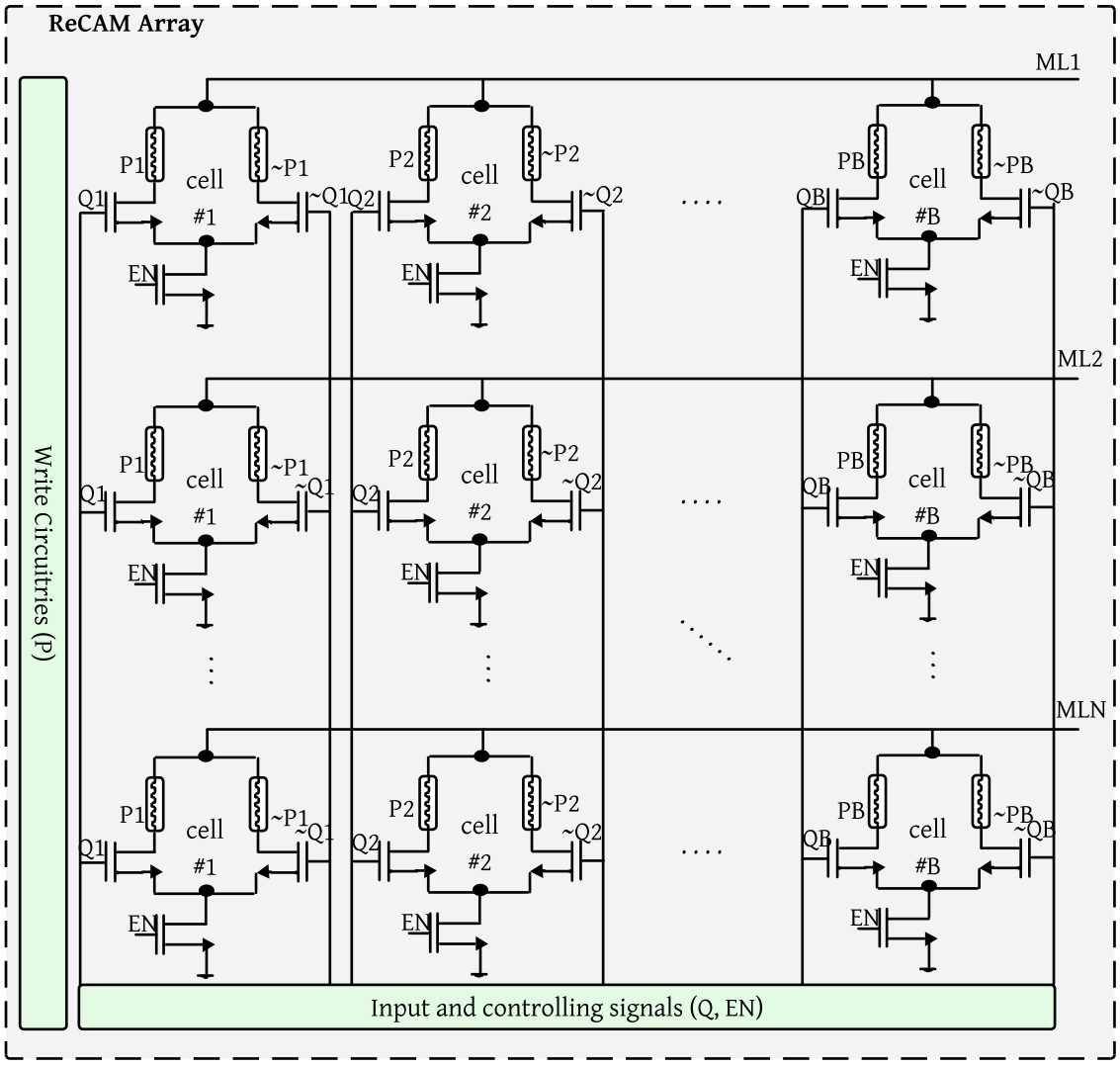
2.4. Sensing Mechanisms in ReCAM

Since the output of a ReCAM operation is an analog ML voltage, a sensing circuit is required to convert it into a digital value. This digital value is then interpreted as an all-match, an all-mismatch, or a multi-level similarity result. Therefore, sensing is a critical component in determining the reliability and quality of ReCAM.

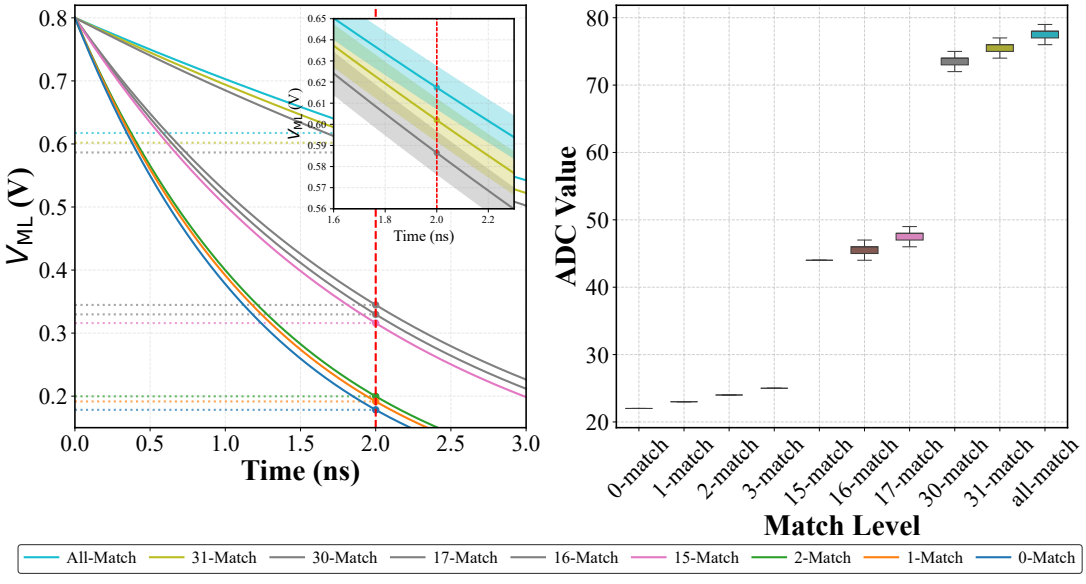
2.4.1. Conventional Voltage-Domain Sensing

Conventional ReCAM architectures primarily rely on voltage-domain sensing to evaluate the similarity between stored and input patterns [51], [52]. During the search operation, the ML is first precharged and then conditionally discharged through resistive paths determined by the comparison result between the stored prototype and the query pattern. The ML voltage is sampled at a predefined sensing time; a higher residual ML voltage indicates more matching bits, while a lower ML voltage indicates more mismatches.

To distinguish multiple similarity levels, ReCAM architectures commonly employ an ADC or a set of voltage comparators to digitize the sampled voltage [51], [52]. The resulting digital code is subsequently mapped to a similarity level or inverse Hamming distance. In practical implementations, a single ADC may be shared among multiple MLs to reduce area and power overhead.



(a)



(b) (c)

Figure 2.4.: Preliminaries of ReCAM and conventional voltage-domain sensing. (a) ReCAM array organization. (b) and (c) ML voltage decay and corresponding ADC output distributions under variation

2.4.2. Limitations of Conventional Sensing

The primary limitations of conventional voltage-domain sensing are its high sensitivity to analog non-idealities and scalability challenges. Device variations can shift and widen the ML voltage distributions, while interconnect parasitics distort the ML discharge behavior, particularly in large arrays. As the array dimension increases, the number of possible similarity levels also increases, causing the voltage separation between adjacent levels to become progressively smaller. Consequently, distinguishing neighboring similarity levels becomes increasingly difficult and highly susceptible to noise and process variations.

Furthermore, ADC-based sensing introduces significant hardware overhead in large-scale ReCAM arrays. Accurately resolving closely spaced ML voltage levels often requires high-resolution ADCs, which increase area, power consumption, and sensing latency [11], [53]. In addition, ADC offsets, quantization noise, reference-voltage fluctuations, and comparator mismatches can further degrade sensing accuracy and reduce decision margins [53], [54]. These challenges are particularly critical in analog CiM architectures, where reliable computation depends directly on accurately resolving infinitesimal ML voltage differences amid process variations and runtime non-idealities.

Therefore, reliable ReCAM operation requires not only robust memory devices but also efficient sensing methodologies capable of compensating for these analog imperfections.

2.5. Reliability Challenges in ReCAM

Although ReCAM offers attractive benefits for analog similarity search, its reliability is compromised by interactions among ReRAM device non-idealities, CMOS peripheral circuits, and interconnect parasitics.

2.5.1. ReRAM Defects

ReRAM devices are vulnerable to several manufacturing defects and abnormal resistance states [55], [56]. A common class of defects is stuck-at faults, where the device becomes permanently stuck in either HRS or LRS. A stuck-at-HRS defect prevents the device from switching to LRS, while a stuck-at-LRS defect prevents switching to HRS. Such defects can directly corrupt the expected match or mismatch behavior of a ReCAM cell.

Other defects include deep-HRS and deep-LRS states. Deep-HRS may occur when the filament is excessively ruptured, resulting in abnormally high resistance. Deep-LRS may occur when the filament becomes too thick or stable, resulting in abnormally low resistance. In addition, undefined intermediate resistance states may arise from irregular filament formation or incomplete switching.

In addition to ReRAM defects, transistor and interconnect defects can also affect ReRAM functionality. Access transistors may suffer from stuck-on or open faults, causing unintended conduction or preventing the intended discharge path. Interconnect defects, such as resistive opens or weak connections, can further distort ML discharge behavior.

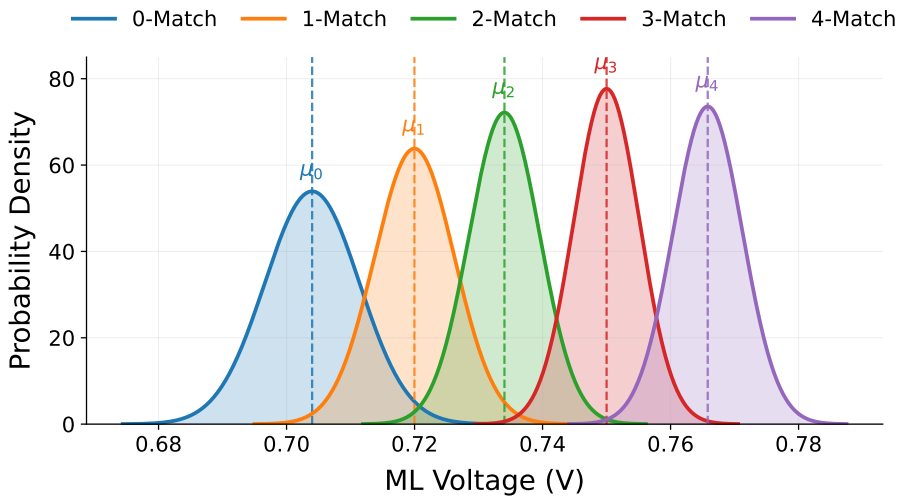


Figure 2.5.: Overlapping between the similarity levels due to process variations (PV)

2.5.2. Process Variations (PV)

PV is a major challenge in analog ReCAM architectures. Because the similarity result is encoded as an analog ML voltage, variations in ReRAM resistance, transistor parameters, and interconnect parasitics can shift or broaden the voltage distributions associated with different similarity levels [15], [16], [18] as shown in Figure 2.5.

PV can be broadly classified as global and local [57], [58]. Global variations affect an entire chip in a similar direction and may shift all ADC outputs while preserving the relative ordering between similarity levels. In contrast, local variations affect devices differently within the same array, potentially creating overlap between adjacent similarity distributions.

Such overlap is particularly critical because the sensing circuitry can no longer reliably distinguish neighboring similarity levels. Consequently, process variation may lead to incorrect similarity decisions even when no permanent hard defect exists [3], [59].

2.5.3. Interconnect and Peripheral Non-Idealities

In addition to device-level variations, ReCAM reliability is affected by circuit-level non-idealities. The ML and other interconnects exhibit resistive and capacitive parasitics that alter discharge behavior and introduce location-dependent effects. Cells farther from the sensing circuitry may contribute differently to the sensed ML voltage than cells closer to it [60], [61].

Peripheral circuits also introduce non-idealities. Comparators and ADCs suffer from offset, noise, finite resolution, and reference-voltage inaccuracies. These imperfections may shift effective decision boundaries between similarity levels. When voltage margins become small, such non-idealities can directly lead to incorrect similarity decisions [62], [63].

2.5.4. Aging and Drift

Runtime degradation further affects the long-term reliability of ReCAM. Repeated switching operations may alter the filament structure and gradually shift the resistance distribution. Even without frequent switching, resistance drift can change the programmed state over time due to material relaxation and vacancy redistribution [16], [64], [65].

Aging and drift gradually reduce sensing margins by shifting or widening ML voltage distributions. These effects may not immediately cause functional failure, but they can progressively increase the probability of decision failure during operation. Therefore, one-time manufacturing testing and static calibration may be insufficient to guarantee reliable ReCAM operation throughout the device lifetime [15], [59].

2.6. Manufacturing Test and Quality Assurance

As semiconductor technologies continue to scale, manufacturing testing has become an essential stage in the chip development cycle, as illustrated in Figure 2.6. As shown in Figure 2.6, the development cycle spans design, fabrication, manufacturing, testing, and validation stages, with testing serving as the critical gate between fabrication and deployment. The objective of manufacturing testing is to identify defective chips after fabrication, ensure correct functionality, and improve overall product quality and reliability [66], [67].

Conventional digital memories are typically tested using March algorithms and pass/fail criteria that verify correct read and write operations under predefined conditions [26], [68]. These techniques are highly effective for binary memory systems in which outputs can be classified as either correct or faulty. However, applying such methodologies directly to analog ReCAM architectures is considerably more challenging, since similarity is expressed as a continuous ML voltage rather than a binary digital output [15], [16].

Unlike conventional memories, ReCAM operation depends on analog sensing margins and similarity-level separation. As a result, defects and variations do not always manifest as catastrophic failures, but may instead appear as shifted voltage distributions, similarity-level overlap, or degraded sensing accuracy [15], [62], [63]. For example, global variations may shift all ADC outputs while preserving the relative ordering between similarity levels; although such chips may fail fixed pass/fail criteria, they may still remain functionally correct after calibration. In contrast, local variations may create overlap between adjacent similarities distributions, making reliable sensing significantly more difficult [16], [69].

In addition to process variations, ReCAM arrays are vulnerable to both hard and soft faults. Hard faults originate from permanent physical defects, whereas soft faults arise dynamically due to temperature fluctuations, aging, resistance drift, or sensing noise [15], [64]. Furthermore, some chips may appear correct at manufacturing time but fail later due to drift, environmental stress, or gradual sensing-margin degradation, making runtime monitoring an essential component of quality assurance. Consequently, reliable ReCAM operation requires distinguishing among permanent defects, recoverable variations, and runtime degradations.

These characteristics make conventional go/no-go testing insufficient for analog ReCAM architectures. Discarding all chips that fail fixed thresholds may unnecessarily reduce

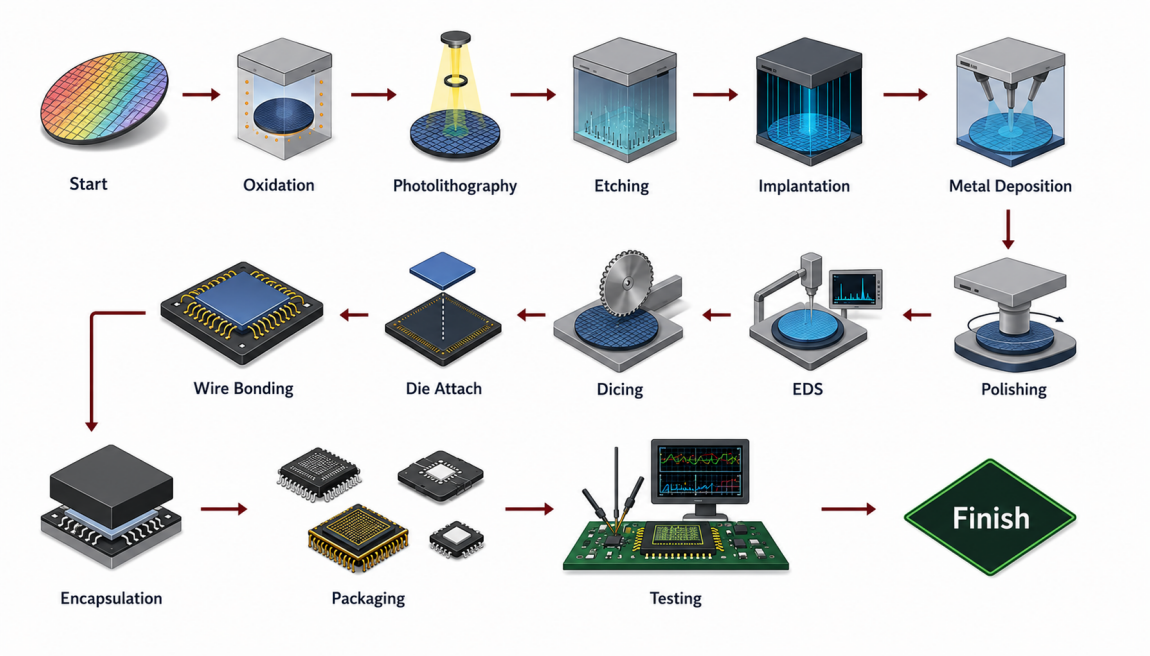


Figure 2.6.: Chip development cycle highlighting manufacturing, testing, and validation stages

yield, while insufficient testing may allow unreliable chips to escape detection. Therefore, quality assurance in ReCAM requires a comprehensive framework that combines manufacturing testing, variation-aware calibration, fault diagnosis, robust sensing methodologies, and runtime monitoring — collectively addressing the challenges that motivate the contributions of this dissertation [15], [16].

2.7. Summary

This chapter presented a comprehensive background on emerging NVM technologies, ReRAM device physics, the CiM paradigm, CAM and ReCAM architectures, and analog sensing mechanisms. In addition, the fundamental reliability and quality-assurance challenges associated with analog similarity computation, including device variability, fabrication defects, circuit non-idealities, process variations, and runtime degradation, were discussed.

These challenges demonstrate that reliable ReCAM operation cannot be guaranteed using conventional memory testing alone, as traditional digital test methods are insufficient to capture analog fault manifestations and variation-induced errors. Instead, ensuring dependable and scalable analog similarity computation requires comprehensive methodologies that jointly address testing, calibration, diagnosis, robust sensing, and runtime monitoring. The following chapters address these challenges through the contributions developed in this dissertation (**C1–C7**), each targeting a specific aspect of ReCAM quality assurance.

Part II.

Contributions

3. Reliability Analysis and Mitigation for Analog NVM-CiM: Foundation for ReCAM Reliability

This chapter provides the system-level foundation for the ReCAM-focused reliability framework developed in the remainder of the dissertation. While the analysis is presented for analog NVM-CiM more generally, the underlying reliability mechanism, namely decision failure caused by distorted analog sensing signatures, directly applies to ReCAM architectures. In ReCAM, the output voltage encodes the similarity between the stored and input patterns. Any variation, parasitic effect, or sensing inaccuracy that shifts this analog response can therefore result in incorrect similarity decisions. For this reason, the decision-failure framework introduced in this chapter serves as the starting point for the ReCAM-specific testing, diagnosis, calibration, sensing, and runtime-monitoring methodologies developed in Chapters 4–9.

Despite the promising advantages of analog CiM architectures, their reliability remains a major challenge due to the strong sensitivity of analog computation to device- and circuit-level non-idealities. Unlike conventional digital systems, where logic levels provide high noise margins, analog CiM operations are directly influenced by variations in device resistance, sensing circuitry, and interconnect behavior. As a result, process variation, resistance drift, sensing offsets, and crossbar parasitics can significantly degrade computation accuracy and lead to incorrect output decisions.

The reliability challenge of analog NVM-CiM extends across multiple abstraction layers, from the underlying memory technology and circuit implementation to architecture-level execution and application-level accuracy. Therefore, analyzing these variations requires a holistic cross-layer approach rather than isolated device- or circuit-level investigations. The interaction between technology-level imperfections and application behavior plays a critical role in determining the overall robustness of analog CiM architectures.

This chapter focuses on the reliability analysis of analog NVM-CiM and investigates how low-level imperfections propagate through the computing stack and impact application correctness. Through this comprehensive reliability investigation, this chapter provides insights into the relationship between technology-level non-idealities and system-level behavior and identifies effective mitigation strategies for reliable and scalable analog CiM architectures.

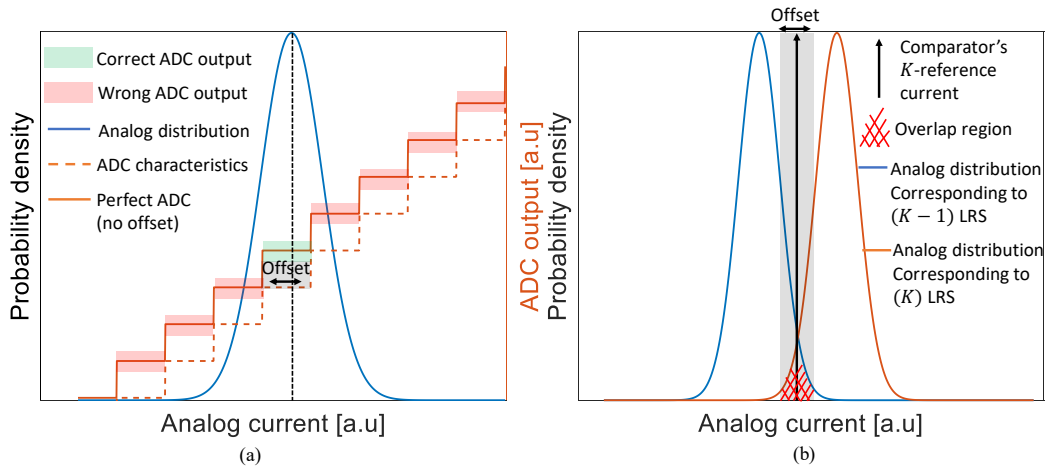


Figure 3.1.: The concept of decision failure in (a) MAC operation with ADC utilization and (b) scouting Boolean operation with comparator, the offset issue is shown in gray

3.1. Motivational Example

The ADC and comparator modules are crucial for the NVM-CiM. However, they are not perfect and show an offset issue. Figure 3.1 (a) and (b) show the offset issue for ADC and comparator, respectively. The offset is a boundary beyond which the ADC or comparator cannot operate reliably. 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 exhibits resistive and capacitive (RC) parasitic effects. Especially in more advanced technology nodes, reducing the cross-sectional area increases the parasitic resistance [70]. In the NVM-CiM, this increase in the resistive parasitic component interferes with the effective resistance of the NVM devices and impairs sensing of the analog signal at the crossbar output.

The aforementioned technology- and circuit-level factors can ultimately cause erroneous sensing of NVM states that participate in the computation, resulting in an incorrect output from the NVM-CiM module, known as *decision failure*. Assume that the expected correct output value of a reliable NVM-CiM operation with I_x as the input set is O_x. Affected by the decision failure, by having I_x, with the probability of $P_{1..N_T}$, the output value can be sensed incorrectly as $\tilde{O}_x^{1..N_T}$. For each input state, we model this hardware-level error in the form of an *expected value* as formulated in Eq. 3.1, in which N_T represents the total number of possible output states.

$$\mathbb{E}[E_{HW}^{DF}] = \sum_{i=1}^{N_T} P_i \cdot |\tilde{O}_x^i - O_x| \tag{3.1}$$

As shown in Figure 3.1 (a), in the case of MAC operation by utilizing the ADC, for one input state, the expected distance between the correct ADC output and the observed incorrect output is the $\mathbb{E}[E_{HW}^{DF}]$. Whereas in the case of scouting Boolean operation by utilizing the comparator (Figure 3.1 (b)), for neighboring input states, the area of the overlap region indicates the $\mathbb{E}[E_{HW}^{DF}]$.

3.2. Problem Statement

The reliability of analog NVM-CiM architectures is strongly affected by device- and circuit-level non-idealities that distort the analog computation results and lead to incorrect output decisions. Unlike conventional digital systems, where computation errors can often be isolated and corrected at a single abstraction layer, reliability challenges in analog CiM propagate across multiple design stacks, from the memory technology and sensing circuitry to the architecture and application levels.

sAt the technology level, process variation, resistance drift, and temperature sensitivity introduce uncertainty in the programmed resistive states of NVMs. Consequently, the analog current generated during computation deviates from its ideal value. At the circuit level, additional imperfections, including sensing offset, interconnect parasitics, crossbar dimensions, and operand activation patterns, further distort the generated analog signals and impair reliable sensing. These combined effects can result in *decision failure*, where the sensed output of the CiM operation differs from the expected correct result.

The impact of these reliability challenges becomes more critical in large-scale analog CiM architectures, where the accumulation of variation and parasitic effects increases the overlap between neighboring computation states and reduces sensing margins. Furthermore, errors generated at the hardware level can propagate through the architecture stack and eventually affect application-level correctness and inference accuracy. However, existing reliability studies mainly focus on isolated abstraction levels, such as device modeling, circuit optimization, or application-aware training, without providing a comprehensive end-to-end investigation of error generation and propagation.

Therefore, a systematic cross-layer reliability analysis framework is required to accurately model the origin of analog CiM errors, analyze their propagation across different abstraction levels, and evaluate their impact on real-world applications. In addition, effective mitigation techniques are needed at both the hardware and application levels to reduce decision failure and improve the robustness of analog NVM-CiM.

3.3. Contributions

The main contributions of this chapter can be summarized as follows:

- A comprehensive cross-layer reliability analysis framework for analog NVM-CiM architectures are presented, covering the technology, circuit, architecture, and application levels.
- The impact of technology-level non-idealities, including process variation, resistance variability, and device imperfections on the reliability of analog CiM operations are modeled and analyzed.
- The influence of circuit-level imperfections, such as sensing offsets, interconnect parasitics, crossbar size, and operand activation patterns, on analog computation accuracy and decision failure is systematically investigated.

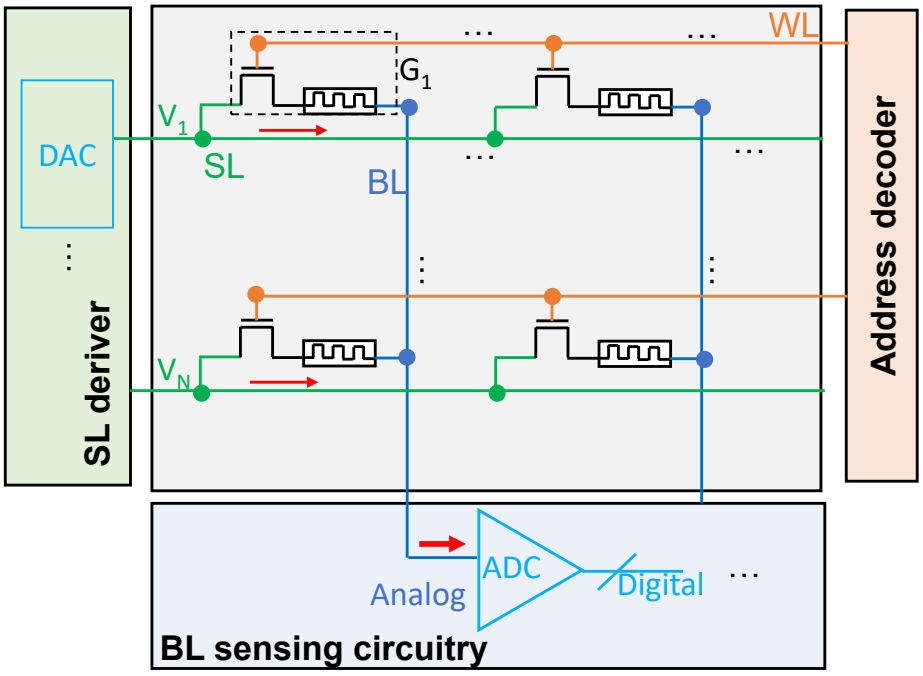


Figure 3.2.: Analog NVM-CiM in the crossbar structure

- Hardware-level errors are statistically modeled, and their propagation across higher abstraction layers is analyzed using a CiM-enabled full-system architecture simulation framework.
- Fault injection using real-world workloads and applications is performed to quantitatively evaluate the error-masking capability and robustness of analog NVM-CiM.
- Hardware- and application-level mitigation techniques are explored and evaluated to reduce decision failure and improve the overall reliability of analog CiM architectures.

3.4. Analog NVM-CiM

The multiply-accumulate (MAC) operation can be performed in a crossbar structure of the NVM and in an analog way. A MAC operation has a general form of $[V_1..V_N] \times [G1..G_N]^T$. As shown in Figure 3.2, the first vector of operands ($[V_1..V_N]$) can be encoded as the voltage and applied to the source-line (SL) of the crossbar structure. Also, the second vector ($[G1..G_N]$) can be encoded as the conductance of the NVM devices. Governing by *Ohm's law*, the current passing through each branch of the NVM cell is $V \times G$, and according to *Kirchhoff's current law (KCL)*, the current passing through the bit-line (BL) is the summation of the current in all branches, i.e., it is the result of the MAC operation in an analog way. The result of the MAC operation can be generated in a single shot, enhancing performance and energy efficiency compared to digital computation.

The analog NVM-CiM module is typically part of a digital computer, so to maintain communication, a Digital-to-Analog Converter (DAC) at its input and an ADC at its output

are required. In Figure 3.2, the DAC module is loading the SLs ($V_1..V_N$) and the ADC is converting the analog current into a digital correspondence. Depending on the application, the ADC can be either a full-fledged multi-bit ADC or a lower-cost one-bit comparator.

In addition to MAC, the execution of Boolean kernels can also be accelerated using NVM-CiM. *Scouting logic* is a way to perform the NVM-CiM for Boolean logic operations in the form of *K out of N* [71]. In this concept, the first step is to activate multiple rows by selecting their word-line (WL)s; then a voltage is applied to the NVM crossbar, and the BL current at the output is compared with a *K-reference current*. If less than K devices are in LRS, the output of the comparison is binary '0', otherwise, it is binary '1'.

During the NVM-CiM for scouting Boolean operation, using a one-bit comparator is sufficient, while, based on the required precision, during the NVM-CiM for MAC, utilization of the (multi-bit) ADC may be required.

3.5. 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 $\mathbb{E}[E_{HW}^{DF}]$ by accounting for multiple technology- and circuit-level parameters. At the architectural level, we extend the widely used gem5 infrastructure by adding the analog NVM-CiM to enable end-to-end,

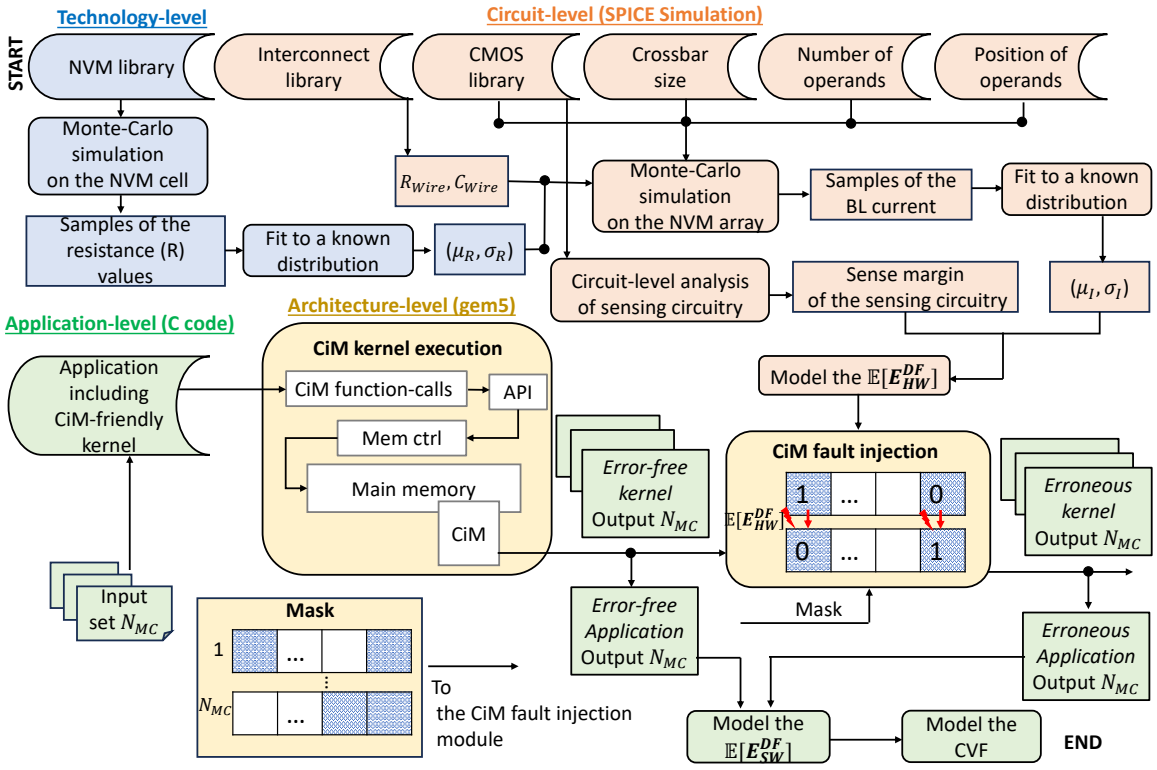


Figure 3.3.: The end-to-end flow for reliability analysis during NVM-CiM, μ and σ show the mean and standard deviation of a statistical distribution, N_{MC} shows the number of Monte Carlo simulations, $\mathbb{E}[E_{HW}^{DF}]$ and $\mathbb{E}[E_{SW}^{DF}]$ show the hardware- and software-level expected errors, respectively

full-system 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 3.3 shows our end-to-end reliability analysis flow, and its details will be discussed in the following subsections.

3.5.1. 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 multiple technology- and circuit-level parameters. At the technology level, the effect of the process variation on the NVM devices is dominant, which results in their resistance distribution. At the circuit level, the number of activated rows, the ADC/comparator offset, 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.

3.5.1.1. NVM technology

Generation of the CiM output in the periphery circuitry only includes sensing of the current [72]. 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 (μ_{LRS}, σ_{LRS}), (μ_{HRS}, σ_{HRS}) can be obtained.

3.5.1.2. 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.

3.5.1.3. Offset of the ADC/comparator

As discussed in Section 3.1, 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 to sense the cumulative resistance of 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 $\mathbb{E}[E_{HW}^{DF}]$.

3.5.1.4. 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 [70].

3.5.1.5. Crossbar size

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

3.5.1.6. 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 the two input states and varies across the different hardware configurations discussed. Therefore, the error trend is not similar between these two CiM kernels.

3.5.2. 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 paper, we develop our framework based on CiM-enabled gem5 as detailed in [73]. 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 writing to 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 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 3.3, fault injection to account for decision failures 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.

3.5.2.1. 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 $\mathbb{E}[E_{HW}^{DF}]$.

3.5.2.2. 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 against decision failures, we can run 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 Eq. 3.2.

$$\mathbb{E}[E_{SW}^{DF}] = \frac{1}{N_{MC}} \times \sum_{i=1}^{N_{MC}} \frac{|\text{error-free}_i - \text{erroneous}_i|}{\text{error-free}_i} \tag{3.2}$$

To measure the masking capability of the CiM-friendly application, we define the *CiM vulnerability factor (CVF)* as in Eq. 3.3, in an analogous way to how the architectural vulnerability factor (AVF) was defined [74]. CVF is simply the probability that the error in the output of the CiM module causes a noticeable error (according to Eq. 3.2) in the output of the program. A smaller CVF indicates a stronger masking capability for an application.

$$CVF = \frac{\mathbb{E}[E_{SW}^{DF}]}{\mathbb{E}[E_{HW}^{DF}]} \tag{3.3}$$

3.5.2.3. 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 3.5.1), only the number of 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 technologies, along with the crossbar size, are predetermined during the circuit-level design phase and modeled in $\mathbb{E}[E_{HW}^{DF}]$.

3.6. Results and Discussion

3.6.1. Technology-level analysis

For our end-to-end simulations, we use the tools and parameters outlined in Table 3.1. For the NVM technology, we use the measurement-verified Verilog-A models of two representative technologies: STT-MRAM and ReRAM. STT-MRAM provides relatively small $\Delta\mu_R (= \mu_{\text{HRS}} - \mu_{\text{LRS}})$ and $\sigma_{\text{LRS, HRS}}$ 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, lie between the STT-MRAM and ReRAM [78].

Table 3.1.: End-to-end simulation setup and tools.

Technology-Level Parameters	
MTJ Model [75]	Radius = 20 nm Barrier Material = MgO RA = 7.5 $\Omega\mu m^2$ TMR = 150% HRS, LRS = 11.56 kΩ, 5.967 kΩ
ReRAM Model [23], [76]	JART VCM Read Variability Filament Radius = 45 nm Disc Region Length = 0.6 nm HRS, LRS = 105.51 kΩ, 1.975 kΩ
Circuit-Level Parameters	
Simulation Tool	Cadence Virtuoso
CMOS Technology	GlobalFoundries 22FDX
Interconnect Technology [77]	22 nm and 5 nm technology nodes
Architectural-Level Parameters	
Instruction Set Architecture (ISA)	x86
CPU Model	In-order processor Four-stage buffered pipeline Frequency = 1 GHz
Main Memory Size	256 MB
Application-Level Parameters	
MAC Benchmark	2mm (PolyBench)
Boolean Benchmark	Database Query

3.6.2. Circuit-level analysis

We perform the circuit-level simulation using SPICE and parameters outlined in Table 3.1 to quantitatively account for the impact of the following circuit-level parameters on $\mathbb{E}[E_{HW}^{DF}]$.

3.6.2.1. 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.

3.6.2.2. MAC operation

For the ADC during the MAC operation, we select a 9-bit ADC with a voltage source of 1 Volt from [79]. 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

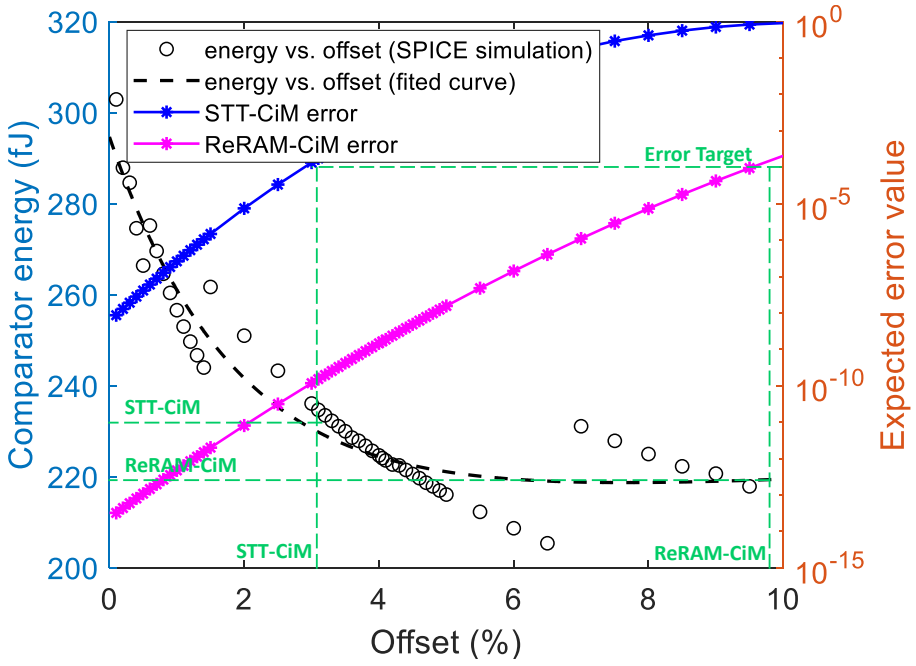


Figure 3.4.: The trade-off between energy, offset, and expected error values, for scouting Boolean NVM-CiM with two operands, crossbar size: 128×128, interconnect node: 5 nm, to achieve a comparable $\mathbb{E}[E_{HW}^{DF}]$ with an equal number of operands, the position of operands for STT-CiM and ReRAM-CiM are closest to and farthest from the sensing circuitry, respectively

at the cost of more energy consumption. The offset-energy trade-off of the selected ADC from [79] is already optimized.

3.6.2.3. 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 3.4 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 3.4, 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 comparator's energy consumption can also be reduced for ReRAM. According to Figure 3.4, selecting the offset as $\sim 3\%$ can provide an optimum offset-energy trade-off.

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

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

Figure 3.5 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 3.5.1). Because the ADC with fixed quantization levels is used for the MAC operation and due to the smaller σ_R values in

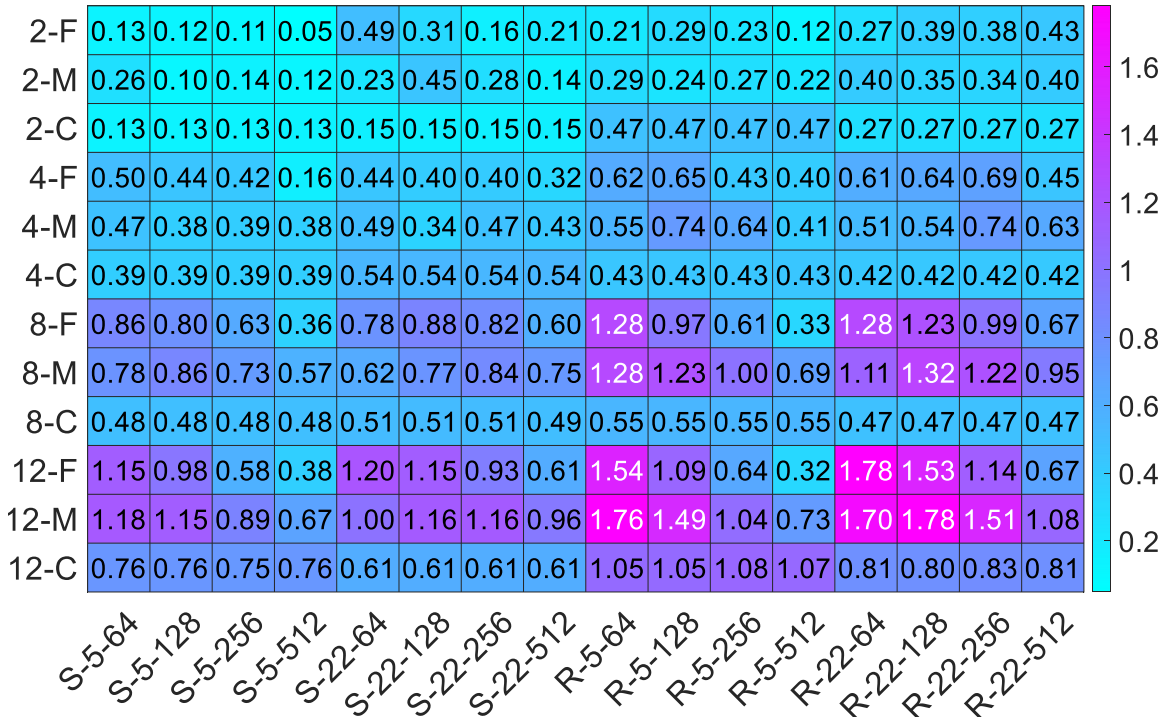


Figure 3.5.: Heat map of $\mathbb{E}[E_{HW}^{DF}]$ under different hardware parameters during the MAC operation, X labels: S/R, STT-MRAM, ReRAM. 5/22, interconnect node. 64/128/256/512, crossbar size. Y labels: 2/4/8/12, number of operands. F/M/C, Farthest, Middle, and Closest positions

the STT-MRAM, this technology yields a lower $\mathbb{E}[E_{HW}^{DF}]$ than ReRAM. Besides, the higher number of operands exacerbates the $\mathbb{E}[E_{HW}^{DF}]$.

Moreover, Figure 3.6 shows the *worst-case* $\mathbb{E}[E_{HW}^{DF}]$ for the scouting Boolean operation (offset: 3%). For the Boolean kernel, the reference is chosen from one of two distinct distributions. Hence, any parameter that increases the sensing margin decreases the $\mathbb{E}[E_{HW}^{DF}]$. Therefore, selecting an NVM technology with a higher $\Delta\mu_R$, and decreasing the number of operands can decrease the $\mathbb{E}[E_{HW}^{DF}]$. Besides, using interconnect technology with lower parasitic resistance, decreasing the crossbar size, and positioning the activated rows closest to the sensing circuitry are other means of decreasing $\mathbb{E}[E_{HW}^{DF}]$.

3.6.3. 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, which performs matrix-matrix multiplication, and its CiM-friendly kernel is 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.

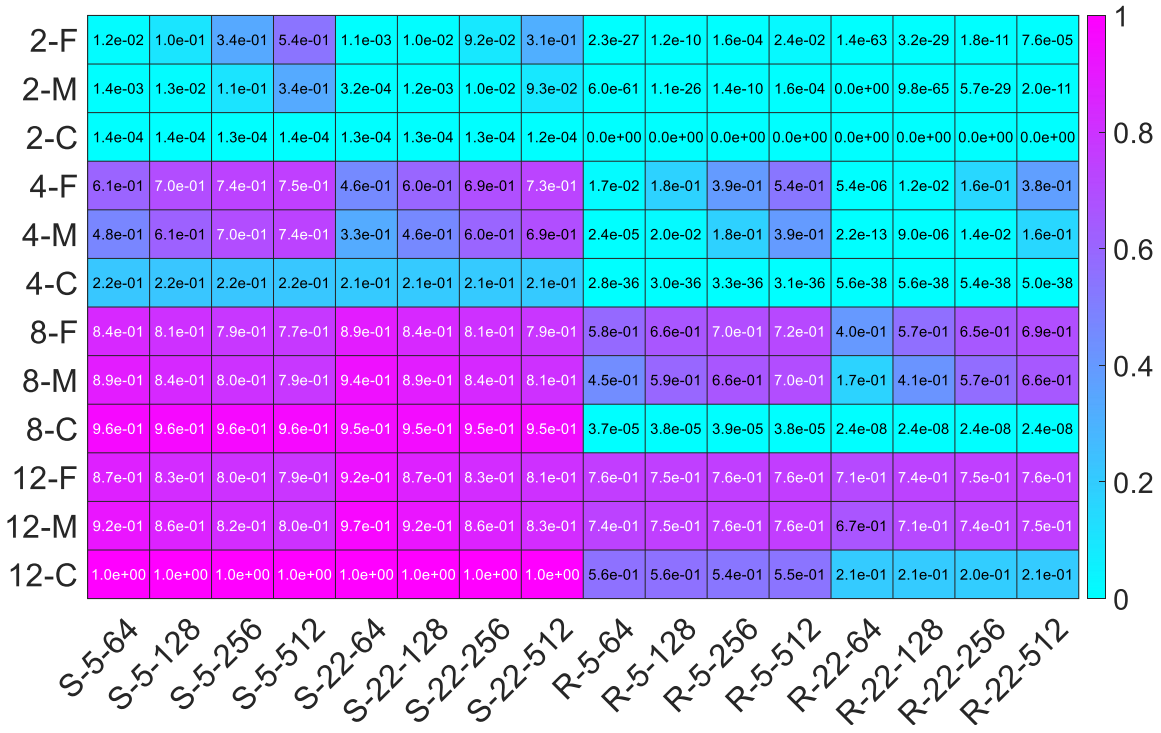


Figure 3.6.: Heat map of $\mathbb{E}[E_{HW}^{DF}]$ under different hardware parameters during NVM-CiM for scouting Boolean operation, the offset of the comparator is 3%, for X and Y labels, refer to the caption of Figure 3.5

Tab. 3.2 shows the $\mathbb{E}[E_{HW}^{DF}]$, CVF, $\mathbb{E}[E_{SW}^{DF}]$, and normalized execution time during the execution of the *2mm* benchmark. As the square-matrix size increases, the *2mm* benchmark benefits more significantly from 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, Tab. 3.3 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 3.6, increasing the number of operands beyond eight is not feasible due to the extremely high $\mathbb{E}[E_{HW}^{DF}]$.

In Tab. 3.2 and 3.3, we consider both the STT-MRAM and the ReRAM technologies; however, only with an interconnect node of 22 nm and a crossbar size of 128×128. As discussed in Section 3.5.2.3, the CVF is only impacted by the number of the operands, and the effect of other hardware parameters is modeled in $\mathbb{E}[E_{HW}^{DF}]$ (1, 2, 4-6 in Section 3.5.1). The $\mathbb{E}[E_{SW}^{DF}]$ for any hardware configuration presented in Figure 3.5 and 3.6 can be obtained by multiplying its $\mathbb{E}[E_{HW}^{DF}]$ into the respective CVF value from Table 3.2 or 3.3.

Table 3.2.: 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 (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	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 3.3.: 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

3.6.4. 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 Tab. 3.3). For the MAC operation, though the increase in the number of operands exacerbates the $\mathbb{E}[E_{HW}^{DF}]$, as shown in Tab. 3.2 for the *2mm* benchmark, the decreasing trend

Table 3.4.: 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

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-CiM, where $\mathbb{E}[E_{SW}^{DF}]$ can be significant even with more than two operands, implementing a hierarchical approach to scouting Boolean operations is promising. For an 8-input Boolean STT-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*.

Tab. 3.4 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 Tab. 3.4, using the hierarchical mapping configurations slightly increases the CVF. However, as the $\mathbb{E}[E_{HW}^{DF}]$ in the case of STT-CiM-2 and STT-CiM-4 is considerably less than that of STT-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 Tab. 3.4, hierarchical execution of Boolean operations and necessary adjustments in the application using the corresponding CiM kernels incur higher execution time. This can be explained by the increase in the number of CiM instructions when using alternative hierarchical configurations.

3.7. Summary

This chapter presented a comprehensive cross-layer reliability analysis framework for analog NVM-based CiM systems, spanning from the technology level to the application level. At the technology level, the effects of NVM device variability and process variation were investigated. At the circuit level, hardware-level decision failure was statistically modeled for both MAC- and scouting Boolean-based CiM operations, accounting for sensing circuitry requirements, interconnect parasitics, crossbar dimensions, operand activation patterns, and operand positions within the array. The modeled hardware-level errors were then propagated to higher abstraction levels through fault injection into a CiM-enabled full-system simulation framework, enabling quantitative evaluation of the error-masking capability and robustness of real-world applications.

Furthermore, this chapter explored both hardware- and application-level mitigation techniques to reduce decision failure and improve the overall reliability of analog NVM-CiM architectures. The presented analysis demonstrates the importance of a holistic technology-to-application reliability investigation and provides valuable guidelines for designing more robust and scalable analog CiM systems.

Most importantly, the results reveal that reliability in analog NVM-CiM is not limited to data retention or storage correctness, but is fundamentally linked to the correct interpretation of analog computation signatures. Sensing-induced decision failures, therefore represent a primary source of computational-quality degradation in analog CiM architectures. This observation is particularly critical for ReCAM, where the match-line response encodes the similarity metric and small sensing deviations can change the final search result.

Motivated by these insights, the remainder of this dissertation focuses specifically on ReCAM architectures and addresses the main sources of decision failure across the man-

ufacturing and runtime lifetime of the array. Chapter 4 first develops a defect-oriented fault model and March-like test methodology for ReCAM arrays. Chapters 5 and 6 extend the analysis to fault diagnosis and process-variation-aware testing. Chapter 7 introduces post-manufacturing calibration to recover chips affected by global variations, while Chapters 8 and 9 investigate alternative sensing and runtime-monitoring techniques to improve robust operation after deployment. Together, these contributions (**C2–C7**) establish a comprehensive quality-assurance framework for reliable similarity computation in ReCAM.

4. Fault Modeling and Testing of ReRAM-based CAM Array

Ensuring the quality of ReCAM is a critical challenge due to the analog nature of similarity-search operations and the intrinsic non-idealities of both ReRAM and CMOS technologies, which are absent in conventional CAMs. ReCAM arrays exhibit unique fault behaviors arising from resistive-switching variability and analog-sensing dependencies. These challenges require dedicated fault-modeling and testing methodologies that accurately capture ReCAM-specific failure mechanisms. Therefore, this chapter presents a comprehensive framework for ReCAM fault modeling and introduces a March-like testing methodology tailored for analog similarity-search architectures.

4.1. Motivational Example

Figure 4.1 illustrates the similarity-search behavior of a ReCAM array under ideal conditions. Unlike conventional SRAM-CAM, which relies on binary comparison signals, ReCAM performs analog in-memory similarity computation using the ML discharge characteristics. During a search, the ML is precharged and then discharged through a resistive path determined by the relationship between the stored pattern and the input query. A match forces the current through HRS, resulting in a slow discharge and a relatively high ML voltage at the sensing time, whereas a mismatch causes discharge through LRS, producing a faster voltage drop. The residual ML voltage, therefore, encodes the degree of similarity between the two patterns.

In an ideal ReCAM array, different similarity levels produce distinct, separable ML voltages that can be digitized accurately by the sensing circuitry. However, Figure 4.1 highlights that this separation is highly sensitive to defects and analog perturbations. A resistive defect inside even a single ReCAM cell can alter the discharge rate of the entire ML, shifting the resulting ML voltage toward a neighboring similarity level. Although the cell may still behave correctly during digital read/write operations, its analog discharge characteristics become distorted. Such analog deviations are not detected by conventional digital memory tests and may cause a 1-match pattern to be incorrectly interpreted as a 2-match pattern after ADC quantization. This type of error becomes increasingly likely in large ReCAM arrays because multiple defective or highly variable cells accumulate along the shared ML, causing distribution overlap between adjacent similarity levels.

These observations demonstrate that ReCAM failures often arise not from digital read/write errors but from subtle, undetected distortions in analog similarity computation. As a result, ensuring reliable ReCAM operation requires dedicated fault modeling

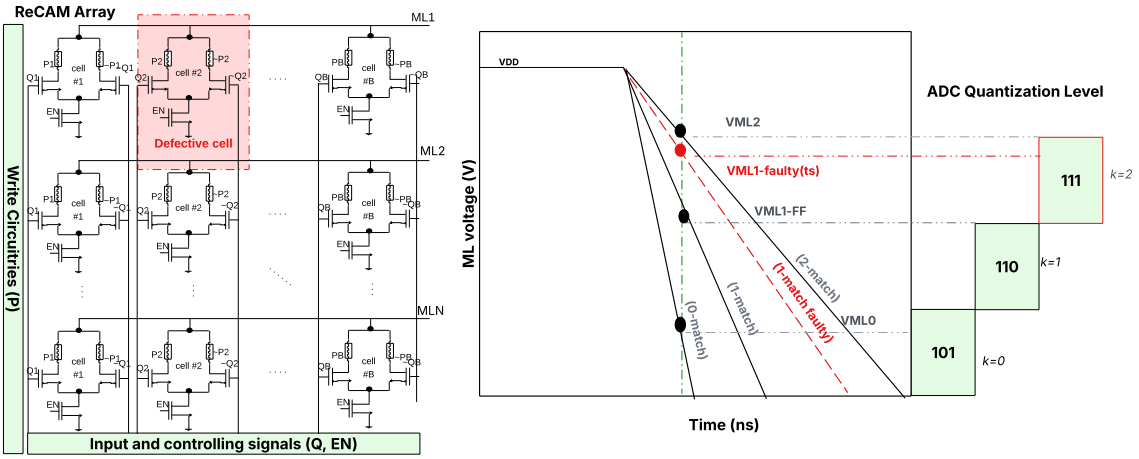


Figure 4.1.: Illustration of conventional voltage-domain sensing in ReRAM-based CAM (ReCAM) arrays under ideal and defect-aware conditions. In fault-free operation, ML discharge produces well-separated voltage levels corresponding to different similarity levels (k-match). In the presence of defects, ML discharge shifts and sensing margins shrink, causing ADC quantization errors and incorrect similarity detection.

and test methodologies capable of capturing both digital and analog defect behaviors that traditional SRAM-oriented tests cannot detect.

4.2. Problem Statement

Classical memory test methodologies—including March tests—were designed for SRAM-based memories and CAM arrays whose operations are strictly digital. These tests detect faults by verifying whether cells can reliably store and return binary values. However, these assumptions break down in ReRAM-based CAM architectures. In ReCAM, similarity evaluation is performed in the analog domain, and the final output depends on the ML discharge behavior, device resistance values, and the accuracy of the peripheral sensing circuitry. Variations in ReRAM resistance, parasitic interconnect elements, CMOS transistor defects, and sensing-circuit non-idealities all influence the ML discharge trajectory. Many of these variations do not manifest as digital read/write failures, but instead perturb the analog similarity signature.

Furthermore, integrating ReRAM devices with CMOS introduces additional defect mechanisms absent in SRAM-CAM. ReRAM-specific defects such as stuck-at-HRS/LRS, deep-HRS/LRS, and undefined intermediate states produce analog distortions that traditional digital fault models cannot capture. Similarly, access-transistor failures and resistive-open defects at the ML interface alter the equivalent discharge resistance, causing ML voltage shifts that escape detection in digital testing. As a result, traditional March tests provide inadequate fault coverage because they cannot observe analog deviations that corrupt similarity detection without violating digital correctness.

Therefore, new testing methodologies are required to address three fundamental challenges:

- accurately modeling the defects unique to ReCAM architectures,

- analyzing the analog impact of these defects on match-line discharge and similarity computation, and
- developing March-like test sequences capable of activating, observing, and detecting ReCAM-specific analog fault behaviors.

Addressing these challenges is essential for ensuring reliable similarity-search functionality in ReRAM-based CAM architectures.

4.3. Contributions

This chapter introduces a comprehensive testing methodology tailored specifically to ReRAM-based CAM arrays. The main contributions are summarized as follows:

- A systematic fault-modeling framework for ReCAM architectures is developed to capture the electrical impact of single physical defects at the device, cell, and interconnect levels. The framework includes ReRAM resistance faults, CMOS access-transistor faults, and ML connectivity defects.
- A detailed defect-impact analysis is performed to characterize how ReCAM-specific defects distort ML discharge behavior and analog similarity responses. Using circuit-level simulations and defect-injection experiments, this work identifies analog fault behaviors that traditional SRAM-based testing methodologies cannot detect.
- A dedicated March-like test algorithm is proposed for ReCAM architectures. The algorithm extends conventional March-test principles by incorporating analog observation conditions, ML discharge monitoring, and activation sequences designed to expose ReCAM-specific fault behaviors.
- The proposed testing methodology is evaluated with respect to fault coverage, scalability, and extensibility. The results demonstrate complete coverage of the considered ReCAM fault models while remaining applicable to large-scale arrays and future ReCAM designs.

4.4. ReCAM Defects

Although ReCAM architectures provide compact and energy-efficient similarity-search functionality, they remain vulnerable to several physical defect mechanisms that affect both their digital and analog behavior. Due to the analog nature of ML discharge and similarity sensing, even small electrical deviations can distort the final similarity result and reduce sensing reliability. In this work, the considered defects are broadly categorized into two groups based on their location and impact on the ReCAM architecture: *intra-cell defects* and *inter-cell defects*. Figure 4.2 illustrates representative defect locations considered in this work, while Table 4.1 summarizes the primary fault mechanisms and their corresponding effects on ReCAM operation.

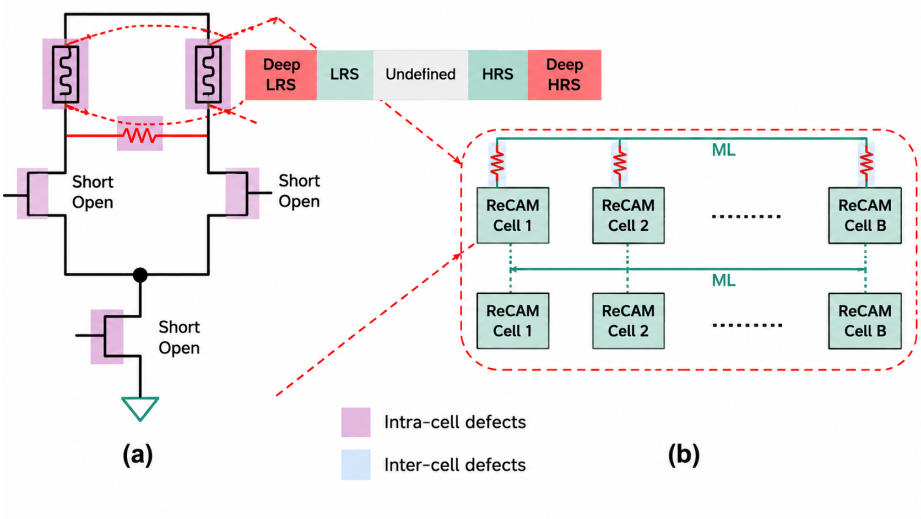


Figure 4.2.: Schematic of (a) intra-cell defect locations and (b) inter-cell defect locations within a ReCAM cell and ReCAM array, respectively.

4.4.1. Intra-Cell Defects

Intra-cell defects originate within individual ReCAM cells and directly affect the local storage and comparison behavior. These defects may arise from imperfections in the ReRAM devices, access transistors, or internal cell connections.

4.4.1.1. ReRAM Device Defects

ReRAM devices are susceptible to fabrication imperfections, filament instability, and switching degradation [55], [56]. One common failure mechanism is the *stuck-at-LRS* or *stuck-at-HRS* fault, where the device becomes permanently trapped in the LRS or HRS, respectively. In addition, abnormal filament formation may produce excessively low or high resistance values, commonly referred to as *deep-LRS* and *deep-HRS* states. Furthermore, irregular filament geometry or incomplete switching may produce intermediate resistance states between HRS and LRS that are undefined [39], leading to ambiguous sensing behavior during similarity computation.

4.4.1.2. Transistor Defects

Access transistors may also experience defects such as *stuck-on* and *stuck-open* faults [80]. A stuck-on transistor continuously conducts current regardless of the gate voltage, potentially creating unintended ML discharge paths. Conversely, a stuck-open transistor blocks the intended current path and prevents the corresponding cell from participating correctly in the similarity-search operation.

4.4.1.3. Bridging Defects Within the Cell

Bridging defects occur when unintended resistive connections form between internal nodes of the ReCAM cell. For example, a resistive short between the drains of the left and

Table 4.1.: Fault models considered in the ReCAM architecture, including defects originating from CMOS transistors, ReRAM devices, and interconnect structures

Fault Origin	ID	Fault Description
CMOS Transistor Defects		
CMOS	F1	T_L stuck-on
CMOS	F2	T_L stuck-open
CMOS	F3	T_R stuck-on
CMOS	F4	T_R stuck-open
CMOS	F5	T_{EN} stuck-on
CMOS	F6	T_{EN} stuck-open
ReRAM Device Defects		
ReRAM	F7	R_L stuck-at-LRS
ReRAM	F8	R_L stuck-at-HRS
ReRAM	F9	R_R stuck-at-LRS
ReRAM	F10	R_R stuck-at-HRS
ReRAM	F11	R_L deep-LRS
ReRAM	F12	R_L deep-HRS
ReRAM	F13	R_R deep-LRS
ReRAM	F14	R_R deep-HRS
ReRAM	F15	R_L undefined region
ReRAM	F16	R_R undefined region
Interconnect Defects		
Interconnect	F17	Resistive open between the ML and the ReCAM cell
Interconnect	F18	Resistive short between the drains of T_R and T_L

right access transistors can distort the discharge current and alter the analog ML behavior. Such defects may significantly affect the sensed similarity level even when the stored data remains logically correct.

4.4.2. Inter-Cell Defects

Inter-cell defects primarily affect the shared ML infrastructure and the connectivity between neighboring ReCAM cells. Among these defects, resistive open defects caused by partially filled or unfilled vias are particularly critical. These defects introduce a large series resistance between a cell and the ML or may completely disconnect the cell from the shared ML structure.

As a result, one or more cells may become electrically isolated and fail to contribute correctly to the similarity computation process. Unlike conventional digital memories, these defects not only affect logical functionality but also modify the analog ML discharge

characteristics, leading to inaccurate similarity levels and unpredictable sensing behavior at the array level.

Although other non-idealities such as transient faults, minor process variations, and small parameter drifts may also occur in practice, they are not explicitly modeled in this work. Their impact is typically mitigated by the shared ADC structure and by system-level calibration techniques that compensate for small analog deviations.

4.5. March Test Algorithm

Manufacturing testing ensures correct functionality and sufficient yield in memory arrays. March algorithms are widely used in industry for their structured, deterministic sequences of write and read operations that uncover digital fault types such as stuck-at, transition, and coupling faults [81]. These algorithms apply a defined sequence of operations to every cell in ascending and descending address orders, ensuring high coverage of digital failure mechanisms.

Previous work in ReRAM testing has proposed variations of March algorithms to detect ReRAM device faults [82], [83], [84], [85], [86], [87], [88], [89], [90]. Likewise, extensive studies on CAM testing focus on detecting comparison faults, content-fault behaviors, and timing-related issues in SRAM-based CAM architectures [81], [91], [92], [93]. However, ReCAM architectures fundamentally differ from both ReRAM-only memories and SRAM-based CAMs because the correctness of a ReCAM search operation depends on analog ML discharge behavior and the subsequent sensing process.

In ReCAM, defects may not cause incorrect digital reads or writes but instead distort the evaluation of analog similarity. A defective device may still correctly store and return binary data while producing a shifted ML voltage distribution. Such deviations create similarity-level overlap, leading to misclassification even when no digital error is observed. Consequently, traditional March testing is insufficient because it neither activates nor observes analog fault manifestations.

As a result, testing ReCAM architectures requires a dedicated approach that extends the principles of March algorithms by incorporating analog-aware activation sequences, ML discharge observation, and sensing-aware measurement conditions. Developing and validating such a March-like test methodology forms the central objective of this work.

4.6. Proposed fault modeling for ReCAM

In standard memory testing, the primary operations are read and write, with data stored and retrieved at memory addresses. However, in ReCAM, the testing process differs due to the nature of CAM operations. CAM operations involve data comparison tasks instead of just reading or writing data. For ReCAM, the search inputs are binary patterns that are compared against the stored values. The observed ML voltage output results from a match or a mismatch. This output is then processed through an ADC. Given the operational differences between memory and ReCAM, the first step in generating ReCAM-specific tests is to model defects originating from Tr, ReRAM, and intra-cell via-related sources,

Table 4.2.: Technology- and circuit-level simulation parameters

Category	Parameter Details
Technology-Level Parameters	
ReRAM Model	JART VCM Read Variability [23]
Filament Radius	45 nm
Disc Region Length	0.6 nm
N_{init} (HRS, LRS)	9×10^{26} m^{-3}, 3×10^{26} m^{-3}
N_{init} (Undefined Region)	65×10^{26} m^{-3} [94]
N_{init} (Deep-HRS, Deep-LRS)	5×10^{26} m^{-3}, 1×10^{29} m^{-3} [94]
Circuit-Level Parameters	
Simulation Tool	Cadence Virtuoso
CMOS Technology	GlobalFoundries 22FDX
Supply Voltage (V_{DD})	800 mV
Temperature	27°C
ADC [95]	8-bit resolution
Interconnect Parameters per ReCAM Cell (22 nm) [47], [48]	ML: $R = 11.95\ \Omega$, $C = 16.63$ aF
ReCAM Array Size	64-bit

and then analyze their fault behaviors in ReCAM's outputs. To achieve this, we perform electrical single-defect injections into ReCAM cells.

We start our fault modeling method with a single ReCAM cell and extend it to the entire ReCAM array. For each faulty scenario, we apply binary input patterns (Q) against the stored prototypes (P) for the entire ReCAM array and then compare the resulting faulty outputs from the shared ADC (Out_{ADC}) against the Fault-Free (FF) outputs as depicted in Figure 2.4(a). These comparisons enable us to classify fault behaviors based on equivalence and dominance, thereby facilitating the grouping of similar faults. By doing that, we can determine which input combinations are needed to detect all faulty scenarios.

4.6.1. Fault modeling of single ReCAM cell (intra-cell faults)

To understand the fault behaviors and impact of the modeled defects listed in Table 4.1 on a single ReCAM cell depicted in Figure 4.2 (a), we use the simulation setup outlined in Table 9.1. The intra-cell defects, i.e., F1-F16 and F18 (see Table 4.1), are modeled in a single ReCAM cell. We apply all input combinations of P and Q listed in Table **??** to the ReCAM cell. The values of P and Q are applied in a complementary manner to the ReRAM and transistor elements, i.e., P and Q on the left side and $\bar{P}$ and $\bar{Q}$ on the right side of the ReCAM cell. Then, we can observe whether a fault manifests by comparing the faulty outputs to the FF outputs. We perform comprehensive electrical-level simulations using SPICE to analyze the impact of these defects for all fault scenarios. Then we compare the corresponding waveforms and ADC outputs with those of the FF.

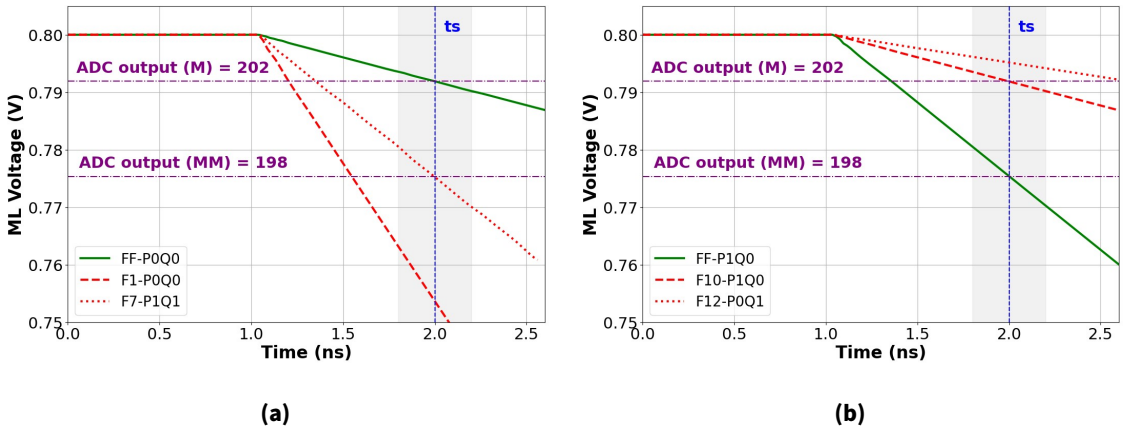


Figure 4.3.: Waveform simulations of match line (ML) voltage when (a) injecting F1 (T_L stuck-on) and F7 (R_L stuck-at-LRS) vs fault-free (FF) of two input combinations ($P = 0, Q = 0$) and ($P = 1, Q = 1$) expecting match (M) and also (b) injecting F10 (R_R stuck-at-HRS) and F12 (R_L Deep HRS) vs FF of two input combinations ($P = 1, Q = 0$) and ($P = 0, Q = 1$) expecting mismatch (MM) at sampling time (t_s)

For faults F1 and F7, the ML voltage for a match scenario in Figure 4.3 (a) deviates from the expected match and is interpreted incorrectly as a mismatch instead. Conversely, under faults, F10 and F12 in Figure 4.3 (b), the ML voltage for a mismatch scenario deviates from the expected mismatch and is incorrectly interpreted as a match. Such deviations from the expected voltage behavior directly indicate the presence of faults in the ReCAM cell. Under the F18 fault scenario, the ML voltage remains consistently low due to the low ohmic resistance of the discharge path, regardless of the input patterns, thereby uniquely identifying this fault.

4.6.2. Fault modeling of ReCAM array (inter-cell faults)

As illustrated in Figure 4.2 (b), in a ReCAM array, multiple ReCAM cells are connected to a shared ML. This section aims to extend our one-bit ReCAM fault modeling into a multi-bit ReCAM array. To ensure a realistic electrical-level simulation setup for this experiment, we model the ML's resistive-capacitive (RC) parasitic.

The first experiment in this regard is injecting the inter-cell defect (F17) into the ReCAM array. During this experiment, we consider various locations for F17 as shown in Figure 4.2 (b). Based on our observations, the presence of F17 results in an extremely slow discharge path regardless of its location. This means that the effect of inter-cell resistive open (F17) on the ReCAM output is independent of the cell's location, whether it is the first, middle, or last cell in the row. So, in the presence of F17, the Out_{ADC} has a maximum value.

Besides the injection of F17, we also inject other intra-cell defects into the ReCAM with B ReCAM cells connected to the ML, following a one-fault-injection-at-a-time fashion and through a comprehensive electrical level simulation. The bounded length of the ML (see Table 9.1) leads to similar trends between the faulty behavior of the one-bit ReCAM cell and the B-bit ReCAM array. This way, we can build our testing methodology based on the fault modeling of the one-bit ReCAM cell.

4.6.3. Fault Grouping

A detailed fault modeling analysis identifies and categorizes multiple equivalence-collapse patterns, grouping equivalent faults that exhibit the same behavior across different fault scenarios, as described in Table 4.3. Here, 'M' represents a match and 'MM' represents a mismatch. Notably, faults {F1, F9, F13, F16, F18} produce identical faulty outputs, as do faults {F2, F5, F6, F8, F12, F17}, faults {F3, F5, F7, F11, F15, F18}, and faults {F4, F6, F10, F14, F17}. We noticed that for each fault injection scenario, there is typically one faulty output across all possible input combinations, except for F5 and F6, and also F18 and F17, where there are two faulty outputs, as shown in Table 4.3. Therefore, they can be grouped, with each group representing a set of faults, as follows:

- G1 = {F1, F9, F13, F16, F18}
- G2 = {F2, F5, F6, F8, F12, F17}
- G3 = {F3, F5, F7, F11, F15, F18}
- G4 = {F4, F6, F10, F14, F17}

The full results of this analysis are summarized in Table 4.3. Additionally, as indicated in this table, selecting the appropriate input combination is crucial for fault identification. In the next section, we elaborate on the development of the test methodology for the ReCAM array.

4.7. ReCAM Test

4.7.1. Preliminaries and Notations

This paper assumes that the ReCAM cell performs two primary operations: WRITE and COMPARE. The P data is programmed into the ReCAM cells during the WRITE operation, setting them to either the LRS or HRS state. After injecting a defect into either the Transistor or ReRAM components, the outputs of the ReCAM cells are compared. The COMPARE operation compares the Out_{ADC} with the FF output when the Q search data are applied to all ReCAM cells, detecting potential faults. The following notation is used to explain the algorithm procedures.

- $w_P d$: Write the prototype data pattern d in the ReRAM as LRS or HRS state to a specific cell, and this data pattern could be either an all-0s or an all-1s pattern.
- $c_Q d$: Compare Out_{ADC} with FF output using specific data patterns d for the search query Q in all ReCAM cells. These patterns can include all-0s, all-1s, walking-0, or walking-1.

Table 4.3.: Fault behaviors of a single ReCAM cell. Red entries indicate faulty outputs. M and MM denote Match and Mismatch, respectively. The defect size ranges for F17 and F18 are also indicated.

Q	P	FF	Faulty Output																
			F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
																			> 100 kΩ	< 1 kΩ
0	LRS	M	MM	M	M	M	M	M	M	M	MM	M	M	M	MM	M	M	MM	M	MM
1	HRS	M	M	M	MM	M	MM	M	MM	M	M	M	MM	MM	M	M	MM	M	M	MM
0	HRS	MM	MM	MM	MM	M	MM	M	MM	MM	MM	M	MM	MM	MM	M	MM	MM	M	MM
1	LRS	MM	MM	M	MM	MM	M	M	MM	M	MM	MM	MM	M	MM	MM	MM	MM	M	MM

4.7.2. Generating test vectors for single cell

To achieve 100% fault coverage, as each combination of binary values for the P and Q inputs represents distinct fault scenarios affecting the output of the ReCAM operation. For instance, faults in group G1 can be detected using the test vector $(P = 0, Q = 0)$, while faults in groups G2, G3, and G4 require specific test vectors $(P = 0, Q = 1)$, $(P = 1, Q = 1)$, and $(P = 1, Q = 0)$ respectively.

Thus, grouping these faults into equivalence classes reduces the number of test vectors needed for fault detection, reducing test time and complexity. For testing the ReCAM array in Figure 2.4 (a), we use only two data values for P, either 0 (LRS) or 1 (HRS), extending them to all 0's or all 1's across all B-bits in the ReCAM array to minimize the costly write operations required for ReRAMs. Moreover, by modifying Q data to include all 1's, all 0's, walking 1's, and walking 0's, we can detect all potential fault groups in each cell and identify which cells have faults in the ReCAM array.

4.7.3. Symmetric March-like Test Algorithm

The proposed symmetric March-like test algorithm detects all the faults in the ReCAM array. The test is symmetric because the same sequence is performed when writing 0 or 1 in the ReCAM array. This test algorithm involves multiple March elements (ME) that progressively write prototype data and compare Out_{ADC} to detect and identify the faulty cell. Each sequence of these elements targets a different fault group scenario.

Thus, the sequence operations of symmetric march-like test algorithms are as follows:

$$
\begin{aligned}
&ME1 : \left(w_P\,\mathbf{0},\ c_Q\,\mathbf{0}\right);\\
&ME2 : \left(c_Q\mathbf{01\ldots1};\ c_Q\mathbf{10\ldots1};\ \ldots;\ c_Q\mathbf{11\ldots0}\right);\\
&ME3 : \left(c_Q\,\mathbf{1}\right);\\
&ME4 : \left(c_Q\mathbf{10\ldots0};\ c_Q\mathbf{01\ldots0};\ \ldots;\ c_Q\mathbf{00\ldots1}\right);\\
&ME5 : \left(w_P\,\mathbf{1},\ c_Q\,\mathbf{1}\right);\\
&ME6 : \left(c_Q\mathbf{10\ldots0};\ c_Q\mathbf{01\ldots0};\ \ldots;\ c_Q\mathbf{00\ldots1}\right);\\
&ME7 : \left(c_Q\,\mathbf{0}\right);\\
&ME8 : \left(c_Q\mathbf{01\ldots1};\ c_Q\mathbf{10\ldots1};\ \ldots;\ c_Q\mathbf{11\ldots0}\right).
\end{aligned}
$$

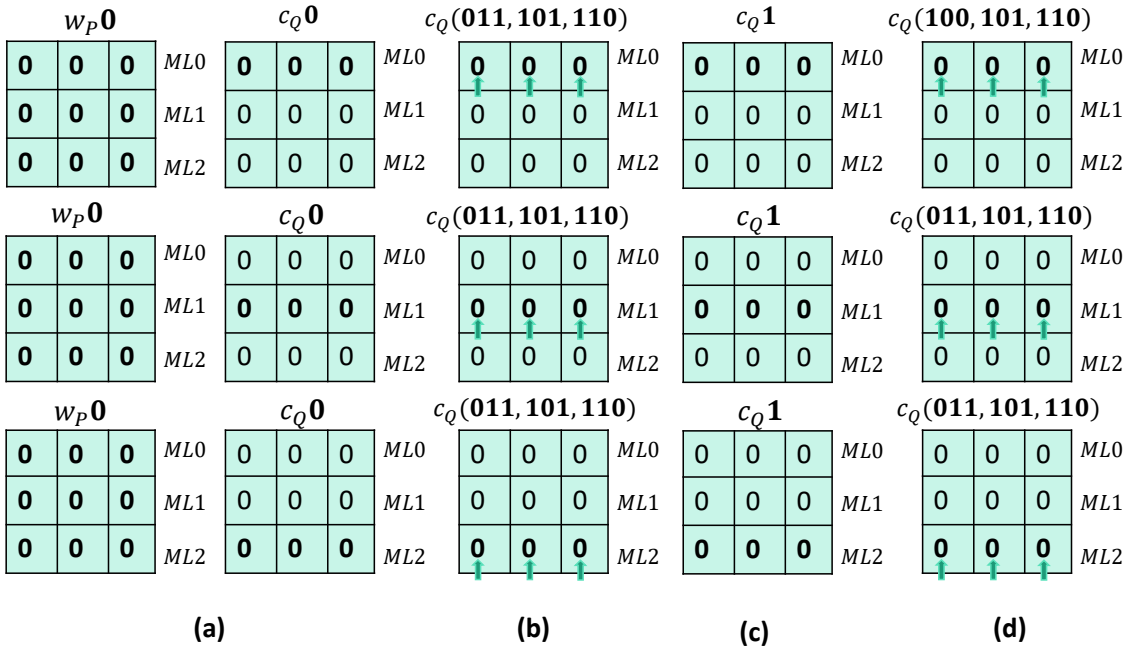


Figure 4.4.: 3×3 FF scenario of ReCAM array when performing (a) ME1, (b) ME2, (c) ME3, and finally (d) ME4, respectively

The proposed March-like test algorithm covers all determined fault types in the ReCAM array. We define the faults into four groups G_1, G_2, G_3, G_4. These groups correspond to fault types encountered during different ReCAM operations, specifically during matching and mismatching between P patterns and Q patterns.

4.7.3.1. Fault Groups

- G_1: The output where $(P = 0, Q = 0)$ is expected to be a match, but a mismatch occurs.
- G_2: The output where $(P = 0, Q = 1)$ is expected to be a mismatch, but a match occurs.
- G_3: The output where $(P = 1, Q = 1)$ is expected to be a match, but a mismatch occurs.
- G_4: The output where $(P = 1, Q = 0)$ is expected to be a mismatch, but a match occurs.

Thus, grouping these faults into equivalence classes reduces the number of test patterns needed for fault detection, reducing test time and complexity.

Lemma 1: The proposed March-like test sequence achieves 100% coverage for all fault groups.

Proof: The March-like test algorithm comprises several MEs designed to activate and detect faults in each group.

G_1 detection (ME1 and ME2)

- In **ME1**, all prototype bits P are programmed as 0 (LRS), and the search query Q is set to all 0s. A match is expected for all cells, and a fault in G_1 is identified if any mismatch occurs
- In **ME2**, executes the compare operation using a walking-0 sequence in Q across all B cells connected to each ML, i.e., compare bit-by-bit cell. For example, it compares the first cell with the comparand $c_Q1...10$, meaning the first bit in the ReTCAM cell is compared only, then compares the second cell with $c_Q1...10$, and so on, until the last B-bit cell in the row. If any ReCAM cell reports a mismatch, it is a fault in G_1.

Thus, ME1 and ME2 detect all faults in G_1.

G_2 detection (ME3 and ME4)

- In **ME3**, since all P bits remain programmed as 0, and the search query Q is set to all 1's. It is expected to be a mismatch; any match detected indicates a fault in G_2.
- In **ME4**, executes the compare operation using a walking-1 sequence in Q, while all values of P are kept at 0, across all B cells connected to each ML. Any match, rather than a mismatch, detects faults in G_2.

Thus, ME3 and ME4 detect all faults in G_2. Figure 4.4 shows an example of the FF scenario of the ReCAM matrix 3×3 when performing the sequence of March elements of (a) ME1, (b) ME2, (c) M3, and then (d) M4, respectively. Similarly, from M5 to M8, the same sequences of operations are performed, but this time, when a Write 1 operation is executed.

G_3 detection (ME5 and ME6)

- In **ME5**, all P bits are now programmed as 1 (HRS), and the search query Q is set to all 1's. A match is expected for all cells, and a fault in G_3 is identified if any mismatch occurs.
- In **ME6**, like ME2 walking 1's are applied. Any mismatch, rather than the expected match, detects faults in G_3.

Thus, ME5 and ME6 detect all faults in G_3.

G_4 detection (ME7 and ME8)

- In **ME7**, since all P bits remain programmed as 1, and the search query Q is set to all 0's. It is expected to be a mismatch; any match detected indicates a fault in G_4.
- In **ME8**, like ME2 walking 0's are applied. Any mismatch instead of the expected match detects faults in G_4.

Table 4.4.: March elements corresponding to ReCAM test operations.

Operation	Fault Response	March Element
$w0/c0$	MM	ME1, ME2
$w0/c1$	M	ME3, ME4
$w1/c1$	MM	ME5, ME6
$w1/c0$	M	ME7, ME8

Thus, ME7 and ME8 detect all faults in G_4.

Hence, each March element in the sequence detects specific all-fault groups G_1, G_2, G_3, G_4 by writing either 0 or 1 into the stored prototype data P and by manipulating the search query data Q using different binary patterns. Therefore, the test achieves 100% fault coverage for the entire ReCAM array. Table 4.4 summarizes the failed response of a ReCAM cell undergoing comparison after write operations and the corresponding March element detecting the faults.

The proposed test algorithm includes; To perform the test operations, ME1 and ME5 require 1 Write (in parallel across all MLs) and $2N$ comparisons (due to the shared ADC). Additionally, ME2, ME4, ME6, and ME8 require B-bit comparisons per each ML, ME3 and ME7 require only $2N$ comparisons. Summing up to 2 Write operations only, and $4N(1 + B)$ Compare operations to achieve 100% fault coverage. This proposed March-like test algorithm can also be adapted to ReCAMs based on other CAM cell designs, such as [96], [97], [98], [99]. While the general principles of the March-like test sequence remain the same, specific adjustments are necessary to account for the specific structure of a CAM cell.

4.7.4. MBIST in ReCAM Architecture

Implementing Memory BIST in the ReCAM architecture can be augmented on top of existing approaches commonly used to test NVMs to enhance defect coverage, as demonstrated in [100], [101]. In ReCAM, the shared ADC block replaces the sense amplifier in typical NVM memories. Additionally, the proposed March-like test algorithm can be implemented in the memory BIST controller as a Finite State Machine (FSM)-based pattern generator, which is necessary to systematically test match and mismatch scenarios. The outputs for this generator are used to fill the CAM array with prototype and query data. To support our testing methodology, a comparator is needed to determine whether the Out_{ADC} for the test patterns matches the expected value. This enables us to enhance comprehensive testing coverage using our proposed algorithm while minimizing hardware overhead in MBIST implementations. The details of such MBIST realization will be in our future work.

4.8. Summary

This chapter presented a comprehensive fault modeling and testing framework for ReCAM architectures. First, the operational principles of ReCAM and the analog similarity-search mechanism were introduced. Then, the major defect mechanisms affecting ReCAM reliability were analyzed and classified into intra-cell and inter-cell defects, including ReRAM defects, transistor faults, bridging defects, and resistive-open defects affecting the match-line infrastructure.

To efficiently detect these faults, a March-like testing methodology tailored for ReCAM architectures was proposed. The developed test algorithm exploits the match and mismatch behavior of ReCAM cells to identify both digital and analog fault manifestations while maintaining low testing complexity. The corresponding fault behaviors and detection March elements were systematically derived and evaluated for all considered fault models.

Simulation results demonstrated that the proposed methodology achieves complete fault coverage with fewer test operations, making it suitable for large-scale ReCAM arrays. In particular, the proposed algorithm requires only two write operations and $4N(1 + B)$ compare operations for an $N \times B$-bit ReCAM array, thereby reducing test overhead while preserving full fault detectability. Overall, this chapter establishes a scalable and efficient foundation for reliable testing of ReRAM-based CAM architectures under both device-level and array-level defects.

5. PVTVAD: PVT-Variation-Aware Fault Diagnosis of ReRAM-based CAM Arrays

Process, Voltage, and Temperature (PVT) variations alter ReRAM resistance distributions and sensing behavior, making the ML voltage signatures highly sensitive to operating conditions. Since CiM-based ReCAM architectures rely on analog sensing, even small variations can distort ML voltages, reduce distinguishability between neighboring fault signatures, and lead to fault masking or misinterpretation of wrong output. Fault diagnosis, therefore, becomes essential not only during manufacturing tests to improve yield, but also during system operation to identify defective cells that may degrade similarity-search accuracy.

Unlike conventional memories, diagnosing faults in ReCAM arrays is considerably more challenging because multiple cells contribute simultaneously to a shared analog ML response. As a result, the faulty cell location cannot be directly observed and must be inferred from small deviations in ML voltage caused by the defect. Under PVT variation, these voltage differences become increasingly overlapped, reducing diagnosis resolution and making precise fault localization difficult, especially for large arrays.

Accurate fault localization in ReCAM arrays is particularly challenging because all cells connected to the same row share a common ML. Consequently, the faulty cell location can only be inferred from small voltage differences observed at the ML output. As array dimensions increase, these location-dependent voltage differences become even smaller under PVT variation, requiring highly sensitive and variation-aware diagnostic methodologies.

To address these challenges, this chapter presents a variation-aware Design-for-Testability (DfT) framework for precise fault localization in ReCAM arrays. The proposed architecture combines a tunable operational amplifier (OpAmp) with adaptive ADC-based sensing to amplify small ML voltage differences and improve diagnostic resolution under PVT variation. An iterative multi-step diagnosis flow dynamically adjusts gain and sampling time to progressively isolate the faulty cell location while minimizing sensing overhead and diagnosis latency.

Experimental results demonstrate that favorable voltage and temperature operating points can be leveraged to accelerate diagnosis and improve the overlap of similarity levels signatures. In addition, a partitioning-based diagnosis strategy significantly reduces the required amplification gain for large arrays while maintaining accurate localization capability. Across multiple fault scenarios and PVT corners, the proposed framework achieves precise fault localization with low diagnosis latency and minimal hardware overhead.

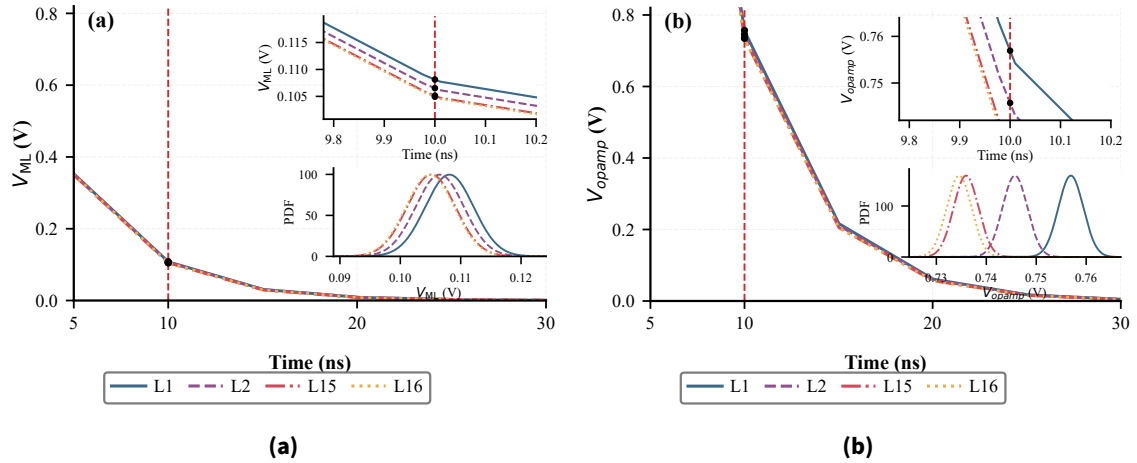


Figure 5.1.: (a) ML voltages for the all-match pattern in a 16-bit ReCAM with fault F1 at L1, L2, L15, and L16 (L1 furthest from sensing). Insets show a zoom near $t_s = 10$ ns (top) and the PDFs (bottom), highlighting a strong overlap due to PV and ML parasitic RC. (b) Corresponding OpAmp outputs, where amplification improves separation across locations.

5.1. Motivation Example

Figure 5.1(a) highlights the main diagnosis challenge in ReCAM arrays, where faults located at different cells produce only infinitesimal ML voltage differences when a single defect (F1) is introduced at two different locations in a 16-bit ReCAM array under the all-match case. Specifically, the voltage difference between inserting the fault in the first and second cell is as small as 0.1 mV. These tiny discrepancies arise from the resistive–capacitive (RC) parasitics along the ML, which cause position-dependent signal attenuation relative to the ADC input. Due to PVT variations, the ML voltage distributions overlap almost completely, further collapsing the distinguishability between the two cases and resulting in identical quantized outputs. This makes diagnosing each faulty cell extremely challenging. In contrast, Figure 5.1(b) shows the corresponding OpAmp outputs, where amplification increases the separation between the two cases. As a result, their distributions become more distinguishable, producing different quantized ADC values and enabling effective diagnosis of the faulty cell's location.

To overcome this issue, we integrate a DfT circuitry that employs a tunable ideal OpAmp voltage controller with high gain, enhanced by a trimming circuit for adjustable gain and sampling time selection. The OpAmp amplifies the voltage differences between ML voltages at various faulty locations, enabling the ADC to generate distinct outputs for each defective cell. This enhances diagnostic accuracy by allowing precise fault localization across the ReCAM array. Figure 5.1(b) demonstrates that applying an ideal gain of $5X$ sufficiently amplifies the voltage differences, significant enough to be distinguishable by the ADC. Especially in the last cells, which are physically closer to the ADC. This observation suggests that faults near the ADC are detected more easily than those farther away.

5.2. Problem Statement

Fault diagnosis in ReCAM architectures is particularly challenging because faults must be inferred from small ML voltage deviations shared across multiple cells, making accurate localization highly sensitive to variations.

This challenge becomes significantly more severe under process, voltage, and temperature (PVT) variations, which distort the ML voltage distributions associated with different fault locations and similarity levels. Consequently, neighboring ML voltage signatures may overlap, reducing diagnostic resolution and making it difficult to distinguish between faulty and fault-free behaviors.

Moreover, conventional March-based diagnostic techniques are not well-suited to ReCAM architectures because they primarily rely on binary pass/fail observations and do not account for analog ML behavior or variation-induced distribution shifts. As ReCAM array dimensions increase, the voltage differences between fault locations become even smaller, often requiring excessive sensing gain and multiple diagnosis iterations to achieve reliable localization.

Therefore, there is a critical need for a variation-aware diagnostic methodology capable of:

- accurately localizing faulty cells under PVT variation,
- adaptively improving ML distinguishability,
- minimizing sensing overhead and diagnosis latency, and
- maintaining robust diagnosis capability across multiple operating corners.

To address these challenges, this work proposes a PVT-aware DfT framework that combines adaptive gain control, configurable sampling times, and iterative diagnosis to enhance fault-localization resolution in ReCAM arrays.

5.2.1. Contributions

The main contributions of this chapter are summarized as follows:

- A comprehensive DfT architecture is proposed for precise fault localization in ReRAM-based CAM arrays under process, voltage, and temperature (PVT) variations.
- An iterative diagnosis flow is introduced that dynamically adjusts OpAmp gain and sensing time to improve ML signature distinguishability and progressively localize faulty cells.
- The impact of PVT variations on ML voltage distributions, ADC separability, gain requirements, and diagnosis latency is thoroughly analyzed across multiple operating corners.

- A scalable partition-based localization strategy is proposed to reduce the required sensing gain for large ReCAM arrays by approximately 80%, improving scalability and sensing robustness.
- Simulation results demonstrate accurate fault localization within only 2–8 diagnosis iterations and approximately 2.29× lower diagnosis latency than March-based diagnosis approaches, while incurring only 0.36% area overhead and up to 6.3% test-power overhead.

5.3. DFT scheme for CAMs and NNMs

DfT techniques for ReRAM devices are designed to address the unique challenges posed by defects and fault mechanisms specific to ReRAM technology, which differ from those in charge-based memory. Previous research has focused on developing fault models and defect characterization methods tailored to the physical and operational properties of ReRAM cells. For instance, DfT approaches leveraging the computational capabilities of ReRAM in CiM setups can accelerate the testing process, reducing costs and improving scalability [102]. Additionally, test-generation techniques have been proposed to detect faults such as stuck-at faults and retention failures, ensuring effective fault detection in emerging technologies like ReRAM devices [94], [103].

Furthermore, diagnosis plays a critical role in improving yield by enabling the identification and bypassing of faulty cells at runtime. In this regard, some works address NVM with standard memory functionality [104], [105] as well as CAM diagnostics [106], [107]. Despite their different implementations, the common point between these two groups is the use of March test-based algorithms in their diagnostic methods. This, however, requires a long March sequence for diagnosis, resulting in a high diagnosis time.

Furthermore, despite the promising characterizations of ReCAM, to the best of our knowledge, no existing work addresses diagnostic methods for this structure. This is the shortcoming that this paper aims to address. Moreover, this paper further tackles the high timing complexity associated with the March sequence.

5.4. Fault diagnosis Framework

Unlike traditional memory, CAM operations focus on data comparison rather than conventional read-and-write accesses. In ReCAM, the input patterns Q's are compared with the stored values P's. The result of this comparison—whether a match or mismatch—is reflected in the ML voltage, which is subsequently digitized using an ADC. To develop our diagnostic approach for ReCAM, we first define faults that can cause errors during the match and those that can cause errors during mismatch. We perform electrical single-defect injections into the ReCAM cells to simulate these faults. For each fault scenario, Q is applied against the stored P across the entire ReCAM array of all-match and all-mismatch, respectively. The O_{ADC} outputs are compared with the fault-free ones in each location independently.

Table 5.1.: Simulation Setup and Parameters

Category	Parameter Details
Technology-Level Parameters	
ReRAM Model [23]	JART VCM Read Variability
Filament Radius	45 nm
Disc Region Length	0.6 nm
N_{init} HRS, LRS	9 m, 3×10^{26} m^{-3}
N_{init} Undefined Region [94]	65 m, 1×10^{26} m^{-3}
N_{init} Deep HRS, Deep LRS [94]	5 m, 1 k, 1×10^{26} m^{-3}
Circuit-Level Parameters	
Simulation Tool	Cadence Virtuoso
CMOS Technology	GlobalFoundries 22FDX
Supply Voltage (V_{DD})	800 mV
Temperature	27°C
ADC [95]	4-channel, 8-bit resolution
Interconnect per ReCAM Cell (22 nm) [47], [48]	ML: $R = 11.95\ \Omega$, $C = 16.63$ aF
Match-Line Length	64-bit
Monte Carlo Samples	1,000
Performance and Area Parameters	
ADC Area	0.095 mm^2
CAM Cell Area [108]	0.0242 μm^2
Feedback Resistor Area	144 μm^2
OpAmp Area	196 μm^2
ReCAM Array Power	0.108 mW
ADC Power	50.8 mW
OpAmp Power	0.806 mW

5.4.1. DfT Implementation

Conventional ReCAM sensing has limited diagnostic resolution because, under process variations (PV) and ML parasitics, faults injected at different cell locations may produce very similar or even identical ADC outputs (O_{ADC}), as illustrated in Fig. 5.1. PV further broadens the statistical distributions of each fault-location signature, increasing overlap and reducing the distinguishability between neighboring faulty locations.

To quantify the diagnosis resolution, the separability between two fault-location signatures i and j can be expressed as

$$R_{ij} = \frac{|\mu_i - \mu_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}}, \tag{5.1}$$

where μ_i and μ_j denote the mean ADC outputs, while σ_i and σ_j represent the corresponding standard deviations under PV. Larger values of R_{ij} indicate better fault-signature separability and therefore higher diagnosis resolution. As PV increases, the variance terms R_{ij} decrease, making direct single-shot fault localization increasingly difficult.

To improve distinguishability, we introduce the DfT circuitry shown in Fig. 5.2, where a tunable OpAmp amplifies small ML voltage differences before digitization. By increasing the separation among location-dependent signatures, the ADC can resolve faults that would otherwise remain indistinguishable. Rather than attempting to identify all fault locations using a single fixed gain and sampling time, which would require excessive amplification and incur significant overhead, we employ a multi-step diagnostic strategy.

In each diagnosis setup $D^k = (G^k, t_s^k)$, the OpAmp gain and sampling time are adaptively increased until a unique faulty-cell signature is obtained. This progressive strategy forms the basis of the decision-tree diagnosis flow and avoids overdesign associated with a single aggressive sensing configuration. Under stronger PV conditions, additional diagnosis iterations, higher gain settings, longer sampling times, or array partitioning may become necessary to maintain sufficient fault-signature separability.

The gain G is tuned through the trim-based feedback structure of the OpAmp [109], [110], while the sampling time t_s is adjusted through programmable delay elements in the sample-and-hold circuitry. At each step, the measured O_{ADC} is compared against the expected signatures, and the decision tree determines whether the faulty location is identified or an additional diagnosis setup is required.

In this work, the OpAmp is modeled as an ideal tunable-gain amplifier; therefore, internal OpAmp non-idealities under PV, such as gain error, offset, or bandwidth degradation, are not explicitly considered. Instead, the analysis focuses on the impact of ReRAM D2D variability and CMOS PVT variations on the observable ML signatures and the resulting diagnosis resolution.

5.4.2. Fault Classification

Table 5.2 summarizes the classification of the considered ReCAM faults based on their observable behavior under representative input conditions. Although the faults originate from different sources, including transistors, ReRAM devices, and interconnects, their impact at the CAM level can be grouped into two main categories. The first group includes faults that disturb the match condition by introducing an unintended discharge, resulting in a Match-to-Mismatch transition. The second group includes faults that weaken or block the discharge under mismatch conditions, leading to a Mismatch-to-Match transition.

Based on this observation, the faults can be abstracted into these two dominant behavioral classes. Therefore, in this work, F1 and F2 are used as representatives of the two categories. F1 models the case where a fault causes extra discharge during a match operation, while F2 captures the case where the discharge is reduced or suppressed during a mismatch. Other faults listed in Table 5.2 show similar behavior under appropriate input patterns and thus fall into the same two classes.

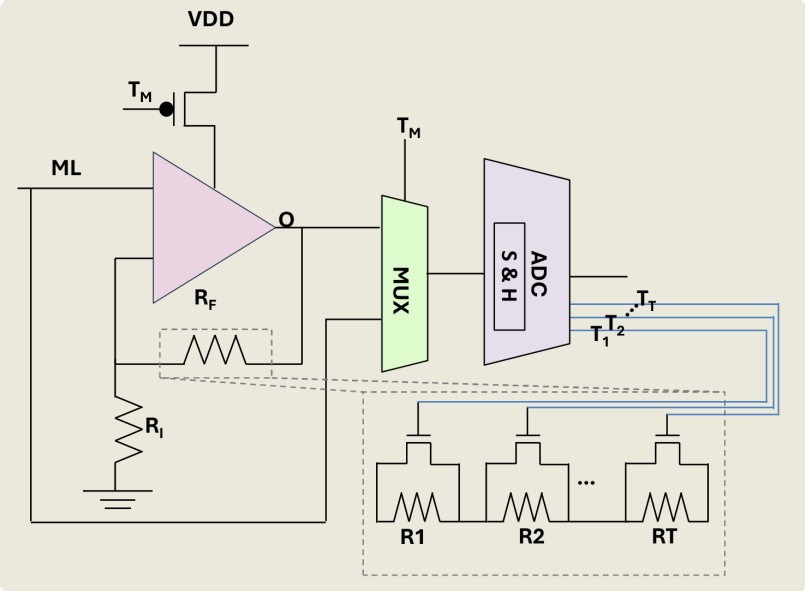


Figure 5.2.: DfT circuitry to generate different Gain (G) and sampling time (t_s), using a tunable OpAmp structure with trimming concept using a controllable resistor as feedback resistance, T_M is the signal for enabling the test mode

5.4.3. Diagnosis Test Pattern

In this work, we classify the test input binary patterns applied to the ReCAM array as either all-match or all-mismatch. For the all-match category, we consider uniform patterns consisting entirely of '1's or '0's applied to both P and Q across the entire array. In the all-mismatch category, the patterns in P and Q are opposite. This means that if P consists entirely of '0's, then Q consists altogether of '1's, and vice versa.

The all-match category can also include the 'X' (don't care) pattern, which we implement in our diagnosis for arrays of any feasible size. In this case, the 'X' pattern is implemented by programming both ReRAMs of a cell to HRS, representing the presence of 'X' data in the prototype. Similarly, in the query phase, the 'X' data can be represented by grounding both transistors (T_L and T_R). As a result, when 'X' data is present, the ML voltage remains high, which is interpreted as a match. Moreover, this pattern is particularly useful after identifying a faulty cell, as it can be used in the bypass process to exclude the faulty cell from the array as a defect-tolerance strategy. By marking the faulty cell as 'don't care', it is prevented from influencing subsequent comparisons, thereby improving fault tolerance and maintaining the correct functionality of the ReCAM array in the presence of faults.

The use of only all-match and all-mismatch patterns is motivated by the fault classification presented in Table 5.2. As discussed, all considered faults can be grouped into two dominant observable behaviors: faults that disturb the match condition (Match-to-Mismatch) and faults that affect the mismatch condition (Mismatch-to-Match). Therefore, it is sufficient to excite the array under both pure-match and pure-mismatch scenarios to reveal both classes of faults.

Using mixed or random patterns would introduce a combination of match and mismatch cells along the match-line, leading to intermediate ML voltage levels that depend on the

Table 5.2.: Classification of ReCAM faults according to their observed behavior under representative test conditions

Fault	Origin	Fault Type	(Q,P)	Observed Behavior
Match → Mismatch (False Mismatch)				
F1	CMOS	T_L stuck-on	(0,0)	Match → Mismatch
F3	CMOS	T_R stuck-on	(1,1)	Match → Mismatch
F5	CMOS	T_{EN} stuck-on	(1,1)	Match → Mismatch
F7	ReRAM	R_L stuck-at-LRS	(1,1)	Match → Mismatch
F9	ReRAM	R_R stuck-at-LRS	(0,0)	Match → Mismatch
F11	ReRAM	R_L Deep LRS	(1,1)	Match → Mismatch
F13	ReRAM	R_R Deep LRS	(0,0)	Match → Mismatch
F15	ReRAM	R_L Undefined	(1,1)	Match → Mismatch
F18	Interconnect	Resistive short	(0,0)	Match → Mismatch
Mismatch → Match (False Match)				
F2	CMOS	T_L stuck-open	(1,0)	Mismatch → Match
F4	CMOS	T_R stuck-open	(0,1)	Mismatch → Match
F6	CMOS	T_{EN} stuck-open	(1,0)	Mismatch → Match
F8	ReRAM	R_L stuck-at-HRS	(1,0)	Mismatch → Match
F10	ReRAM	R_R stuck-at-HRS	(0,1)	Mismatch → Match
F12	ReRAM	R_L Deep HRS	(1,0)	Mismatch → Match
F14	ReRAM	R_R Deep HRS	(0,1)	Mismatch → Match
F16	ReRAM	R_R Undefined	(1,0)	Mismatch → Match
F17	Interconnect	Resistive open	(1,0)	Mismatch → Match

number and positions of mismatches. Under PV, these intermediate levels tend to overlap, making it difficult to isolate the effect of a single faulty cell. In contrast, uniform patterns enforce a controlled operating condition in which the ML behavior is dominated either by full discharge or by no discharge. This significantly improves the observability of fault-induced deviations and enables reliable differentiation between fault locations.

5.4.4. Experiment setup

To evaluate the impact of a single fault at different locations within the ReCAM array, we perform comprehensive electrical-level simulations using SPICE. We focus on the representative scenarios that affect the all-match condition by injecting F1 in a different location each time, as shown in Table5.2 Experiments are conducted using all-match input patterns across a very wide range of process, voltage, and temperature (PVT) conditions to capture the effects of these variations on the diagnosis process. While F1 is analyzed under the all-match scenario, the proposed diagnosis methodology is generic and can be extended to mismatches and other patterns as well. The experiment begins with a single ReCAM cell to determine input patterns that effectively detect these faults.

For instance, when injecting F1, the appropriate input test vector is $(P = 0, Q = 0)$, which is a matching case. Conversely, for F2, the appropriate input test vector is $(P = 1, Q = 0)$, which is a mismatch case as listed in Table 5.2. To extend these observations to the entire ReCAM array, we apply all-match and all-mismatch binary test vectors and independently observe the O_{ADC} at each location to pinpoint faulty cells.

In this work, we introduce the faulty behavior in ReRAM cells using the Defect-Oriented model, as described in [23]. This model accounts for physical defects by altering technology parameters, affecting the device's electrical characteristics. Specifically, defects in ReRAM

Algorithm 1 PVT-Aware Statistical Fault Localization

Require: Offline MC dataset $\mathcal{D}_{\text{MC}}$
Require: Tester measurements $\mathcal{D}_{\text{S}}$
Require: Separation threshold Δ_{PVT}
Ensure: Fault localization result (T_D, L, t_s, G)

Phase I: Offline Reference Construction
for all T_D **do**
  for all t_s **do**
    Compute $V_{\max}(T_D, t_s)$ from all MC samples
    Select gain $G(T_D, t_s)$ using $V_{\max}(T_D, t_s)$
    for all locations L **do**
      Convert $V_{\text{ML}}(T_D, t_s, L)$ to O_{ADC}
      Compute $\mu(T_D, t_s, L)$ and $\sigma(T_D, t_s, L)$
    end for
  end for
end for
Store lookup table $\{\mu, \sigma, G\}$

Phase II: On-Tester Diagnosis
for all T_D **do**
  for all candidate locations L **do**
    for all t_s in increasing order **do**
      Read $V_{\text{ML,S}}(T_D, t_s, L)$ from tester
      Convert $V_{\text{ML,S}}$ to O_{ADC} using $G(T_D, t_s)$
      Compute $\text{sep}(L, t_s)$ against stored references
      if $\text{sep}(L, t_s) \geq \Delta_{\text{PVT}}$ **then**
        report (T_D, L, t_s, G)
        break
      end if
    end for
    if no t_s satisfies Δ_{PVT} **then**
      Select t_s with maximum $\text{sep}(L, t_s)$
      report best available (T_D, L, t_s, G)
    end if
  end for
end for

devices are modeled by adjusting the initial oxygen vacancy concentration (N_{init}) in the ReRAM device's disk, which is part of the JART model v1bas outlined in [23]. All simulations follow the setup described in Table 9.1.

5.4.5. PVT Setup

To study the impact of PVT variations on ReCAM array diagnosis, we evaluate the all-match input pattern for array sizes ranging from 8-bit to 64-bit cells under multiple

operating conditions in the presence of a representative fault type (e.g., F1), as introduced in the previous section.

We vary both the ambient temperature (T) and the supply voltage (V_{DD}). Starting from the nominal corner at 27°C and $V_{DD} = 0.8\,\text{V}$, we examine (i) fixed-voltage corners with varying temperatures (0°C and 85°C), and (ii) fixed-temperature corners with varying V_{DD} (0.6 V and 1.0 V). For each temperature–voltage operating point, device-to-device variations in both the transistor and ReRAM devices are incorporated through 1,000 Monte Carlo (MC) simulations. The transistor variability follows the models summarized in Table 9.1, while ReRAM variability is captured statistically using the model in [23].

5.5. ReCAM Diagnosis Algorithm

Previously, we first developed a diagnosis flow under ideal conditions, assuming only hard faults with the absence of PVT variation [111]. In this work, however, ReCAM performs analog computation, and the ML voltage used to infer Hamming distance is highly sensitive to PVT variations. Such variations can mask faulty cells or lead to incorrect localization. To address this challenge, we propose a PVT-aware statistical fault-localization method that explicitly accounts for distribution overlap. Instead of relying on single deterministic values, the proposed approach evaluates the separability between statistical distributions, as summarized in Algorithm 1.

5.5.1. PVT-Aware Statistical Fault Localization Algorithm

The proposed PVT-aware diagnosis consists of two sequential phases: (i) offline statistical reference construction, and (ii) on-tester fault localization. The first phase builds a PVT-aware reference model via MC simulations, while the second phase uses it during diagnosis to identify the faulty cell under varying conditions.

Phase I: Offline PVT Reference Construction Prior to on-chip diagnosis, an offline characterization step is performed to capture the impact of process, voltage, and temperature variations on the ML behavior. For each diagnosis test T_D, sampling time t_s, and fault location L, MC simulations are performed, and the corresponding ADC outputs are collected.

Let $\mathcal{A}_{T_D,t_s,L}$ denote the set of ADC samples for configuration (T_D, t_s, L). The statistical reference for each location is characterized by its mean and standard deviation:

$$\begin{aligned} \mu(T_D, t_s, L) &= \text{mean}(\mathcal{A}_{T_D,t_s,L}), \\ \sigma(T_D, t_s, L) &= \text{std}(\mathcal{A}_{T_D,t_s,L}). \end{aligned} \tag{5.2}$$

For each (T_D, t_s), the maximum ML voltage over all MC samples and locations is first determined:

$$V_{\max}(T_D, t_s) = \max_{L,r} \{\text{ML}(T_D, t_s, L, r)\}\,. \tag{5.3}$$

Based on this value, the gain used for amplification and ADC conversion is selected as

$$G(T_D, t_s) = \max\left(1, \left\lfloor \frac{V_{DD}}{V_{\max}(T_D, t_s)} \right\rfloor\right). \tag{5.4}$$

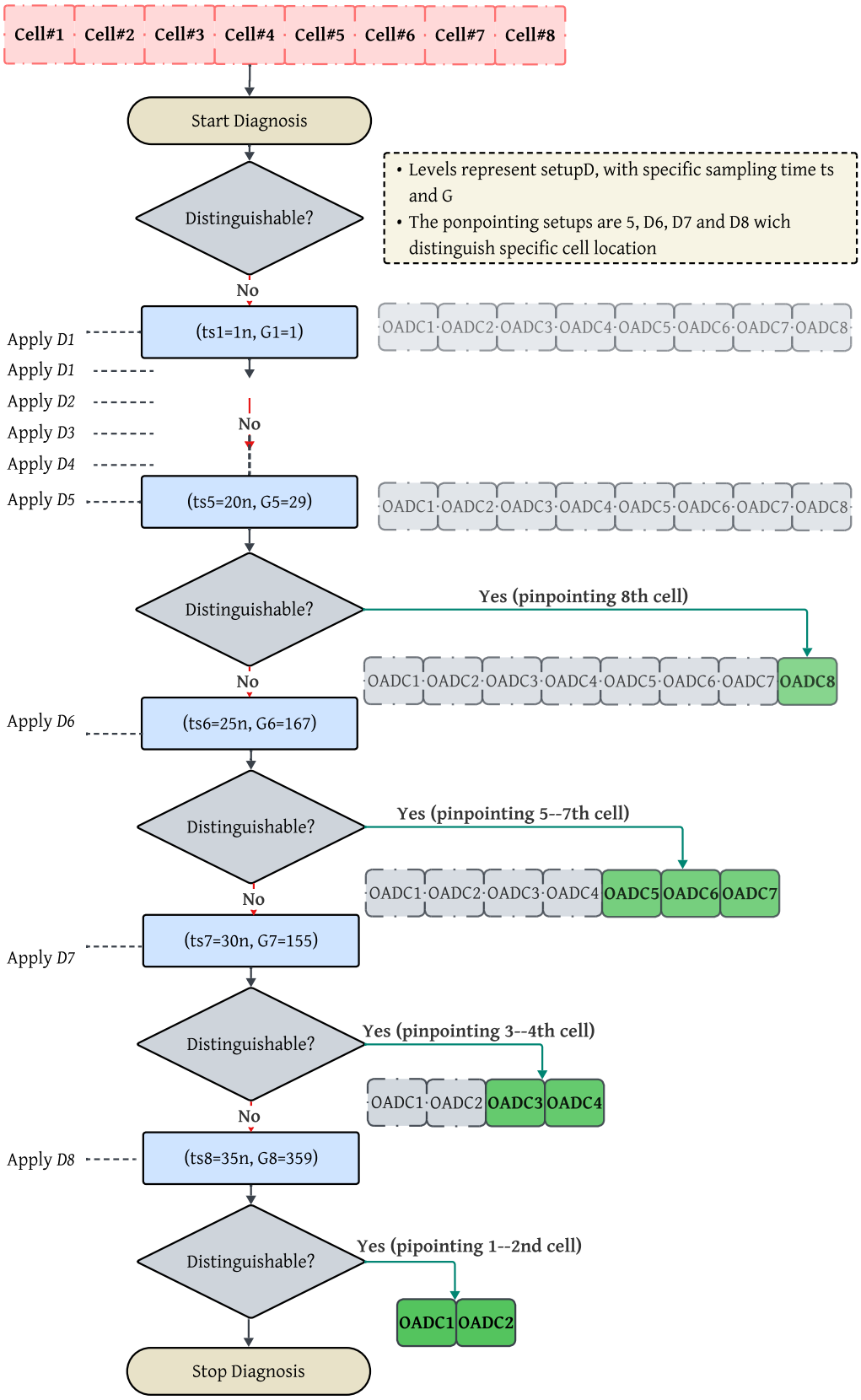


Figure 5.3.: Decision-Tree algorithm with multiple gains (G) and corresponding sampling time (t_s) setup while all-match is applied and injecting F1 in each location of the 8-bit ReCAM cells connected to the match line (ML)

The gain at each step is selected as the minimum amplification required to maximize use of the ADC dynamic range while keeping the amplified ML signal within its input limits.

This procedure generates a lookup table

$$(T_D, t_s, L) \rightarrow (\mu, \sigma, G),$$

which serves as the PVT-aware statistical reference used during diagnosis. Since this characterization is performed offline, no statistical computation is required at test time.

Phase II: On-Tester Fault Localization During diagnosis, measured ML values from the chip under test are converted into ADC outputs using the same gain $G(T_D, t_s)$ obtained in Phase I. For each candidate location L and sampling time t_s, the measured ADC output is compared against the reference distributions of all other locations.

To quantify distinguishability, a separation metric is defined as

$$\mathrm{sep}(L, t_s) = \min_{j \neq L} \left| \frac{O_{\mathrm{ADC}}(L, t_s) - \mu(T_D, t_s, j)}{\sigma(T_D, t_s, j)} \right|. \tag{5.5}$$

A location is considered uniquely identifiable when

$$\mathrm{sep}(L, t_s) \geq \Delta_{\mathrm{PVT}}, \tag{5.6}$$

where $\Delta_{\mathrm{PVT}} = 3$ corresponds to a three-sigma confidence criterion. Starting from smaller sampling times, the algorithm searches for the earliest t_s satisfying this condition. This follows a decision-tree-like progression, where larger t_s values progressively improve separation when earlier configurations are insufficient.

If no sampling point satisfies the threshold, a fallback rule selects

$$t_s^*(L) = \arg\max_{t_s} \mathrm{sep}(L, t_s), \tag{5.7}$$

which provides the best available discrimination.

For each diagnosed fault, the algorithm reports the faulty location, the selected sampling time t_s^*, corresponding gain G, achieved separation, and whether fallback selection was invoked.

Diagnostic resolution is quantified by the separability of adjacent fault-location signatures, which decreases as PV variation increases. Algorithm 1 summarizes the complete two-phase PVT-aware fault localization flow. This procedure effectively constructs a decision-tree-like flow in which progressively larger t_s values improve separation until a single faulty location can be pinpointed. It integrates seamlessly with the no-PVT diagnosis flow and remains fully compatible with the March-like testing procedure employed in the ReCAM array.

5.5.2. Bypassing Faulty Cells

The need to diagnose faulty cells in ReCAM arrays extends beyond traditional failure mode analysis, yielding improvements. Identifying the exact location of defective cells

Table 5.3.: PVT diagnostic scenarios for ReCAM arrays under the all-match input. The table reports the required gain (G), sampling time (t_s), iteration count, and their ratios in the worst case scenario (Last Identification) with respect to the ideal no-PVT baseline

ReCAM Size & Test Pattern			First Identification Cell			Last Identification Cell			PVT vs. Ideal Ratios		
Size	**(V, T)**	**Test Pattern**	t_s (ns)	G	# iters	t_s (ns)	G	# iters	G	t_s	# iters
Ideal (No-PVT) Baseline											
8-bit	(0.8 V, 27°C)	(0,0)	5	1	2	30	90	7	1.0×	1.0×	1.0×
16-bit	(0.8 V, 27°C)	(0,0)	5	2	2	35	4102	8	1.0×	1.0×	1.0×
32-bit	(0.8 V, 27°C)	(0,0),(0,x)	5	2	4	30	4102	16	1.0×	1.0×	1.0×
64-bit	(0.8 V, 27°C)	(0,0),(0,x)	5	2	8	35	4102	32	1.0×	1.0×	1.0×
Directly Diagnosable Without Partitioning (PVT Corners)											
	(0.8 V, -40°C)	(0,0)	20	16	5	40	522	9	5.80×	1.33×	1.29×
	(0.8 V, 0°C)	(0,0)	5	1	2	40	533	9	5.92×	1.33×	1.29×
	(0.6 V, -40°C)	(0,0)	20	20	5	35	338	8	3.76×	1.17×	1.14×
8-bit	(0.6 V, 0°C)	(0,0)	5	2	2	25	60	6	0.67×	0.83×	0.86×
	(0.6 V, 27°C)	(0,0)	5	2	2	30	139	7	1.54×	1.00×	1.00×
	(0.6 V, 85°C)	(0,0)	5	2	2	35	291	8	3.23×	1.17×	1.14×
	(1.0 V, -40°C)	(0,0)	15	5	4	40	363	9	4.03×	1.33×	1.29×
	(1.0 V, 0°C)	(0,0)	5	1	2	40	372	9	4.13×	1.33×	1.29×
	(1.0 V, 27°C)	(0,0)	5	1	2	35	149	8	1.66×	1.17×	1.14×
	(1.0 V, 85°C)	(0,0)	5	1	2	40	313	9	3.48×	1.33×	1.29×
	(0.8 V, 0°C)	(0,0)	5	2	2	35	4221	8	1.03×	1.00×	1.00×
	(0.6 V, 0°C)	(0,0)	5	3	2	40	22 172	9	5.41×	1.14×	1.13×
	(0.6 V, 27°C)	(0,0)	5	3	2	30	1664	7	0.41×	0.86×	0.88×
16-bit	(0.6 V, 85°C)	(0,0)	5	3	2	40	19 034	9	4.64×	1.14×	1.13×
	(1.0 V, 0°C)	(0,0)	10	5	3	30	867	7	0.21×	0.86×	0.88×
	(1.0 V, 27°C)	(0,0)	5	1	2	30	855	7	0.21×	0.86×	1.00×
	(1.0 V, 85°C)	(0,0)	5	1	2	35	2749	8	0.67×	1.00×	1.00×
Requires Partitioning (PVT Corners)											
32-bit	(1.0 V, 27°C)	(0,0),(0,x)	5	1	4	30	855	14	0.21×	1.00×	0.88×
64-bit	(1.0 V, 27°C)	(0,0),(0,x)	5	1	8	30	855	28	0.21×	0.86×	0.88×

can significantly benefit fault tolerance and reliability [112], [113], [114]. Once the faulty cell is identified, it can be excluded from the comparison to ensure that the remaining fault-free cells perform correctly. This can be achieved by marking the faulty cell as 'don't care' and isolating its contribution to the ML. Such an exclusion can be implemented in the faulty ReCAM by applying a binary ('0') for both Q and $\bar{Q}$, effectively disconnecting the faulty cell from the entire array.

5.6. Diagnosis Evaluation

5.6.1. Diagnosis Flow Results

Figure 5.3 illustrates the operation of the proposed diagnosis flow for an 8-bit ReCAM under an all-match test with an injected fault F1. Starting from an initial diagnosis setup, the algorithm progressively adjusts the sampling time t_s and gain G across successive diagnosis setups D^i, increasing the distinguishability among candidate fault locations. At each step, the observed ADC signature is evaluated, and candidate locations that are inconsistent with the measured response are pruned, thereby refining the search space in a decision-tree-like manner until a single faulty cell is uniquely identified.

The number of diagnosis iterations depends on the faulty cell location due to location-dependent ML parasitics and sensing sensitivity. In particular, cells farther from the sensing circuitry (e.g., location L1) exhibit smaller and harder-to-distinguish signatures, requiring more diagnosis iterations and larger (G, t_s) settings. In contrast, cells closer to the ADC (e.g., L8, L7) can be localized with fewer iterations because their signatures separate earlier in the diagnosis flow. For the example shown, the best-case diagnosis corresponds to the cell closest to the ADC, which is localized in five diagnosis steps, whereas the worst-case diagnosis for the farthest cell requires eight steps.

These results highlight the hierarchical nature of the proposed diagnosis methodology: early setups provide coarse distinguishability among groups of candidate cells, while later setups progressively improve resolution until exact localization is achieved. This adaptive progression avoids relying on a single overly aggressive sensing configuration, while enabling systematic fault pinpointing across the full ReCAM array.

5.6.2. PVT Diagnosis Results

In the diagnostic scenario, the best (worst) case, i.e., the lowest (highest) number of diagnosis iterations, occurs for the cell closest to (farthest from) the ADC. Table 5.3 summarizes the required gain G, sampling time t_s, and diagnosis iterations under different PVT corners.

Under the nominal condition ($T = 27°\text{C}$, $V_{\text{DD}} = 0.8\,\text{V}$), which serves as the no-PVT baseline, the first faulty cell is identified using $(t_s, G) = (5\,\text{ns}, 1)$ in two diagnosis iterations, whereas the farthest cell requires $(30\,\text{ns}, 90)$ and seven iterations. As listed in Table 5.3, even in the worst-case scenario for an ML connected to 16 cells, only two to eight diagnosis steps are required, which remains less than half of conventional March-test-based diagnosis approaches such as [106].

PVT variations primarily affect the required gain, whereas the sampling time and diagnosis iterations vary more moderately. For example, at $V_{DD} = 0.6\,\text{V}$, the gain required for identifying the last cell increases from 60 at 0°C to 291 at 85°C. Similarly, the added negative-temperature corner at −40°C shows that low temperature does not uniformly degrade diagnosis; rather, its impact depends on the supply voltage and mainly shifts the optimal (G, t_s) operating point. In general, lower temperatures tend to require higher gain and, in some cases, slightly more diagnosis iterations for first identification, while preserving the overall diagnosability of the array.

For the 8-bit array, a favorable trade-off appears at $(V_{DD} = 0.6\,\text{V}, 0°\text{C})$, where the farthest location is identified using $(G = 60, t_s = 25\,\text{ns})$ in only six iterations. A similar trend is observed for the 16-bit array, where some PVT corners require substantially larger gain values while the iteration count changes only marginally. These results indicate that adaptive gain tuning is the dominant mechanism for compensating PVT-induced overlap among ADC outputs, while t_s provides secondary refinement to improve separability.

To extend diagnosis to longer MLs, we apply a partitioning-based diagnosis strategy. The ML is divided into multiple chunks, each consisting of, e.g., 16 bits, and each chunk undergoes the diagnosis flow consecutively. To realize this chunking, algorithmic-level masking is applied to the test pattern by assigning “don’t care” values to all other chunks using binary ‘0’ values for both Q and $\bar{Q}$. This process continues until all chunks complete

Table 5.4.: DfT-Induced Overhead in ReCAM Performance Metrics (Baseline: No-DfT Design for Power and Area, March Test for Latency)

ReCAM Array Size			Test Power (W)			Area (mm^2)			Test Latency (ns)		
N_{ML}	N_{CAM}	N_{ADC}	No DfT	With DfT	Ov. (%)	No DfT	With DfT	Ov. (%)	March Test	DfT	Imp.
16	16	1	0.203	0.216	6.3	0.0950	0.0953	0.36	64	112	2.29x
	64	1	0.204	0.216	5.8	0.0950	0.0954	0.36	256	448	2.29x
	16	4	0.203	0.216	6.3	0.3800	0.3810	0.36	16	28	2.29x
	64	4	0.204	0.216	5.8	0.3800	0.3810	0.36	64	112	2.29x
64	16	1	0.814	0.865	6.3	0.0950	0.0954	0.36	256	448	2.29x
	64	1	0.818	0.865	5.8	0.0951	0.0954	0.36	1024	1792	2.29x
	16	4	0.814	0.865	6.3	0.3800	0.3810	0.36	64	112	2.29x
	64	4	0.818	0.865	5.8	0.3800	0.3810	0.36	256	448	2.29x

**The power overhead is calculated for one DfT iteration.*

diagnosis. As shown in Table 5.3, this hierarchical approach significantly reduces the required gain (approximately 80% in larger arrays, such as the 32- and 64-bit cases) while maintaining scalability in diagnosis iterations. Moreover, it exploits favorable voltage and temperature operating points to accelerate diagnosis, enabling robust and scalable fault localization for long MLs. Across all evaluated PVT corners, diagnosability is preserved, while PVT mainly shifts the required sensing operating point rather than causing diagnosis failure.

5.6.3. Complexity and DfT Overhead Analysis

To place the proposed diagnosis framework in perspective, we analyze its timing complexity and hardware overhead relative to conventional March-based diagnosis [115]. Because the ADC is a relatively large and power-intensive component, it is shared among multiple MLs. Let N_{ADC} denote the number of shared ADCs in the array, N_{ML} the number of MLs, and N_{CAM} the number of CAM cells connected to each ML. For conventional March-based diagnosis [115], the timing complexity scales as

$$O(N_{CAM}) \cdot \left(\frac{N_{ML}}{N_{ADC}}\right) \tag{5.8}$$

where the second factor captures time-multiplexed access to the ADC.

In contrast, the proposed diagnosis partitions the ML into chunks of 16 cells and requires at most nine diagnosis iterations in the worst case. By adaptively leveraging favorable voltage and temperature operating points, the number of iterations can be reduced to only seven iterations, further accelerating the diagnosis process, resulting in

$$O\left(\frac{7N_{CAM}}{16}\right) \cdot \left(\frac{N_{ML}}{N_{ADC}}\right) \tag{5.9}$$

Compared with a March-based diagnosis, this reduces timing complexity by approximately 2.29×, demonstrating that adaptive diagnosis significantly lowers localization latency even in the presence of PVT variations.

Table 5.4 summarizes the DfT-induced overhead in terms of area, test power, and diagnostic latency for different ReCAM organizations, including varying numbers of MLs, cells per ML, and shared ADC resources.

The reported overhead is calculated relative to a baseline ReCAM design without DfT support. Specifically, area and power overhead include the tunable OpAmp, feedback-trimming circuitry, and associated test-mode logic. As shown in Fig. 5.2, these blocks are enabled only during diagnosis through signal T_M; therefore, they introduce zero power overhead during functional operation.

Even in the worst-case configuration, assuming a short ML with only 16 connected CAM cells, the added overhead remains limited to only 0.36% area and 6.3% test power. For latency, the baseline comparison is the March-test diagnosis flow, against which the proposed approach achieves the aforementioned 2.29× reduction.

5.7. Summary

This chapter presented a DfT-based fault diagnosis framework for ReCAM arrays, addressing the challenge of accurate fault localization under process, voltage, and temperature variations. The proposed approach combines a tunable OpAmp and ADC to enhance the observability of ML voltages, whose differences are otherwise diminished by ML parasitics and further obscured by PVT-induced variations.

Based on this architecture, an iterative diagnosis methodology was developed that adaptively configures the OpAmp gain G and sampling time t_s until the faulty cell location is uniquely isolated. Experimental results demonstrated that faulty cells can be localized in only 2–8 diagnosis iterations, and up to 9 under extended PVT corners, achieving approximately 2.29× lower diagnosis latency compared to conventional March-based approaches. A key finding is that favorable voltage and temperature operating points can be leveraged to accelerate diagnosis, reducing both the required gain and the number of iterations relative to nominal conditions. Furthermore, for larger arrays, the proposed partitioning strategy reuses calibrated 16-bit diagnosis settings and reduces the required gain by approximately 80%, substantially lowering sensing overhead while maintaining diagnosability for ML widths up to 64 bits.

These benefits are achieved with negligible hardware cost, introducing only 0.36% area overhead and up to 6.3% test-power overhead. Overall, the proposed hierarchical PVT-aware diagnosis framework reduces fault localization complexity, exploits operating-condition awareness to improve diagnostic efficiency, and provides a practical path toward reliable fault diagnosis in future large-scale ReCAM arrays.

6. PV-ReCAM: Process Variation-Aware Testing of ReRAM-based Content Addressable Memory

After establishing defect-oriented testing, the thesis first studies fault observability and localization under PVT variations, then extends the test methodology itself to variation-aware fault detection. Despite the advantages of ReCAM architectures, ReRAM technologies are highly sensitive to PV due to immature fabrication processes and the stochastic nature of resistive switching. In addition to conventional hard defects caused by device and interconnect failures, PV introduces soft defects that alter the analog sensing behavior of ReCAM arrays. As a result, faulty and fault-free behaviors may overlap, making reliable testing significantly more challenging.

This issue limits the effectiveness of conventional March tests, which were originally designed for deterministic digital memories. PV-induced soft defects can distort sensing distributions and even compensate for hard defects, causing faulty cells to appear functionally correct and resulting in test escapes. Consequently, conventional March tests may incorrectly classify defective chips as fault-free, reducing test reliability and defect coverage.

To address these challenges, this chapter investigates the impact of PV on ReCAM fault behavior and proposes a generalized PV-aware March testing methodology that detects both hard defects and variation-induced soft defects. The proposed approach accounts for defect masking and compensation effects caused by PV, achieving robust fault detection with 100% defect coverage under process-variation conditions.

6.1. Motivation Example

6.1.1. Soft Faults

In ReCAM arrays, multiple cells share a common ML, causing PV effects to accumulate, which shifts the equivalent resistance and introduces decision failures—incorrect output usually takes place due to the PV, at the tails of distributions as shown in the overlap region in Figure 6.1.

Although no hard defects are present, accumulated PV shifts the ML voltage beyond the FF boundaries, causing defect-free cells to appear faulty. The susceptibility to PV-induced faults depends on the ML size. For a B-bit ML, it is crucial to identify the most vulnerable voltage distributions.

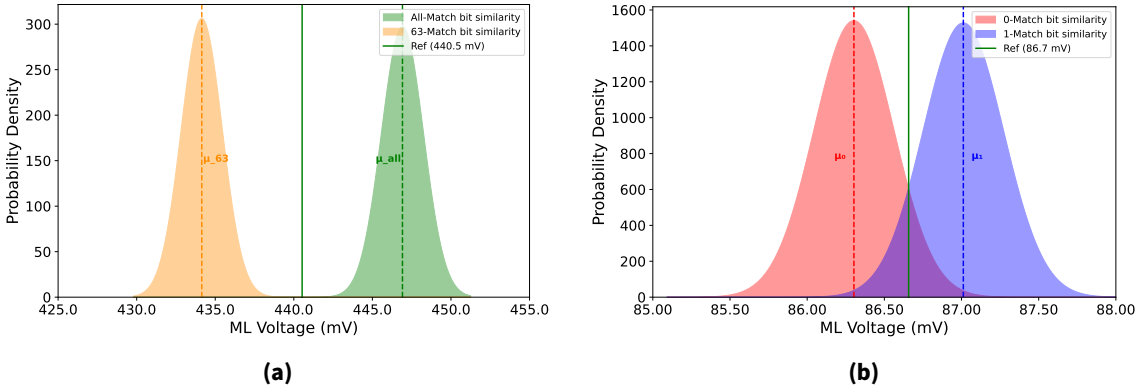


Figure 6.1.: Match-line (ML) voltage distributions for 64-bit ReCAM cells connected to a shared ML: (a) transition from 63-match to all-match, showing no distribution overlap; (b) transition from 0-match to 1-match, showing significant distribution overlap.

As shown in Figure 6.1(a), for a 64-bit ML, the most distinct or equivalently least PV vulnerable voltage separation occurs between all-match and $(B-1)$-match cases. Conversely, the least distinct, i.e., most PV vulnerable separation occurs between 0-match and 1-match scenarios, as seen in Figure 6.1(b), where significant overlap indicates a high probability of incorrect sensing; a reliability issue known as decision failure. Thus, even without hard defects, PV alone can undermine the reliability of analog similarity detection in ReCAM arrays.

6.1.2. Hard Defects

In contrast to PV-induced soft faults, PV masks hard defects. A soft fault induced by PV can compensate for hard defects, causing defective cells to escape detection by conventional March tests. This reveals key limitations of March-like tests under PV conditions. To investigate this, we inject PV by introducing random variations in ReCAM cells using randomly additional series or parallel resistances within the ReRAM stack. This alters electrical characteristics, allowing us to analyze the co-existence of PV-induced soft faults and hard defects.

For instance, if a hard stuck-at-LRS defect exists the right ReRAM (R_R), the ML voltage discharges faster. However, the accumulated PV in defect-free cells could tend to increase the equivalent resistance, compensating for this defect. As shown in Figure 6.2, under all-match input patterns (i.e. $P = Q =$ all 0s), a hard defect introduces minor voltage deviation. Yet, PV compensation restores the ML voltage to FF levels, masking the defect and invalidating March test detection. Thus, PV can obscure hard-fault signatures within specific resistance margins, rendering them undetectable by traditional testing approaches. However, our proposed method targets the adjacent Hamming levels that exhibit the non-equidistant voltage difference. Under certain conditions, PV cannot compensate for a hard defect. Thus, no masking effect exists, and both hard faults and PV-induced soft faults become observable. This is why our method achieves full coverage even when PV partially compensates for a defect.

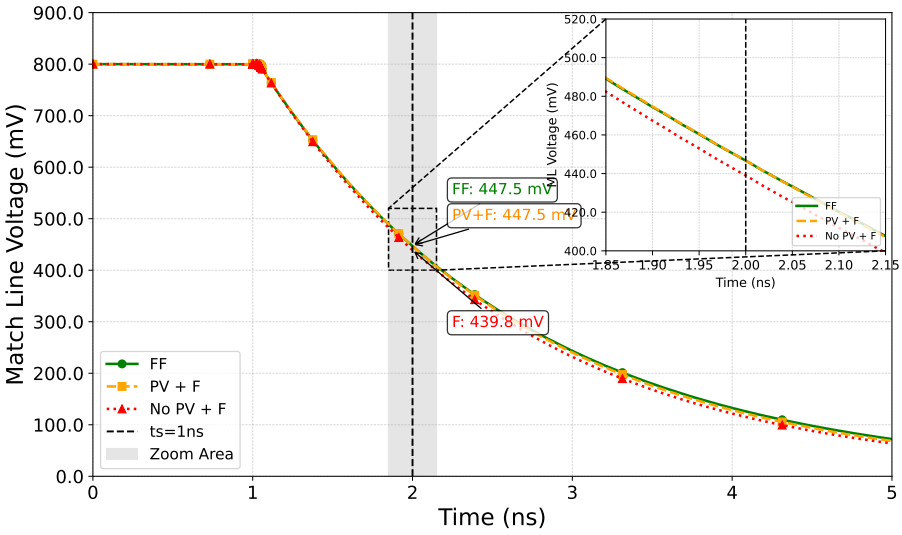


Figure 6.2.: Match Line (ML) voltage behavior over time for a 64-bit ReCAM under three conditions: (1) fault-free (FF), (2) with process variation (PV) plus a single R_{Right} stuck-at-LRS fault (F), and (3) with the same single fault (F) but without the PV with input combination ($P = Q = 0$)

6.2. Problem Statement

Unlike the standard memory, which performs read and write operations, the CAM performs data comparison tasks. Binary query patterns are compared against stored data, and the similarity is reflected in the analog ML voltage. Comparators then digitize this voltage.

To evaluate the limitations of conventional March tests under PV, we inject resistance variations into random ReCAM cells using a Monte Carlo approach. These altered values are sampled from distributions based on technology parameters, then the query patterns (e.g., 0-match, all-match) are applied to observe ML voltage deviations, as detailed in Table 9.1. These deviations reveal that PV can either mask defects or falsely indicate faults, depending on the threshold settings:

1. **Hard defect masked by PV**: Coexisting hard defects and PV may cancel each other, preventing detection by the March test.
2. **Misinterpreted as a hard defect due to PV only**: PV alone shifts ML voltage, causing a soft defect that behaves similarly to hard defects, so that we cannot distinguish between them.

6.3. Contributions

The main contributions of this chapter are summarized as follows:

- The impact of process variation on the functionality and testability of analog ReCAM architectures is analyzed, demonstrating how PV-induced soft defects alter match-line sensing behavior and invalidate conventional March testing assumptions.

- The interaction between hard defects and PV-induced soft defects is investigated, showing how compensation and masking effects result in test escapes and reduced defect coverage under conventional March tests.
- ReCAM fault behaviors under combined hard-defect and process-variation conditions are classified and modeled, highlighting the limitations of traditional deterministic testing approaches in analog CiM architectures.
- A generalized PV-aware March testing methodology is proposed that optimizes test patterns to activate and detect both hard defects and variation-induced soft defects.
- Variation-aware test conditions are developed to expose defect masking and compensation effects caused by process variation in ReCAM arrays.
- Experimental evaluation demonstrates that the proposed PV-aware March methodology achieves 100% defect coverage while significantly improving testing robustness under process-variation conditions.

6.4. PV Mitigation and ReCAM Testing

Previous works have addressed the challenges that PV causes in CiM architectures, but none provide an effective test methodology tailored to analog ReCAM arrays. In this regard, we can classify the related work as follows: (1) works that aim to mitigate the impact of PV at the circuit level, such as design migration from MOSFET to FinFET while accounting for PV [116], and at the software level, which mitigates PV-induced errors through hardware-software co-design and neural network retraining techniques [117].

(2) Works that propose testing solutions. Within this category, there are two subclasses: (2.1) Testing for conventional CAMs, such as [118], [119], which focus on SRAM-based architectures and target static faults during memory or logic operations. (2.2) Testing for CAMs with CiM functionality [81], which typically focus on hard faults such as stuck-at or transition faults. However, these approaches fall short when PV-induced analog variations are present, as they fail to capture soft faults from PV.

6.5. Testing for PV-induced Soft Faults

Existing March tests for ReCAM often fail to detect soft faults caused by PV or its effect with hard defects. Isolating PV-induced faults remains difficult due to the accumulation of PV effects at the array output. This section analyzes PV's impact on ReCAM functionality and presents testing methodologies to overcome these limitations.

6.5.1. Analyzing PV Impact on Match Line

We perform electrical-level simulations, using SPICE, to assess how PV affects various test conditions. A range of similarity patterns—from all-mismatch to all-match—is applied to

Table 6.1.: Simulation Setup and Parameters

Category	Parameter Details
Technology-Level Parameters	
ReRAM Model [23]	JART VCM Read Variability
Filament Radius	45 nm
Disc Region Length	0.6 nm
N_{init} (HRS, LRS)	9 m, 3×10^{26} m^{-3}
N_{init} (Undefined Region) [94]	65 m, 1×10^{26} m^{-3}
N_{init} (Deep HRS, Deep LRS) [94]	5 m, 1 k, 1×10^{26} m^{-3}
Circuit-Level Parameters	
Simulation Tool	Cadence Virtuoso
CMOS Technology	GlobalFoundries 22FDX
Supply Voltage (V_{DD})	800 mV
Temperature	27°C
Interconnect per ReCAM Cell (22 nm) [47], [48]	ML: $R = 11.95\,\Omega$, $C = 16.63$ aF
Match-Line Length	8-, 16-, 32-, and 64-bit
Variation Assumption	$\sigma = 0.003\mu$
Performance and Area Parameters	
CAM Cell Area [108]	0.0242 μm^2
Array Power	0.108 mW
Sampling Time	2 ns
Monte Carlo Samples	1,000

identify the most PV vulnerable combinations which lead to the largest overlap area, and hence higher probability of decision failure.

Figure 6.3 presents (a) the ML voltage and (b) (ΔV_{ML}) trends for 16-bit, 32-bit, and 64-bit ReCAM arrays, highlighting both the most and least distinct voltage distributions relevant for reliable testing.

The susceptibility to PV-induced soft faults depends on two key factors: the voltage distance between adjacent match levels and the standard deviation of the distributions, both of which contribute to decision failure. However, ΔV_{ML} plays a dominant role compared to the standard deviation. Specifically:

- **16-bit ReCAM array:** The ML voltages increase nearly linearly with the number of matching input bits, resulting in approximately equidistant voltage levels and relatively uniform distinguishability between match levels.
- **32-bit ReCAM array:** The voltage differences become slightly nonlinear. Some match levels are more closely spaced than others, resulting in unequal sensitivity to PV throughout the ML; in particular, the least distinct levels occur at intermediate match levels.

- **64-bit ReCAM array:** A non-equidistant trend becomes more evident. The aggregation of more activating cells causes a steeper voltage drop along the ML, making early match transitions (such as 0,1-match) more ambiguous.

  In contrast, the most distinguishable case occurs between the all-match and 63-match inputs.

These observations demonstrate that the number of cells connected to a match line significantly affects its susceptibility to PV-induced faults.

In addition, the location of this 1-match bit within the array significantly affects the ML voltage. To analyze this location dependence, we conduct simulations in which the 1-match bit occurs at different positions: at the beginning (farthest from the comparators), in the middle, and at the end (closest to the comparators). The results indicate that in the most PV prone distributions, the detection scenario occurs when the 1-match bit is located in the beginning, farthest from the comparators. This location yields the smallest voltage difference, whereas placing the 1-match bit at the end of the array results in a more distinct difference, as illustrated in Figure 6.4.

Based on this analysis, a promising solution is to design a specialized test pattern to detect the least-distinct inputs most affected by PV. In other words, we activate the match scenarios, causing minimal changes in ΔV_{ML} and, hence, the most susceptible cases to PV.

The methodology concept consists of the following steps: (1) To efficiently test the PV susceptibility, in ReCAM, we analyze the ML voltage distributions for different numbers of ReCAM cells connected to the ML. (2) In an ML with a length of B, there are $B + 1$ possible matching bit combinations (from 0-matches to B-matches). For each of these combinations, we obtain the distribution of the ML voltage, which is in good agreement with a normal distribution. To distinguish between these distributions, for each pair of consecutive distributions, we set the reference voltage V_{ref} at the midpoint of the means of the two consecutive distributions. The smallest ΔV_{ML} indicates the highest probability of decision failure, due to minimal voltage separation between adjacent Hamming levels. In contrast, the largest ΔV_{ML} corresponds to clearly separated distributions that are less failure-prone to distinguish. Based on Monte Carlo simulations with μ/σ ratio listed in Table 9.1, the minimum detectable shift in ML voltage is approximately 0.5–0.6 mV for the 64-bit array (Fig. 5), which corresponds to a 4–6% change in the effective equivalent resistance of the ML discharge path. Therefore, any PV-induced variation larger than this threshold is guaranteed to be detected.

6.5.2. ReCAM March and Notations

Each ReCAM cell performs two primary operations: WRITE and COMPARE. During the WRITE operation, the prototype data pattern P is programmed into the ReCAM cells, setting each cell to either the LRS or HRS. In the COMPARE operation, the V_{ML} is measured and compared against a sufficient number of V_{ref} using comparators, to finally identify the number of similar bits. This comparison helps detect the minimum voltage difference between adjacent similarity inputs.

The following notation is used to describe the algorithmic procedures:

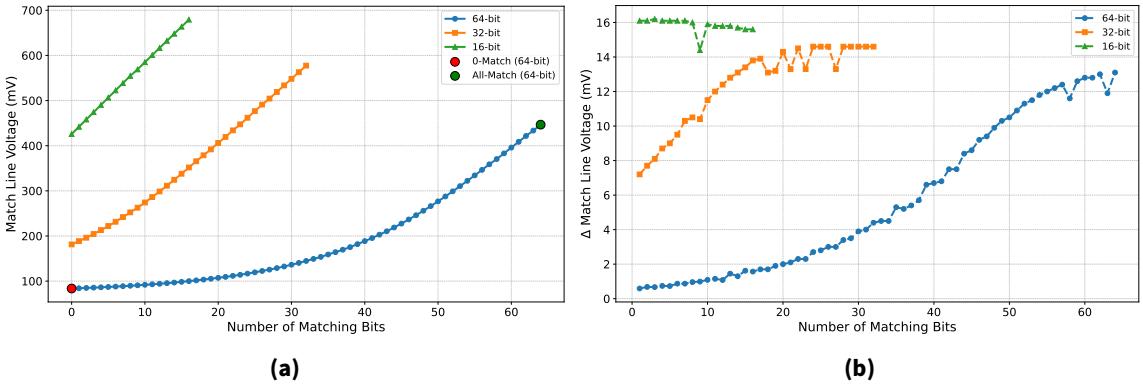


Figure 6.3.: (a) Match Line (ML) voltage levels for different similarity patterns in 64-, 32-, and 16-bit ReCAM arrays, and (b) corresponding non-equidistant voltage differences (ΔV_{ML}), highlighting the 0- to 1-match transition as the least distinct case.

- $w_P\mathbf{d}$: Write the prototype data pattern d (e.g., all-0s or all-1s) into the ReCAM cells, setting each ReRAM element to either LRS or HRS accordingly.
- $c_Q\mathbf{d}$: Compare the outputs of V_{ML} with the V_{ref} values using a specific search data pattern d for the query Q these patterns are walking-0 or walking-1.

As observed from electrical-level simulations, the most failure-prone voltage distributions due to PV differ significantly depending on the array size (B). Specifically, for 64-bit and 32-bit ReCAM arrays, the smallest voltage margin ΔV_{ML} and therefore the highest failure probability occurs between the 0-match and 1-match input patterns as shown in Figure 6.3(b). This makes these cases appropriate for sensitizing PV-induced soft faults.

For smaller arrays, however, this critical pair of patterns shifts. For example, in the 16-bit ReCAM, the highest sensitivity to PV is observed near the center of the match range, particularly between the 8-match and 9-match inputs (i.e., $B/2$ vs. $B/2+1$ match) as shown in Figure 6.3(b). Meanwhile, in the 8-bit array, the most failure-prone case occurs between the 7-match and the all-match (8-match) patterns (i.e., B vs. $B-1$ match). Therefore, accurately identifying the adjacent match configurations with the smallest ΔV_{ML} is crucial for effective PV fault sensitization. This observation motivates the development of size-dependent, adaptive PV-aware March-like algorithms that dynamically adjust their test targets based on array size and the observed voltage distribution behavior.

6.5.3. Size-dependent PV-Aware March-like Test Algorithm

The proposed March-like test algorithm is designed to detect all PV-induced soft faults within the ReCAM array. In the ReCAM operation, since the ML voltage reflects the number of similar bits between the prototype and query, the absolute values of the bits usually do not matter, and only the number of matching bits does. However, the actual values under PV influence the ML voltage margins. Therefore, to maximize PV fault activation, we identify the prototype and query combinations that result in the smallest voltage margins and use those for testing. The same sequence is then repeated across the entire array to create a March-like test pattern. If a cell passes under the smallest margin

Table 6.2.: Comparison of Fallout, Detected Faults, and Test Coverage for Different ReCAM Array Sizes

#ReCAM Cells	ML Size	Fallout Rate	Detected Faults		Coverage	
			PV-Aware Test	ReCAM March [115]	PV-Aware Test	ReCAM March [115]
1×10^{11}	8-bit	1.37×10^{-4}	1.37×10^{7}	4.48×10^{5}	100%	100%
1×10^{11}	16-bit	5.00×10^{-1}	5.00×10^{10}	6.31×10^{1}	100%	1.26×10^{-7}%
1×10^{11}	32-bit	1.30×10^{-5}	1.30×10^{6}	9.70×10^{0}	100%	7.46×10^{-4}%
1×10^{11}	64-bit	1.22×10^{-1}	1.22×10^{10}	5.10×10^{4}	100%	4.20×10^{-4}%

Fallout refers to the ratio of chips failing the test, i.e., chips with at least one detected fault.

condition, it is guaranteed to pass under any condition with a larger margin. Thus, the algorithm guarantees complete coverage of PV-induced faults with minimal test effort.

The test algorithm involves multiple March elements (ME) that progressively write prototype data and compare V_{ML} values with a predefined set of V_{ref} to detect and identify the least distinct distribution, i.e., the pair with the smallest difference between their voltage means (μ). Each sequence of these elements targets a different input matching scenario. Thus, the PV-aware March-like test is applied to longer ML sizes (e.g., 32, 64, or more), where 0-match and 1-match distributions are hardest to distinguish. The operation sequence is as follows:

ME1: (w_P **0**, c_Q **1**) – Write all cells with `0`, then confirm by comparing with `1` to ensure a 0-match across the entire array.

ME2: 1-match input patterns from the first bit to the last bit of the array:

- c_Q **0 1 1 ... 1** – Apply 1-bit match at location 0,
- c_Q **1 0 1 ... 1** – Apply 1-bit match at location 1,
- c_Q **1 1 0 ... 1** – Apply 1-bit match at location 2,
- Continue this walking-0 pattern until c_Q **1 1 1 ... 0** having 1-bit match at end of the array.

Conversely, for the shorter MLs, i.e., 8 for which the least distinct distributions happen for all match and 1-mismatch, we propose the following sequence:

ME3 (w_P **1**, c_Q **1**) – Write and confirm all cells are `1`, then compare with `1` to ensure an all-match scenario.

ME4 0-match input patterns from the last bit to the first bit of the array:

- c_Q **1 ... 1 0** – Apply 1-bit mismatch at location 63,
- c_Q **1 ... 0 1** – Apply 1-bit mismatch at location 62,
- c_Q **1 ... 1 1** – Apply 1-bit mismatch at location 61,
- Continue this pattern until c_Q **0 1 1 ... 1** – having 1-bit match in the first of the array.

Please note that the ML size is a known parameter for the built-in self-test (BIST) unit, which can determine the proper sequence.

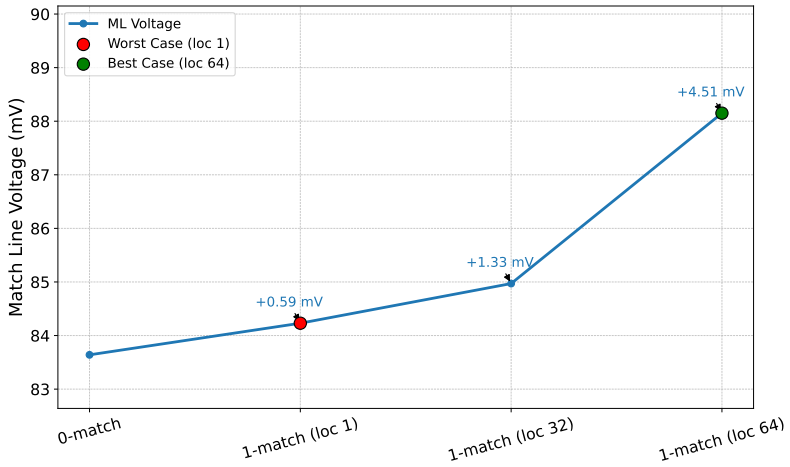


Figure 6.4.: Sensitizing and performing ME2 across a 64-bit ReCAM array to show the ML voltage trends by placing 1-match bit at different positions in the ReCAM array (first, middle, and end), versus the 0-match case. The red circle highlights the 1-match in the first cell, indicating the smallest voltage difference. Annotations indicate the differences (ΔV_{ML}) between consecutive 0-match and 1-match input combination

For longer MLs, e.g., $B = 64$, we use two March elements (ME1, ME2): one to confirm the 0-match pattern consistently, and the other as a walking 1-bit match element to identify this pair of least distinct voltage distributions corresponding to the highest decision failure probability and smallest ΔV_{ML}. Thus, the test begins by programming P to 0s, followed by comparing it with all 1s in Q to establish the baseline of V_{ML} for a 0-match scenario, sensitized by ME1.

To identify the least distinct distribution, we apply a walking-0 pattern to Q, where a single 0 is placed at varying bit positions against an otherwise all-ones input (i.e., a 1-match input). This pattern is applied from the first cell in the array, farthest from the comparators, to the last cell in the array of e.g., size 64 cells connected to an ML which ME2 sensitizes.

As shown in Figure 6.4, the farthest cell from the comparators is prioritized for PV testing, as it exhibits the minimum ΔV_{ML} for adjacent distributions across the entire array. The resulting V_{ML} is compared against a predefined reference voltage V_{ref} at each step. To identify the most vulnerable locations—where a minimal ΔV_{ML} could be misinterpreted as another distribution. This comparison helps determine the position at which applying a 1-bit match yields the smallest voltage difference between the two selected distributions.

To ensure a comprehensive testing algorithm for various ML sizes, we propose ME3 and ME4, which apply to smaller arrays, such as 16-bit and 8-bit. The key difference between ME1 and ME3 lies in their baselines: ME1 uses 0-match as the baseline, whereas ME3 uses all-match Additionally, the primary difference between ME2 and ME4 lies in the walking element: ME2 employs a 1-bit match, whereas ME4 utilizes a 1-bit mismatch. This walking position uniquely identifies the defective cell, enabling per-cell localization even though the ML is shared. Therefore, the test is applied to each ML in the array, providing bit-level fault localization within each ML.

Moreover, For $B = 16$, the most PV prone cases occur between patterns with $B/2$ and $B/2 + 1$ matches, i.e., 8-match and 9-match. To target this critical region, similarly, ME5

can be inherited from ME1 or ME3 by having $B/2$ or $B/2 + 1$ (i.e., 8 or 9-match input). At the same time, ME6, inherited from ME2 or ME4, generates either a 1-bit mismatch or a 1-bit match to achieve the least distinct distributions.

Also, for physical defects, ME1 and ME4 target conditions in which expected matches are misinterpreted as mismatches, while ME2 and ME3 detect cases in which expected mismatches are incorrectly interpreted as matches.

This universal testing methodology accounts for both physical defects and PV, ensuring 100% defect and PV coverage and offers a scalable, location-aware approach to detect and localize PV-induced faults in ReCAM arrays.

6.5.4. Evaluation

To quantitatively evaluate the impact of PV on fault detection, we analyzed the fallout obtained and test coverage across various ReCAM array sizes, as summarized in Table 6.2. Fallout refers to the proportion of chips that fail the test (i.e., if any test within a chip fails, the entire chip is considered failed). To calculate the fallout, we first quantify the overlap region, which is the likelihood of incorrect results. If $B = 64$, the largest overlap happens between 0-match and 1-match distributions as illustrated in Figure 6.1(a). Let μ_0, σ_0 and μ_1, σ_1 denote the means and standard deviations of the 0-match and 1-match distributions, respectively. Let $V_{\text{ref, } 0\to1}$ represent the decision threshold used to distinguish between them. The probability that a 0-match input is interpreted incorrectly as a 1-match is given by:

$$P_{e,0} = 1 - \text{CDF}\left(\frac{V_{\text{ref, } 0\to1} - \mu_0}{\sigma_0}\right) \tag{6.1}$$

$$P_{e,1} = \text{CDF}\left(\frac{V_{\text{ref, } 0\to1} - \mu_1}{\sigma_1}\right) \tag{6.2}$$

where $V_{\text{ref, } 0\to1} > \mu_0 \text{and} \mu_1$

$$P_{Fallout} = \max(P_{e,0}, P_{e,1}) \tag{6.3}$$

The overall fallout is defined by (3) as the maximum probability of incorrect interpretation of (1) and (2). This metric is essential for evaluating the reliability of decision thresholds in the presence of PV.

A key insight from Table 6.2 is that for $B = 8$, the conventional March test achieves the same fault coverage as our proposed PV-aware test. This is because the least distinct distributions at $B = 8$ are already accounted for in the conventional March algorithm.

However, for longer MLs, such as $B = 16$, where the least distinct distributions occur, i.e., near the middle of the match range, and for $B = 64$, which occurs between 0-match and 1-match. In these challenging cases, our PV-aware test achieves full fault coverage, outperforming traditional March tests.

Our proposed PV-aware testing approach consistently achieves 100% defect coverage across all array sizes because it tests for the minimum ΔV_{ML} in two match cases. In contrast, while the conventional March test may be effective under no-PV conditions, it proves ineffective in the presence of PV, often failing to detect faults altogether. Please

Table 6.3.: Comparison of Testing Algorithms for Related Works

Test Algorithm	Write Ops	Compare Ops	Defect Coverage
March C-8 (8T SRAMs) [118]	4	$12n_{\text{rows}}$	PV-induced only
March IMC-8T [119]	6	$\approx 12n_{\text{rows}}$	PV + Hard
Conventional CAM [81]	$8n_{\text{rows}}$	$2n_{\text{rows}} + 3B + 3$	Hard only
ReCAM March [115] (5-ME)	3	$6n_{\text{rows}} + 2n_{\text{rows}}B$	Hard only
Proposed PV-Aware (2-ME)	1	$2n_{\text{rows}}B$	PV + Hard

note that although smaller ML sizes exhibit lower fallout rates, longer MLs are ultimately more favorable due to the greater amount of analog data flowing through them, thereby enhancing overall performance and energy efficiency.

6.5.4.1. Assessment of time complexity

The proposed test algorithm sensitized by two size-dependent March Elements (ME1, ME2) and (ME3, ME4), requiring one write operation and n-bit comparisons per ML where B is the number of connected cell to the ML, totaling one write and $2 \times n_{rows} \times B$ comparisons for 100% coverage of PV-induced soft defects, where n_{rows} indicates the number of MLs in a ReCAM array.

As listed in Table 6.3 and detailed in [115], the conventional March Test requires 3 write operations and $6 \times n_{rows} + 2 \times n_{rows} \times B$ comparisons, focusing only on hard defects and becoming less efficient with increasing n_{rows} and B. The proposed PV-Aware test is more efficient, requiring only one write operation and $2 \times n_{rows} \times B$ comparisons, specifically targeting both PV-induced soft defects and hard defects.

6.6. Summary

This chapter investigated the impact of process variation on ReCAM functionality and testing reliability. We demonstrated that PV-induced soft faults can distort analog sensing behavior and even compensate for physical hard defects, leading to fault masking and test escapes under conventional March testing methodologies. As a result, traditional ReCAM March tests become ineffective in reliably detecting faults in the presence of PV, significantly reducing defect coverage.

To address these limitations, we proposed a generalized PV-aware March testing methodology capable of detecting both hard defects and variation-induced soft faults. The proposed approach optimizes test patterns while accounting for PV-induced masking and compensation effects, enabling reliable fault activation and detection under process-variation conditions. Experimental results demonstrated that the proposed PV-aware methodology achieves 100% defect coverage while significantly improving the robustness and reliability of testing for analog ReCAM architectures.

7. Post-Manufacturing Calibration for ReRAM-based Content-Addressable Memory

During manufacturing, process variations introduce silicon mismatches that significantly impact the reliability and yield of emerging non-volatile memory technologies. As illustrated in Fig. 7.1, these variations are categorized into *global variations*, which affect the entire wafer, and *local variations*, which occur within individual dies due to device-level mismatches [120]. A comprehensive variation model that distinguishes between global and local variations is therefore essential for achieving robust design and competitive manufacturing yield in emerging NVM technologies, as each variation type affects circuit performance and yield differently.

Conventional pass/fail manufacturing tests rely on fixed decision boundaries and therefore cannot account for chip-to-chip variability. As a result, chips affected mainly by global variation may be incorrectly rejected despite being functionally correct, leading to unnecessary yield loss. To address this issue, this chapter proposes a post-manufacturing calibration methodology integrated into the ReCAM test flow. The proposed approach adaptively determines the boundaries between similarity levels (Hamming distances) for each chip using a lightweight LUT-based calibration scheme. By compensating for global variation on a per-chip basis, the method reduces unnecessary chip fallout and achieves up to an 18% yield improvement under PV conditions.

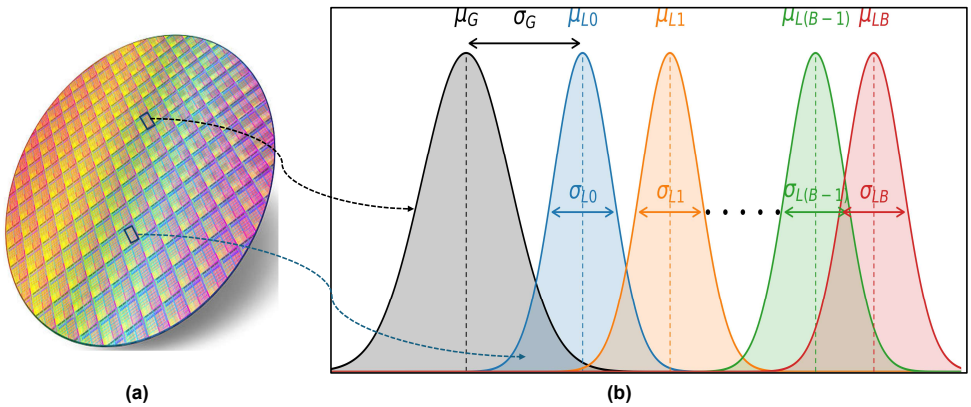


Figure 7.1.: (a) and (b) Illustration of global (die-to-die) and local (within-die) variations across a wafer [120]. The variation between different dies is characterized by the global variation σ_G, whereas variations within each die are distributed around the die median due to local variation σ_L. These variations affect all similarity levels in the ReCAM array, ranging from 0-match to B(all)-match.

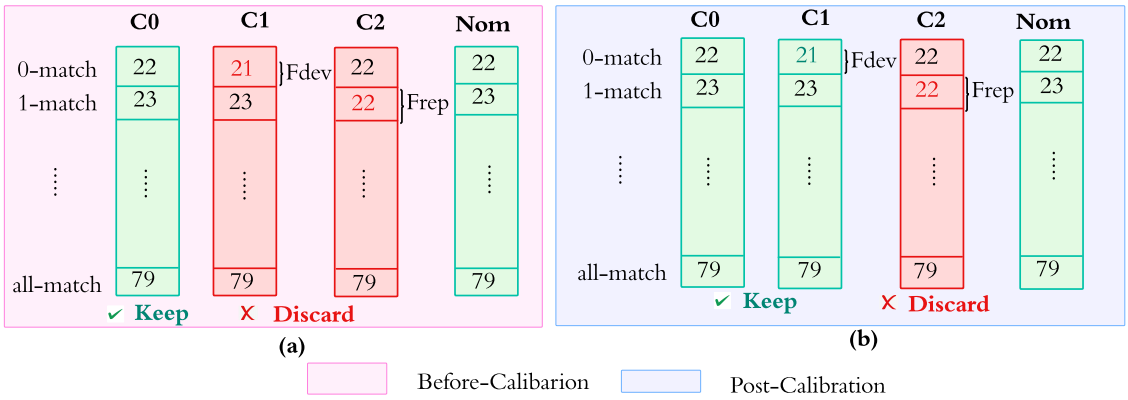


Figure 7.2.: Calibration (a) before and (b) the proposed one for different chip instances compared with Nominal (Nom) one

7.1. Motivation Example

Figure 7.2 illustrates the impact of process variation on conventional ReCAM testing. In the pre-calibration case shown in Fig. 7.2a, the nominal ADC outputs corresponding to the 0-match and 1-match patterns are 22 and 23, respectively. Due to global variation, these ADC values may shift for a given chip, for example, 21 for 0-match and 22 for 1-match, while still remaining distinguishable from each other. However, conventional testing relies on fixed reference values and would therefore incorrectly reject such chips even though their functionality is preserved.

To address this issue, the proposed calibration approach shown in Figure 7.2b determines the similarity boundaries individually for each chip using a lightweight LUT-based mapping. This allows the shifted ADC outputs to be correctly interpreted according to their corresponding similarity levels, thereby preserving the functionality of chips affected by global variation.

In contrast, when local variation causes adjacent similarity levels to overlap within the same chip, for example, when both 0-match and 1-match produce the same ADC value, the two levels become indistinguishable. In such cases, reliable sensing is no longer possible, and the chip must be rejected.

Therefore, the proposed post-manufacturing calibration incorporates a lightweight LUT whose number of entries equals the number of ReCAM cells connected to the ML, enabling per-chip mapping between ADC outputs and their corresponding similarity levels.

7.2. Problem Statement

PV in ReCAM arrays can be broadly classified into local and global variations [121], [122]. Practically, an important source of these variations in ReCAM arrays is the time-dependent drift of ReRAM devices, as the resistance is controlled by applying an electric field; once the field is removed, the resistance relaxes, leading to deviations in resistance values. Local variation arises from device-to-device mismatches within the same die, which

can reduce the voltage or ADC margin across different Hamming distances, potentially causing overlaps in similarity levels within the same chip. When two distinct match levels produce identical ADC outputs, the similarity computation becomes ambiguous, leading to unavoidable functional failures. During manufacturing tests, detecting these cases results in chip fallout. In contrast, global variation appears as a uniform shift in ADC outputs across an entire chip, caused by wafer-level non-uniformities or systematic process skew [116], [123]. Importantly, as long as the ADC ranges across different similarity levels remain non-overlapping, the correct similarity level remains uniquely identifiable and can be fully recovered through per-chip calibration.

7.3. Contributions

The main contributions of this chapter are summarized as follows:

- The impact of both global (die-to-die) and local (within-die) process variations on the sensing behavior and manufacturing yield of analog ReCAM arrays is analyzed.
- It is demonstrated that conventional fixed-boundary pass/fail testing can incorrectly reject functionally correct chips under global variation, leading to unnecessary yield loss.
- A post-manufacturing calibration methodology integrated within the ReCAM test flow is proposed to compensate for chip-to-chip variation.
- A lightweight LUT-based calibration scheme is introduced that adaptively determines similarity-level boundaries on a per-chip basis.
- A distinction is drawn between recoverable global shifts and non-recoverable local overlap between adjacent similarity levels, providing insight into the fundamental yield limitations of ReCAM sensing.
- The proposed methodology is evaluated under process-variation conditions, demonstrating up to 18% yield improvement over conventional testing approaches.

7.4. Variation aware testing in ReCAM arrays

Previous work has explored the challenges of variation and reliability in ReRAM-based computing architectures. However, a key shortcoming is that existing test schemes for analog ReCAM are largely variation-agnostic; they neither model nor mitigate the impact of global or local PV. As a result, the dominant component global variation remains unaddressed, causing functional chips to be unnecessarily discarded and leading to avoidable yield loss. This highlights the need for testing approaches that explicitly account for the mitigation of these variations.

The related literature includes techniques that mitigate PV at the circuit level, such as design migration across technologies that accounts for PV [116], as well as software-level

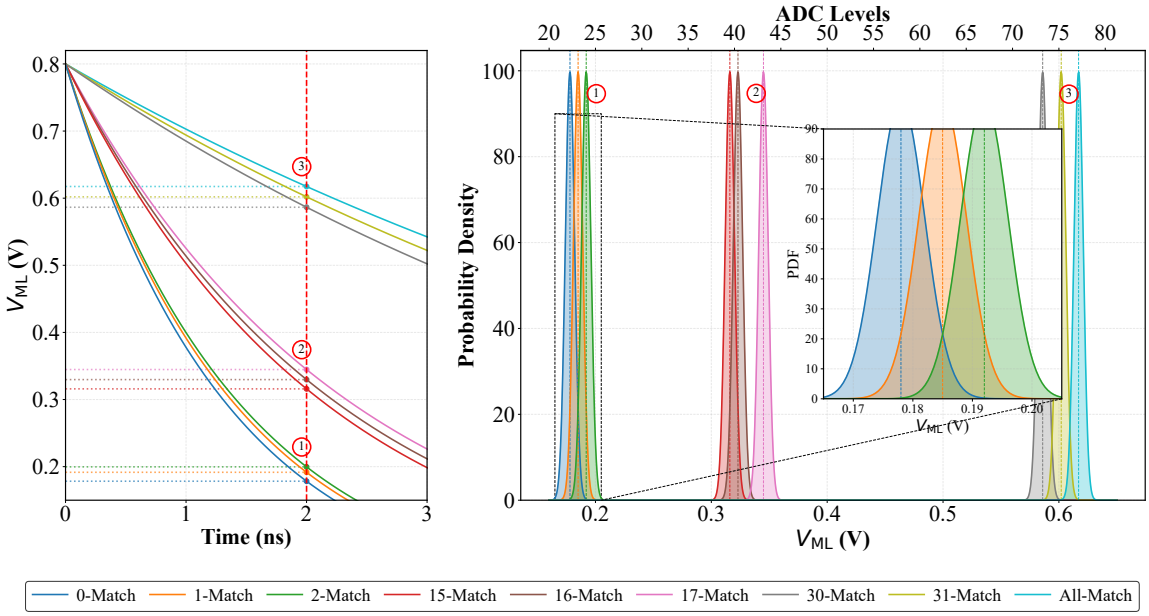


Figure 7.3.: (a) ML voltage decay for low, mid, and high match groups and (b) their voltage distributions for Groups 1–3 with ADC mapping shown on the top axis. The inset zooms in on the closely spaced low-match group 1 (0–2 match similarity level)

approaches in which PV-induced deviations are reduced through hardware–software co-design and neural-network retraining [117]. While effective in their respective contexts, these techniques do not address post-manufacturing test requirements for analog CiM arrays. Other studies focus on testing methodologies. Digital CAM testing [119] primarily targets SRAM-based architectures and static faults, while analog CAM testing [81] typically addresses hard faults such as stuck-at or transition faults. Additional works propose test and post-manufacturing strategies for ReRAM devices [94], [124]; however, these approaches do not incorporate process-variation awareness for analog ReCAM arrays.

Overall, none of the existing methods perform PV-aware testing for analog ReCAM, leaving global variation unmitigated and resulting in unnecessary loss of functional chips. Addressing this gap motivates the development of the global-variation-aware test methodology proposed in this work.

7.5. Proposed calibration methodology within ReCAM Array

7.5.1. Impact of local and global variations on ReCAM

To demonstrate the impact of PV on yield, we present two examples that highlight the distinct effects of local and global variations in ReCAM arrays. These examples show how PV can cause functional chips to be classified as faulty and unnecessarily discarded, contributing directly to yield loss. These experiments are conducted using detailed electrical-level simulations with SPICE tools. To ensure a realistic electrical level simulation setup, the MLs are modeled with resistive-capacitive (RC) parasitics to more closely reflect actual circuit behavior. This simulation follows the setup described in Table 9.1. ReRAM demonstrates a

Table 7.1.: Simulation setup and parameters.

Category	Parameter Details
Technology-Level Parameters	
ReRAM model [23]	JART VCM read-variability model
Filament radius	45 nm
Disc region length	0.6 nm
N_{init} for HRS/LRS	9 m / 3×10^{26} m^{-3}
N_{init} for undefined region [94]	65 m / 1×10^{26} m^{-3}
N_{init} for deep HRS/deep LRS [94]	5 m / 1 k / 1×10^{26} m^{-3}
ReRAM HRS resistance state [125]	$\mu = 105.51$ kΩ, $\sigma = 19.495\%$
ReRAM LRS resistance state [125]	$\mu = 1.975$ kΩ, $\sigma = 5.472\%$
Circuit-Level Parameters	
Simulation tool	Cadence Virtuoso
CMOS technology	GlobalFoundries 22FDX
Supply voltage, V_{DD}	800 mV
Temperature	27°C
ADC [95]	8-bit resolution
Interconnect per ReCAM cell [47]	ML resistance: 11.95 Ω; ML capacitance: 16.63 aF
ReCAM array size	32 cells/ML
Monte Carlo samples	1,000
Performance and Area Parameters	
CAM cell area [108]	0.0242 μm^2
ADC area	0.095 mm^2
LUT area	22.5 μm^2
CAM cell power	0.149 mW
ADC power	50.8 mW
LUT power	2.95 mW
Sampling time	1.00 ns
ADC latency	1.21 ps
LUT latency	112 ps

bit-based programmability advantage but often falls short in large arrays due to stochastic variations, such as drift and noise [126]. To model this, we assume that the deviation in Resistances follows a Gaussian distribution [127], where $R_{\text{HRS,LRS}} \sim \mathcal{N}(\mu_{\text{HRS,LRS}}, \sigma^2_{\text{HRS,LRS}})$. As a result, the resistance values of R_{HRS} and R_{LRS} can differ at the same similarity level for each chip.

Figure 7.3 illustrates the relationship between ML voltage decay and the corresponding distributions for different match levels in a 32-bit ReCAM array in the presence of local variations. The left plot shows the exponential decay of ML voltage over time, where higher match levels exhibit slower discharge due to larger equivalent resistance, resulting in higher voltages at the sampling time (red dashed line). The right plot presents the corresponding probability density functions (PDFs) of ML voltages sampled at t_s, with the top axis indicating the mapped ADC levels. Each distribution corresponds to a specific match level, and the numbered circles (1 to 3) link the voltage decay curves on the left to their corresponding distributions on the right. Group 1 represents the low-similarity match levels (0–2 bit match), Group 2 corresponds to the mid-similarity range (15–17 bit match), and Group 3 covers the high-similarity levels (30–all bit match). The zoomed-in inset highlights Group 1, where the voltage separation is minimal and the distributions significantly overlap, increasing the likelihood of incorrect ADC interpretation. In contrast, Group 3 shows a clear margin between distributions with negligible overlap, allowing ADC values to be distinguished and exhibiting a degree of tolerance to local variations. During local variation simulation, each chip instance uses a different set of random variables. Therefore, as the array size increases, the local mismatch effects across devices tend to cancel within each chip instance, particularly in long-path circuits.

Figure 7.4 illustrates the ML voltage decay and the corresponding ADC values across multiple chips and match levels in a 32-bit ReCAM array under global variation. The left plot shows the exponential decay of ML voltage for various match levels, with a zoomed-in inset focusing on high match levels (e.g., 30-match to all-match), where the voltage separation is minimal, potentially leading to overlapping ADC outputs. The right plot presents the distribution of ADC values for each match level across different corners. As a result, high match levels, such as the all-match case, exhibit wider ADC spreads, whereas the 0-match pattern converges toward a single ADC value. These results highlight how global variation can cause ADC range overlaps between match levels, potentially leading to the chip being discarded. In global variation simulation, all devices share the same set of random numbers across all chip-instance runs, so it can be recovered.

7.5.2. Mapping ADC codes to similarity levels

To mitigate yield loss in ReCAM arrays caused primarily by global variations, we propose a post-manufacturing calibration flow that establishes the digital correspondence between ML voltages, measured through ADC outputs, and their associated similarity levels. Specifically, a lightweight LUT maps the ADC output to the corresponding Hamming distance used in similarity computation on a per-chip basis. The mapping is generated by applying a complete set of similarity patterns, ranging from 0-match to all-match, in order to capture the full range of ML behavior and the resulting ADC variations across different chips. The proposed flow stores the valid ADC ranges for each chip's similarity level, enabling adaptive interpretation of similarity during operation.

In a conventional testing flow without calibration, two major variation-induced failure scenarios can occur. First, die-to-die global variation may shift the ADC distributions of an otherwise functional chip away from the nominal reference values. Second, local variation

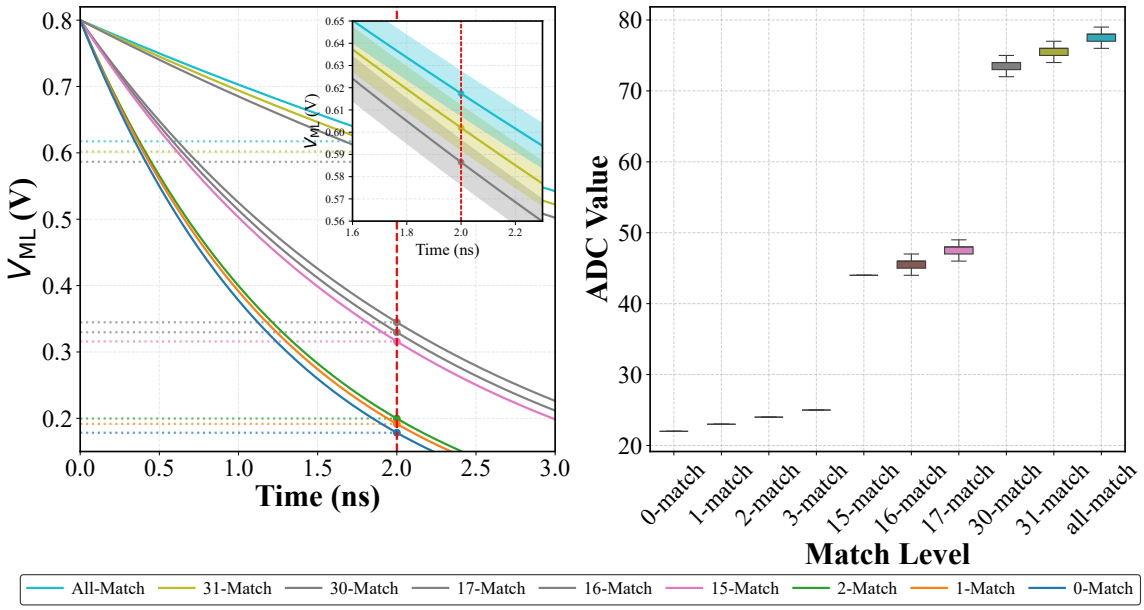


Figure 7.4.: (a) Match-Line (ML) voltage decay in a 32-bit ReCAM under global variation, with inset showing minimal separation for high match levels and (b) Boxplot of ADC outputs across match levels, where high match levels are more sensitive to variation

may cause overlap between adjacent similarity levels within the same chip, preventing reliable distinction between match levels.

Figure 7.2 illustrates these two cases. In Fig. 7.2a, the nominal ADC outputs for the 0-match and 1-match patterns are 22 and 23, respectively. Due to global variation, these values may shift, for example, to 21 and 22, while still remaining distinguishable. Although the chip functionality is preserved, conventional testing based on fixed ADC references would incorrectly reject such chips because the measured values deviate from the nominal ranges. These chips are categorized as F_{dev}.

To address this issue, the proposed calibration approach shown in Fig. 7.2b uses a per-chip LUT to correctly map the shifted ADC outputs to their corresponding similarity levels, thereby recovering F_{dev} chips and improving yield.

In contrast, local variation may cause adjacent similarity levels to overlap within the same chip. For example, both the 0-match and 1-match patterns may produce the same ADC value (e.g., 22), making the two similarity levels indistinguishable. These chips are categorized as F_{rep}, since reliable sensing is no longer possible and calibration cannot recover their functionality.

7.6. variation aware calibration test algorithm

7.6.1. Procedures Notations

Before applying our proposed calibration method, we define some notation to facilitate understanding of the calibration flow. Each ReCAM cell performs one primary operation: WRITE. During the WRITE operation, the prototype data pattern P is programmed into

the ReCAM cells, setting each cell to either the LRS or HRS. and also for Q through V_{DD} or *gnd*, to finally identify the number of similar bits. This specific operation helps obtain the ADC output across different chip instances and also between adjacent similarity inputs.

The following notation is used to describe the algorithmic procedures:

- w_P**d1**: Write the prototype data pattern $d1$ (e.g., all-0s or all-1s) into the ReCAM cells, setting each ReRAM element to either LRS or HRS accordingly.
- w_Q**d2**: Write a specific search data input pattern $d2$ for the query Q; these patterns are incrementing-0 or incrementing-1.

7.6.2. Calibration Test Algorithm

Our proposed calibration test algorithm consists of multiple calibration steps (CS), each designed to write prototype and query data in a controlled manner to identify all combinations of match levels along with their corresponding ADC outputs. Each step targets a specific input matching scenario to systematically characterize the ML voltage. We select one match pattern combination for each similarity level, but other combinations can be applied. The primary objective of the calibration is to populate the LUT. Therefore, after each calibration step, the corresponding LUT entry is filled with the measured value. This algorithm applies to various ML sizes (e.g., 16, 32, 64 bits). The calibration test fails, and the chip is marked as faulty (fallout) when two or more similarity levels map to the same ADC value within the same chip. In such cases, the overlap cannot be corrected, and the correct similarity level cannot be uniquely identified.

The complete sequence of operations for the calibration test algorithm is outlined as follows:

CS1: $(w_P\,\mathbf{1},\ w_Q\,\mathbf{0\ 0\ 0\ \ldots\ 0})$ – Apply **0-match** pattern (no match).

CS2: $(w_P\,\mathbf{1},\ w_Q\,\mathbf{1\ 0\ 0\ \ldots\ 0})$ – Apply **1-match** pattern (bit 0 matched).

CS3: $(w_P\,\mathbf{1},\ w_Q\,\mathbf{1\ 1\ 0\ \ldots\ 0})$ – Apply **2-match** pattern (bits 0–1 matched).

$\vdots$

CS$_k$: $(w_P\,\mathbf{1},\ w_Q\ \underbrace{\mathbf{1\ 1\ \ldots\ 1}}_{\boldsymbol{k}}\ \underbrace{\mathbf{0\ \ldots\ 0}}_{\boldsymbol{B-k}})$ – Apply **k-match** pattern.

$\vdots$

CS(B+1): $(w_P\,\mathbf{1},\ w_Q\,\mathbf{1\ 1\ 1\ \ldots\ 1})$ – Apply **B-match (all-match)** pattern.

For MLs with size B, we use $(B + 1)$ calibration steps. The process begins with CS1, which confirms the 0-match pattern consistently, followed by B incrementing 1-bit match

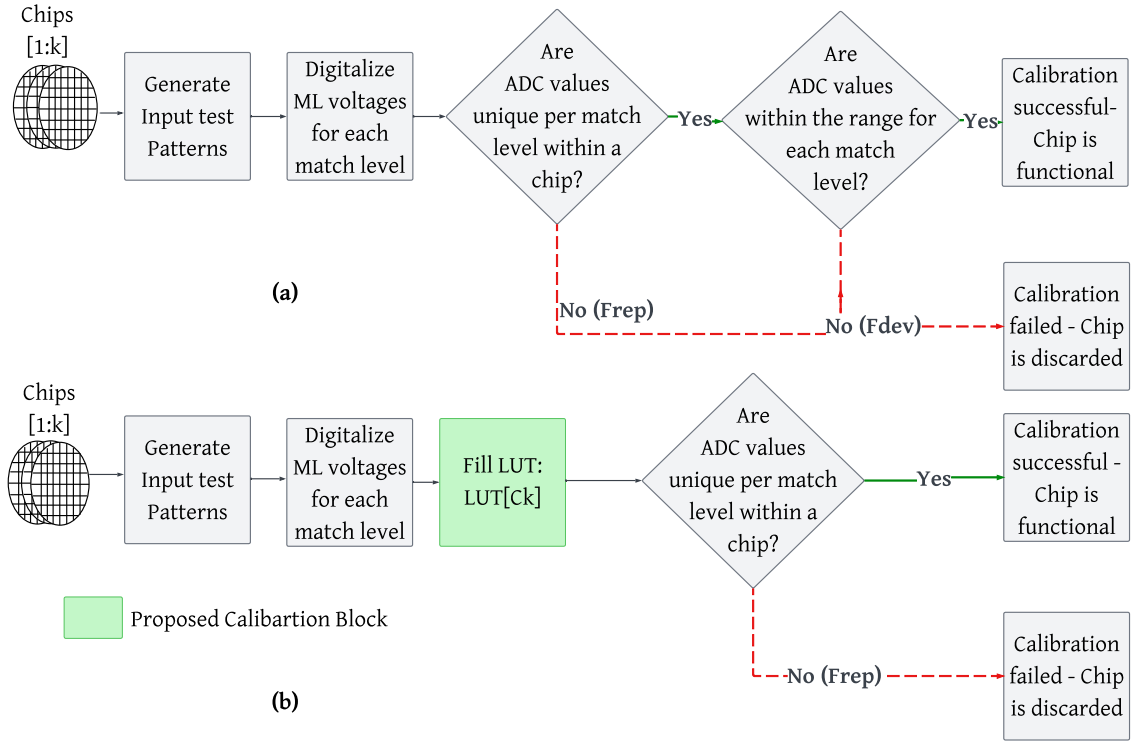


Figure 7.5.: Flow Chart of (a) without and (b) proposed calibration method

patterns to identify the ADC outputs corresponding to each similarity level for every chip instance. The post-calibration process begins by programming the entire array to logic '1' in P and then applying an all-0 pattern in Q to establish the baseline ML voltage corresponding to a 0-match scenario, which is sensitized by CS1.

To capture all B-match similarity levels, we apply an incrementing-1 pattern to Q, where a single '1' is placed at varying bit positions while the rest remain '0', effectively creating 1-match inputs. This pattern is applied sequentially, across all B cells in the ML, and is sensitized using CS($B+1$). This calibration test effectively accounts for global variations. It improves yield by tolerating benign variations that would otherwise result in false positives, a major contributor to yield degradation in ReCAM arrays. The CS_1 sequence requires a single write operation, performed in parallel for all MLs, with a total cost of $(w_P + w_Q)$ for the 0-match pattern. For CS_2 to CS_{B+1}, each sequence requires w_Q writes for the B-bit pattern per ML, and the shared ADC is activated once per sequence (ADC_{activ}). Thus, the total number of write operations required to fill the LUT for a single chip is given by:

$$\text{Complexity} = w_P + (B+1)w_Q + N(B+1)\,ADC_{\text{activ}} \tag{7.1}$$

where N is the number of MLs in the array, and B is the number of CAM cells per ML.

7.6.3. Lookup-Based Calibration

The calibration phase begins with the construction of an LUT for each chip, which stores the ADC output and its corresponding similarity level. Figure 7.5 illustrates the over-

Table 7.2.: Proposed calibration yield across fabrication scenarios

Fabrication Statistical Scenarios (FS)	$^*F_{\text{dev}}$	$^*F_{\text{rep}}$	No-calib (%)	Post-calib (%)	Yield Gain (%)
FS1: Local Measured (HRS: $\mu = 105.51$ **k**Ω**,** $\sigma = 19.495\%$ **LRS:** $\mu = 1.975$ **k**Ω**,** $\sigma = 5.472\%$**)**					
$\sigma_G = 5\%$	63	38	89.9	96.2	6.3
$\sigma_G = 7\%$	141	53	80.6	94.7	14.1
FS2: Local Measured (HRS: $\mu = 105.51$ **k**Ω**,** $\sigma = 20\%$ **LRS:** $\mu = 1.975$ **k**Ω**,** $\sigma = 15\%$**)**					
$\sigma_G = 7\%$	145	57	79.8	94.3	14.5
$\sigma_G = 8\%$	180	63	75.7	93.7	18.0

$^*F_{\text{dev}}$*: deviation failures* F_{rep}*: repetition failures.*

all flowchart of ReCAM calibration, which comprises two main flows: (a) one without calibration and (b) another after applying our proposed calibration method. The latter primarily mitigates the deviation failure. In the pre-calibration flow (Fig. 7.5a), a set of input test patterns is applied to the chip, and the resulting ML voltages for each match level are digitized using an ADC. The chip passes the test only if all ADC values are distinct across similarity levels; otherwise, it is marked as failed. In contrast, the proposed post-manufacturing calibration flow (Fig. 7.5b) introduces an additional verification step in which ADC outputs are compared against pre-defined ranges stored in an LUT specific to each chip.

Each chip is evaluated across all B-match levels, and its ADC outputs are compared to the LUT entries to ensure they fall within distinct ranges for different similarity levels. If all ADC outputs meet this criterion, the calibration is deemed successful, and the chip is classified as functional; otherwise, the calibration fails.

Subsequently, the yield before and after applying the calibration can be calculated as follows:

$$\text{No-Calib. Yield} = 1 - \frac{F_{\text{rep}} + F_{\text{dev}} - F_{\text{com}}}{N_{\text{chip}}} \tag{7.2}$$

$$\text{Post-Calib. Yield} = 1 - \frac{F_{\text{rep}}}{N_{\text{chip}}} \tag{7.3}$$

where F_{com} denotes the number of chips exhibiting common issues of F_{dev} and F_{rep}. N_{chip} is the total number of chip instances. By following this calibration procedure, we isolate and retain only F_{rep} as a critical indicator of functional failure, effectively eliminating the impact of F_{dev} on yield degradation.

7.7. Evaluation and Results

7.7.1. Yield Improvement

The analysis uses several FS scenarios that model the measured local and global variations listed in Table 7.2. To evaluate the effectiveness of the proposed post-calibration method,

Table 7.3.: Area, Power, and Latency Overheads for Different ReCAM Configurations

ReCAM Config.		Area (mm^2)		Power (W)		Latency (ns)		Overhead (%)		
B	N_{ML}	No LUT	With LUT	No LUT	With LUT	No LUT	With LUT	Area	Power	Latency
16	16	0.09501	0.09509	0.2056	0.2174	4.00	4.45	0.095	5.75	10.12
32	16	0.09501	0.09510	0.2058	0.2176	4.00	4.45	0.095	5.74	10.12
64	16	0.09502	0.09512	0.2063	0.2181	4.00	4.45	0.095	5.72	10.12
16	64	0.09502	0.09512	0.8223	0.8341	16.00	16.50	0.095	1.44	2.80
32	64	0.09505	0.09514	0.8232	0.8350	16.00	16.50	0.095	1.44	2.80
64	64	0.09509	0.09519	0.8252	0.8370	16.00	16.50	0.095	1.43	2.80

B: cells per match line; N_{ML}: number of match lines.

we calculate the yield before and after its application. As summarized in Table 7.2, post-manufacturing calibration significantly improves the yield across all evaluated scenarios. For example, in FS1 under moderate global variation ($\sigma_G = 7\%$), the yield improves from 80.6% to 94.7%, reflecting a gain of 14.1%. In more aggressive variation conditions ($\sigma_G = 8\%$) in FS2, our calibration method increases the yield from 75.7% to 93.7%, resulting in an 18.0% improvement. The first case represents more mature (low variation) and the other one less mature (higher variation). These results demonstrate the effectiveness of the calibration approach in mitigating such variation and enhancing the overall yield of ReCAM operation under various manufacturing variations.

7.7.2. Hardware Overhead

Table 7.3 summarizes the area, power, and latency overheads for different ReCAM configurations, defined by the number of match lines (N_{ML}) and the number of CAM cells per match line (B). All values are reported relative to the baseline design (i.e., without the LUT). The LUT-based calibration introduces negligible area and power overheads, with area increasing by only 0.1% and power rising by 1.43–5.75%. Because the LUT size is constant across configurations, its relative contribution decreases as the array size grows.

Unlike area and power, where the ADC dominates, the latency overhead is primarily influenced by the LUT. As a result, smaller arrays experience slightly higher latency penalties. For the smallest configuration (short MLs with 16 cells), the latency overhead is approximately 10.1%, whereas for the largest configuration (64 cells), it decreases to approximately 2.8%. Overall, the proposed calibration scheme requires only one LUT per chip (or per ReCAM array), thereby keeping area, power, and latency overheads minimal, as shown in Table 7.3. The LUT mechanism is integrated directly *in situ* within the ReCAM array, enabling efficient post-calibration without additional architectural modifications.

7.8. Summary

This chapter presented a post-manufacturing calibration methodology to improve the manufacturing yield of ReCAM arrays under both global and local process variations. The proposed approach employs a lightweight per-chip LUT to map ADC outputs to their corresponding similarity levels, enabling adaptive calibration of match boundaries for each chip individually. By compensating for die-to-die global variation, the method successfully recovers chips classified as deviation failures (F_{dev}), while still identifying unrecoverable repetition failures (F_{rep}) caused by overlap between adjacent similarity levels within the same chip. Experimental results demonstrate that the proposed calibration achieves up to 18% yield improvement compared to conventional fixed-boundary testing. In addition, the required LUT introduces negligible hardware overhead, with only about 0.1% area overhead, power overhead ranging from 1.43% to 5.75%, and latency overhead ranging from 2.80% to 10.12%.

8. TDsReCAM: Time-Domain sensing for reliable ReRAM-based Content Addressable Memory

Conventional testing and calibration assume that similarity levels can be sensed reliably. However, sensing itself becomes a dominant source of quality degradation under PV. Therefore, improving the sensing mechanism constitutes an additional dimension of quality assurance. Most existing ReCAM designs sense this voltage-domain output using single-threshold comparators or multi-bit ADC to distinguish among different similarity levels.

Although ADC-based sensing provides a straightforward interface between the analog ReCAM array and the digital backend, it introduces significant area, power, and calibration overhead. This limitation becomes more critical as the array size increases. Larger arrays reduce the voltage spacing between adjacent similarity levels, while ReRAM resistance variation, CMOS process variation, stochastic switching behavior, and interconnect parasitics broaden the corresponding ML voltage distributions. As a result, adjacent similarity levels may overlap, increasing the probability of quantization errors and degrading the bit-error-rate (BER). Maintaining reliable similarity detection under these conditions often requires higher ADC resolution and additional calibration, thereby increasing design complexity and limiting scalability.

To address these limitations, this chapter investigates time-domain sensing as an alternative to conventional voltage-domain ReCAM sensing. Instead of digitizing the ML voltage at a fixed sensing time, the proposed TDC-augmented ReCAM architecture converts the ML voltage decay into a threshold-crossing time. This time interval is then digitized using a low-overhead TDC, thereby shifting the similarity readout from the voltage domain to the time domain. By reducing dependence on high-resolution ADCs, the proposed approach reduces the sensing area and power consumption while improving robustness to PV-induced sensing errors. Compared with the conventional ADC-based baseline, the TDC-augmented ReCAM architecture achieves up to a 3.16× reduction in sensing area and more than four orders of magnitude reduction in sensing power, reaching up to $2.16 \times 10^4\times$, while also improving reliability under process variation.

8.1. Motivational Example

Figure 8.1 illustrates the limitation of conventional voltage-domain sensing in ReCAM arrays. In conventional ReCAM sensing, the residual ML voltage is sampled at a fixed sensing time and digitized using an ADC. Ideally, consecutive similarity levels should

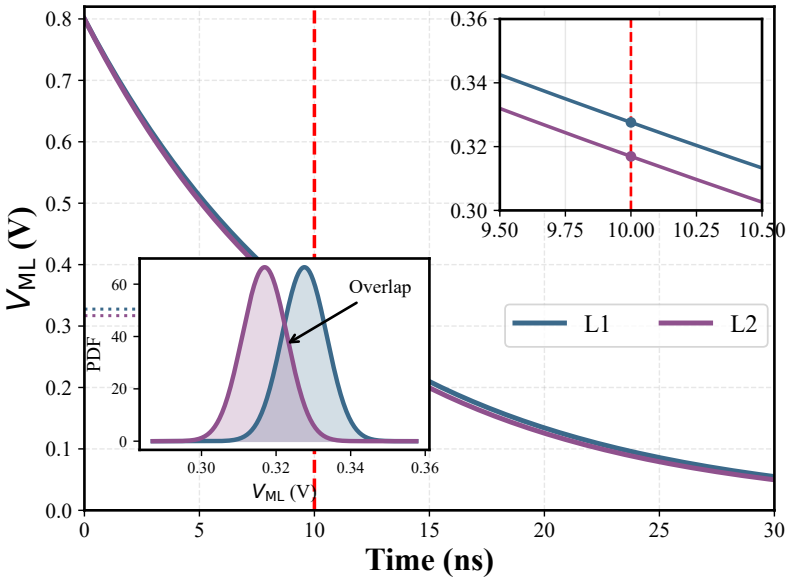


Figure 8.1.: Motivational example showing the limitation of conventional ADC-based voltage-domain sensing in ReCAM. Two adjacent similarity levels, L1 and L2, produce closely spaced ML voltages at the sampling time. Under PV, their voltage distributions overlap, making ADC code mapping not accurate and potentially leading to incorrect similarity detection.

produce clearly separated ML voltages so that each ADC code can be uniquely mapped to the correct similarity level.

However, as shown in Figure 8.1, the ML voltage responses of two adjacent similarity levels, denoted as L1 and L2, can become very close at the sampling time. Under PV, ReRAM resistance variability, and interconnect parasitics, the corresponding voltage distributions broaden and overlap. As a result, an ADC code may no longer correspond to a unique level of similarity. An ML voltage produced by one similarity level can therefore be quantized into a code assigned to a neighboring level, leading to incorrect similarity detection. This limitation becomes more severe as the ReCAM array size increases. Larger arrays accumulate more device variations and parasitic effects along the shared ML, while the voltage difference between consecutive similarity levels decreases. Therefore, distinguishing adjacent levels in the voltage domain requires either careful calibration or higher ADC resolution. However, increasing ADC resolution significantly increases sensing overhead. For example, an 8-bit flash ADC requires 255 comparators and a large resistor ladder, making the sensing block bulky and difficult to scale across many MLs. Moreover, ADC-based sensing often requires a continuous bias current during operation, resulting in high static power consumption.

These observations show that the reliability bottleneck in conventional ReCAM sensing is caused not only by analog ML discharge behavior but also by reliance on fixed-time voltage digitization. When PV causes adjacent ML voltage distributions to overlap, reliable ADC mapping becomes difficult without increasing resolution, calibration effort, area, and power. This motivates the replacement of voltage-domain sensing with time-domain sensing. Instead of assigning similarity based on the residual ML voltage, the proposed TDC-augmented ReCAM architecture measures the time it takes for the ML voltage to

cross a predefined threshold and digitizes this delay with a low-overhead TDC. By shifting the readout from voltage to time, the proposed approach aims to reduce ADC overhead and improve the BER.

8.2. Problem Statement

Conventional ReCAM architectures primarily rely on voltage-domain sensing to extract the similarity information encoded in the ML discharge. After a fixed evaluation time, the residual ML voltage is sampled and digitized using comparators or ADCs. Although this approach provides a straightforward interface between the analog ReCAM array and the digital backend, it becomes increasingly unreliable and costly as the array size grows.

The main challenge is that the voltage separation between adjacent similarity levels decreases with increasing ML length, while ReRAM resistance variation, CMOS process variation, and interconnect parasitics broaden the corresponding ML voltage distributions. This leads to distribution overlap, quantization errors, and degraded BER. A direct solution is to increase the ADC resolution or introduce calibration techniques. However, high-resolution ADCs introduce significant area and power overhead, particularly when high-speed sensing is required to preserve search throughput. In addition, ADC performance is itself sensitive to PV, requiring additional calibration across process, voltage, and temperature conditions.

Therefore, the central challenge addressed in this chapter is how to sense ReCAM similarity information reliably without relying on high-resolution voltage-domain digitization. Since the ML discharge trajectory inherently contains both voltage and timing information, an alternative is to encode similarity in the time required for the ML voltage to cross a predefined threshold. This raises three key questions:

- How can the analog ML discharge behavior in ReCAM be converted into a robust time-domain signature?
- Can time-domain sensing improve similarity-level separability under PV compared with conventional ADC-based voltage sensing?
- What are the area, power, and reliability trade-offs of using TDC-based sensing for ReCAM architectures?

Addressing these questions is essential for developing scalable ReCAM sensing architectures that maintain reliable similarity detection under device variability and circuit-level non-idealities.

8.3. Contributions

This chapter introduces a time-domain sensing methodology for reliable and scalable ReRAM-based CAM operation. The main contributions are summarized as follows:

- A TDC-augmented ReCAM sensing architecture is proposed to convert ML discharge behavior into a time-domain similarity signature. Instead of sampling the residual ML

voltage at a fixed sensing time, the proposed approach measures the ML threshold-crossing time and digitizes it using a TDC-based sensing circuit.

- A detailed comparison between voltage-domain and time-domain similarity sensing is presented. The analysis shows how PV-induced resistance fluctuations and interconnect parasitics affect ML voltage distributions, and how time-domain sensing can improve similarity-level separability.
- A circuit-level simulation framework is developed to evaluate the proposed sensing approach under ReRAM resistance variability, CMOS process variation, and different array sizes. The framework enables quantitative comparison between ADC-based voltage sensing and TDC-based time sensing in terms of sensing margin, BER, area, power, and scalability.
- The proposed TDsReCAM approach is evaluated with respect to reliability and sensing overhead. The results demonstrate that time-domain sensing reduces the dependence on high-resolution ADCs while improving robustness against PV-induced distribution overlap, making it a promising alternative for scalable ReCAM architectures.

8.4. Sensing Techniques for ReCAM and CiM

Voltage-domain sensing using ADCs has been the dominant readout mechanism in ReRAM-based CiM architectures, including ReCAM arrays. Several works adopt ADC-based quantization of match-line or bit-line voltages to infer similarity or computation results from resistive memory arrays [111], [128]. Survey and benchmarking studies also highlight the widespread use of voltage-domain sensing in ReRAM-based CiM [23], [129].

However, prior work has identified key challenges associated with ADC-based sensing, including high sensitivity to PV, substantial sensing overhead, and limited scalability [121], [130]. As array dimensions increase, the required sensing resolution and calibration complexity grow, making voltage-domain sensing less attractive for large-scale and energy-constrained designs [23].

To address these challenges, time-domain computing has emerged as an alternative paradigm for CiM architectures. Instead of directly digitizing analog voltages, time-domain approaches encode computation results as delays or discharge times, which can be digitized using TDCs or counter-based structures [131], [132]. Prior work has demonstrated the effectiveness of time-domain sensing in mitigating analog non-idealities and improving robustness to variations, mainly in vector-matrix multiplication and accumulation operations [133].

Despite these advances, existing time-domain CiM realizations mainly target arithmetic operations and do not address CAM-based similarity search. To the best of our knowledge, TDC-based time-domain sensing has not been systematically explored for ReCAM architectures. This chapter addresses this gap by introducing a TDC-augmented ReCAM architecture that leverages ML discharge time as a similarity signature to improve reliability while reducing sensing area and power overhead.

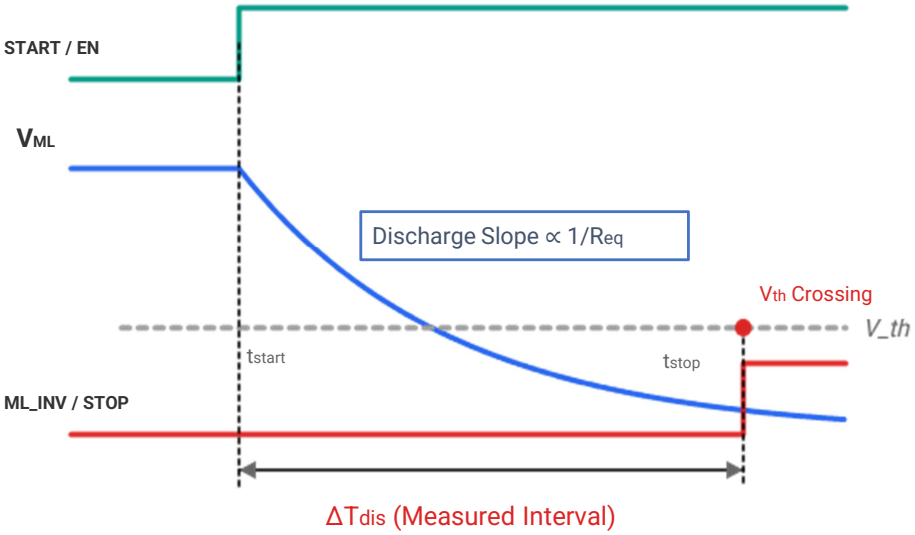


Figure 8.2.: Timing diagram of TDC-based ML sensing. At t_{start}, evaluation begins, and the enable signal (EN) activates the ML discharge path. The ML voltage V_{ML} discharges from the precharged level until it crosses the inverter trip point V_{th} at t_{stop}, causing ML_INV to toggle. The TDC measures the discharge time as a digital time code.

8.5. Proposed Time-Domain Design

The proposed approach performs time-domain similarity sensing by exploiting the ML's discharge behavior. After the precharge phase, the ML voltage decays from the initial level V_{DD} toward ground. As illustrated in Fig. 8.2, the ML discharge time is measured between two reference levels, $V_1 = V_{DD}$ and $V_2 = V_{\text{th}}$. This discharge time depends on the number of matching bits between the stored and input patterns. The resulting time interval serves as a proxy for the Hamming weight of similarity and is digitized using a TDC. To ensure that the discharge-time differences ΔT_{dis} between adjacent similarity levels are resolvable, an explicit load capacitance C_{L} is introduced on the ML. Adjusting C_{L} provides a direct design knob for tuning the ML discharge dynamics so that the timing separation matches the TDC resolution. Finally, a lightweight LUT maps the TDC output codes to the corresponding similarity levels, enabling efficient and robust digital interpretation of the time-domain sensing result.

8.5.1. Time-Domain Architecture for ReCAM Arrays

Based on this principle, a TDC is integrated into the ReCAM array to replace conventional voltage-domain quantization. As shown in Figure 2.4(a), since the rest of the ReCAM architecture operates digitally, time-domain sensing naturally aligns with the system design. To reduce area and energy overhead, the TDC is shared across multiple MLs, similar to the sharing strategy commonly used for ADCs in conventional ReCAM designs [134].

Equation 8.1 expresses ΔT_{dis} as a function of C_{L} and the equivalent discharge resistance R_{eq} of the ReCAM array. During the evaluation phase, the ML voltage V_{ML} decays exponentially due to the combined effect of C_{L} and R_{eq}. The equivalent resistance R_{eq} depends

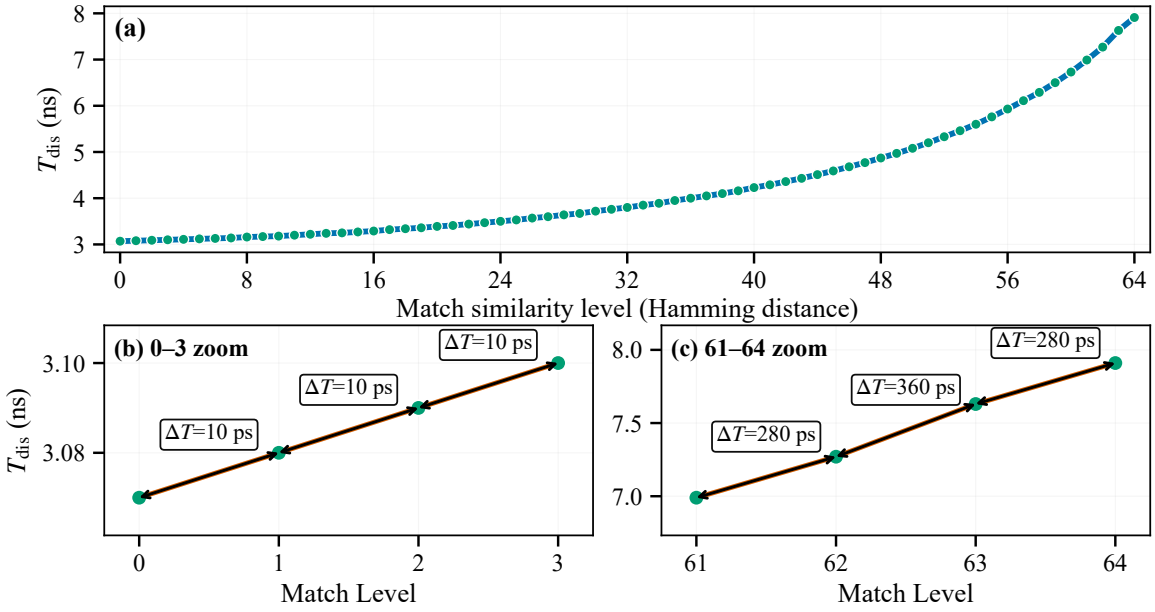


Figure 8.3.: Discharge time T_{dis} versus match similarity level in a 64-bit ReCAM, (a) Full-range response, (b) Low-level similarity group zoom (0–3) with $\Delta T \approx 10$ ps, and (c) High-level similarity group zoom (61–64) with $\Delta T \approx 280$–360 ps

on the number of matching bits between the stored pattern and the input pattern. A higher number of matching bits leads to a larger R_{eq} and, consequently, a slower ML discharge. As a result, the ML voltage remains above the sensing threshold for a longer duration, yielding a larger discharge time interval ΔT_{dis} before crossing the threshold voltage V_{th} of the cascaded inverters.

Note that the logarithmic term captures the voltage drop relative to the sensing threshold, and the resulting time interval ΔT_{dis} is considered in magnitude. Therefore, Figure 8.3 illustrates the relationship between the match-line discharge time T_{dis} and the match similarity level, from 0-match all the way to 64-bit match, in a 64-bit ReCAM array.

$$\Delta T_{dis} = t_{stop} - t_{start} = C_L R_{eq} \ln\left(\frac{V_{th} - V_{ML}(t_{start})}{V_{DD}}\right) \tag{8.1}$$

8.5.2. Reliability Analysis of Time-Domain Sensing

Due to intrinsic ReCAM constraints, including limited ML length, circuit non-idealities, and process variation (PV), the ML discharge time T_{dis} does not map uniquely to the exact number of matching bits. As a result, the timing separation ΔT_{dis} between adjacent similarity levels is inherently non-uniform.

Figure 8.3(a) shows that T_{dis} increases monotonically with the number of matching bits, reflecting the cumulative contribution of parallel discharge paths on the ML. However, the resulting characteristic is strongly nonlinear: adjacent match levels at low similarity exhibit very small timing separations, while substantially larger separations appear at higher similarity levels. This non-uniformity makes low-similarity levels the most vulnerable to PV-induced overlap and therefore represents the worst-case reliability condition.

This effect is illustrated in Fig. 8.3(b), which zooms into match levels 0–3. For the simulation setup in Table 9.1, the minimum inter-level separation is approximately 10 ps, defining the most stringent timing resolution requirement.

In contrast, Fig. 8.3(c) shows that at high similarity levels (61–64), the separation increases to 280–360 ps, providing substantially larger timing margins and reduced sensitivity to PV.

Figure 8.4 further highlights the impact of PV by showing the T_{dis} distributions for the most critical adjacent match-level pairs in a 64-bit ReCAM array. The distributions reveal small but finite mean separations, while their widths capture PV-induced timing uncertainty. These results directly define the minimum timing resolution required from the TDC to reliably discriminate neighboring similarity levels. Despite partial distribution overlap, the available timing margins confirm that, with an appropriate TDC resolution, robust sensing can be achieved even under significant PV.

Table 8.1.: Simulation Setup and Parameters

Category	Parameter Details
Technology-Level Parameters	
ReRAM Model [23]	JART VCM Read Variability
Filament Radius	45 nm
Disc Region Length	0.6 nm
N_{init} (HRS, LRS)	9 m, 3×10^{26} m^{-3}
Circuit-Level Parameters	
Simulation Tool	Cadence Virtuoso
CMOS Technology	GlobalFoundries 22FDX
Supply Voltage (V_{DD})	800 mV
Temperature	27°C
ADC [95]	8-bit resolution
TDC [134]	8.52 ps resolution
Interconnect per ReCAM Cell (22 nm) [47]	ML: $R = 11.95\ \Omega$, $C = 16.63$ aF
Match-Line Length	64-bit
Match-Line Load Capacitance (C_{L})	6.04 pF
Performance and Area Parameters	
CAM Cell Area [108]	0.0242 μm^2
ADC Area	0.095 mm^2
TDC Area	0.003 mm^2
LUT Area	22.5 μm^2
CAM Cell Power	0.149 mW
ADC Power	50.8 mW
TDC Power	0.49 mW
LUT Power	2.95 mW

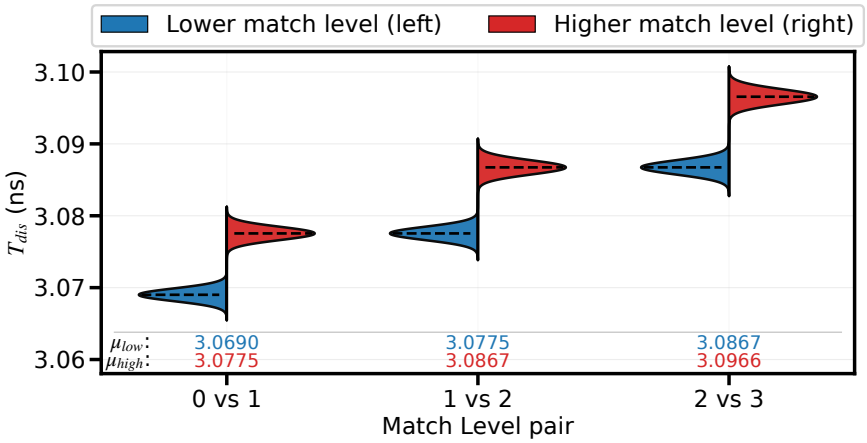


Figure 8.4.: Distributions of the ML discharge time T_{dis} for adjacent match-level pairs (0–1, 1–2, and 2–3) in a 64-bit ReCAM array. Half-violin plots represent the lower and higher match levels, with dashed lines indicating the mean values μ_{low} and μ_{high} of these pairs

8.5.3. Bit-Error Rate Analysis

8.5.3.1. BER in the Time Domain

To evaluate the reliability of time-domain sensing in the presence of process PV, the BER between adjacent match levels is analyzed based on the overlap of their match-line discharge time (T_{dis}) distributions. The BER quantifies the probability of an incorrect decision when a reference threshold is used to distinguish between two neighboring match levels.

Let (μ_i, σ_i) and $(\mu_{i+1}, \sigma_{i+1})$ denote the mean and standard deviation of the T_{dis} distributions corresponding to match levels i and i+1, respectively. The standard deviations include both PV-induced variability and sensing-related uncertainty, accounting for timing jitter and TDC quantization effects. For a given decision threshold t_{ref} such that $\mu_i < t_{\text{ref}} < \mu_{i+1}$, decision errors originate from the overlapping tails of the two distributions.

The probability that a match level i is incorrectly classified as i+1 is given by

$$P_{e,i} = \Pr(T_i > t_{\text{ref}}) = 1 - \text{CDF}\left(\frac{t_{\text{ref}} - \mu_i}{\sigma_i}\right), \tag{8.2}$$

while the probability that a match level i+1 is misclassified as i is

$$P_{e,i+1} = \Pr(T_{i+1} \leq t_{\text{ref}}) = \text{CDF}\left(\frac{t_{\text{ref}} - \mu_{i+1}}{\sigma_{i+1}}\right). \tag{8.3}$$

The BER associated with this decision boundary is defined using a worst-case criterion as

$$\text{BER}_{i \to i+1} = \max\left(P_{e,i},\, P_{e,i+1}\right), \tag{8.4}$$

which represents the maximum probability of incorrect classification between the two adjacent match levels. The reference threshold t_{ref}^{*} is obtained by sweeping t_{ref} and minimizing $\text{BER}_{i \to i+1}$, yielding a balanced trade-off between the two error probabilities.

8.5.3.2. BER in the Voltage Domain

For the ADC-based approach, BER is evaluated directly in the discrete code domain. The sampled ML voltages are first mapped to digital codes using a full-scale range of ML voltages. To account for sensing non-idealities, an input-referred noise component is added to the voltage samples before quantization. The noise standard deviation is set equal to the quantization noise of the ADC converter, which models the combined effect of comparator noise, reference fluctuations, and sampling uncertainty.

After noise injection and quantization, BER is evaluated empirically in the code domain. For a given decision threshold c_{ref}, the misclassification probabilities between two adjacent similarity levels are

$$P_{e,i} = \Pr\{C_i > c_{\text{ref}}\}, \tag{8.5}$$

$$P_{e,i+1} = \Pr\{C_{i+1} \leq c_{\text{ref}}\}, \tag{8.6}$$

where C_i and C_{i+1} denote the output codes associated with match levels i and i+1, respectively.

The BER corresponding to this threshold is defined using a minimax criterion,

$$\text{BER}^{\text{ADC}}_{i \to i+1} = \max\left(P_{e,i},\, P_{e,i+1}\right), \tag{8.7}$$

and the optimal decision threshold is obtained by sweeping c_{ref} over all possible code values and minimizing the worst-case error probability.

8.6. Evaluation and Results

8.6.1. Experiment setup

8.6.1.1. Simulation Framework

Electrical-level simulations are performed in SPICE to evaluate the impact of process variation (PV) on the sensing behavior of the ReCAM array. A detailed circuit-level model is used, including the match line (ML), access devices, and peripheral sensing circuitry, as summarized in Table 9.1. Monte Carlo analysis is employed to capture device- and interconnect-level variability and to assess its effect on the ML discharge dynamics.

A comprehensive set of similarity patterns is applied, spanning from the all-mismatch (0-match) to the all-match (64-match) cases, enabling a systematic study of the dependence of ML discharge behavior on match conditions under PV. To meet the timing resolution requirements of the TDC-based sensing scheme, a parametric sweep is performed on the load capacitance C_{L}.

By analyzing the resulting discharge-time distributions across similarity levels, the framework identifies the most critical adjacent match pairs with the largest distribution overlap, which represent the worst-case scenarios for reliable sensing. For the voltage-domain baseline, an 8-bit high-speed flash ADC is assumed, consistent with prior ReCAM implementations that rely on single-cycle ML digitization [95], [111]. The ADC configuration follows reported designs to ensure a fair and conservative comparison.

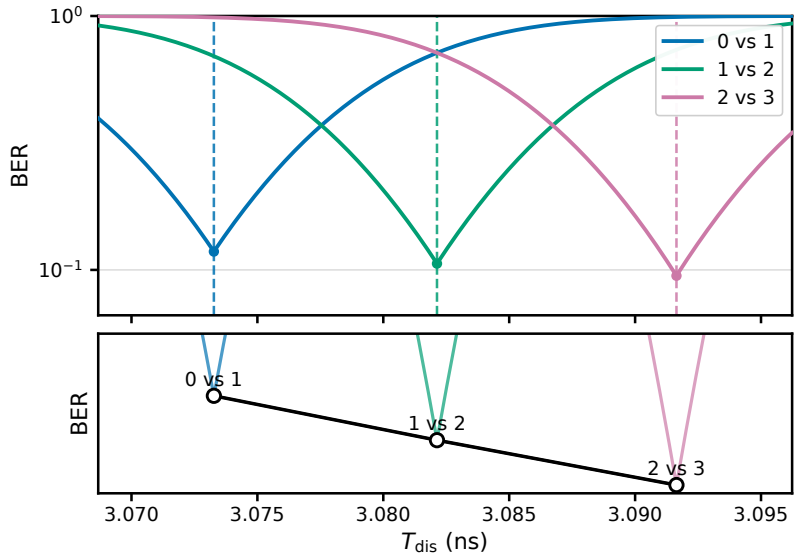


Figure 8.5.: Bit-error rate (BER) as a function of the discharge time T_{dis} for adjacent match-level pairs. The upper plot shows BER curves for the 0 vs. 1, 1 vs. 2, and 2 vs. 3 comparisons, with dashed lines indicating the optimal thresholds t_{opt} that minimize BER. The lower plot zooms in on the low-BER region and highlights the decreasing trend in BER_{min} as match-level separation increases

8.6.1.2. Statistical Analysis

For each similarity condition, the ML discharge time T_{dis} is extracted over a large number of Monte Carlo samples to construct statistical distributions. These distributions are used to quantify the separability between adjacent match levels, which directly determines the probability of sensing errors.

The degree of overlap between neighboring distributions is analyzed to identify PV-vulnerable cases that exhibit the highest likelihood of decision failure. In particular, similarity pairs with small inter-level separation are emphasized, as they constitute the dominant contributors to the worst-case BER. This statistical characterization forms the basis for the subsequent BER evaluation and enables a consistent comparison between voltage-domain and time-domain sensing.

8.6.2. Bit-Error Rate Results

To ensure a fair comparison between ADC- and TDC-based sensing, BER is evaluated by consistently accounting for errors arising from both array-induced variability and sensing-induced non-idealities. For the TDC-based approach, the ML discharge times of the 0-match and 1-match cases are modeled as Gaussian random variables. In addition to PV, sensing uncertainty is incorporated by adding, in quadrature, a timing jitter term and a quantization noise term corresponding to the TDC resolution. The BER is obtained by sweeping the decision threshold and minimizing the worst-case failure probability.

For the ADC-based approach, BER is evaluated directly in the discrete code domain to avoid invalid Gaussian assumptions after quantization. An input-referred RMS noise floor equal to the quantization noise of an 8-bit converter is injected before quantization,

Table 8.2.: Comparison of area, power, and bit-error rate (BER) between conventional ADC-based sensing and the proposed TDC-based sensing architecture for different ReCAM configurations.

ReCAM Config.		ReCAM Core		Total Area			Total Power			BER	
B	N_{ML}	Area	Power	ADC	TDC	Area Red.	ADC	TDC	Power Red.	ADC	TDC
cells/ML	MLs	(μm^2)	(mW)	(mm^2)	(mm^2)		(mW)	(mW)		0-/1-match	
16	16	0.0242	0.149	0.095006	0.030006	3.17×	205.6	0.01520	13.5k×	3.23×10^{-1}	$\ll 10^{-32}$
32	16	0.0242	0.162	0.095012	0.030012	3.17×	205.8	0.01655	12.4k×	8.50×10^{-2}	9.11×10^{-11}
16	64	0.0242	0.149	0.095025	0.030025	3.16×	822.3	0.03812	21.6k×	3.23×10^{-1}	$\ll 10^{-32}$
32	64	0.0242	0.162	0.095050	0.030050	3.16×	823.2	0.04154	19.8k×	8.50×10^{-2}	9.11×10^{-11}

*B denotes the number of cells per match line, and N_{ML} denotes the number of match lines. The ADC-based baseline is derived from [108], [111]. The reported reductions are calculated as the ratio of the ADC-based value to the corresponding TDC-based value. TDC-based refers to the proposed TDsReCAM architecture.

capturing the combined effect of comparator, reference, and sampling noise. The BER is then computed empirically by sweeping the decision threshold over all code values and minimizing the worst-case classification error. This procedure prevents overestimation of ADC performance due to unrealistically small noise assumptions, while ensuring that TDC results reflect both array-induced and sensing-induced timing uncertainties. Figure 8.5 reports the BER as a function of the decision threshold t_{ref} for the worst-case adjacent match-level pairs, namely 0 vs. 1, 1 vs. 2, and 2 vs. 3. Each curve exhibits a clear minimum at an optimal threshold t_{opt}, located close to the midpoint between the corresponding distribution means, which reflects the balanced trade-off between the two tail error probabilities. The upper plot shows the full BER curves, while the lower plot zooms into the operating region around 10^{-1} and connects the optimal points across match-level pairs. As the match level increases, the separation between adjacent distributions, $\Delta\mu = \mu_{i+1} - \mu_i$, grows monotonically, leading to a reduced minimum BER. This confirms that closely spaced distributions at low match levels define the worst-case reliability condition.

8.6.3. Area, Power, and BER Analysis

Table 8.2 compares the area, power, and worst-case BER of ADC- and TDC-based sensing across different ReCAM configurations. For a given array size, the ReCAM core area and power remain unchanged, while the sensing circuitry dominates the system overhead. Replacing the voltage-domain ADC with a time-domain TDC reduces the sensing area by approximately 3.16× across all configurations. This gain results from eliminating bulky analog blocks in the ADC, such as resistor ladders and parallel comparators, and replacing them with a compact digital timing structure.

More importantly, the proposed TDC-based sensing achieves a power reduction exceeding four orders of magnitude, with improvements of up to 2.16×10^4 for large arrays. This reduction stems from the event-driven, fully digital operation of the TDC, in contrast to the continuously biased nature of flash ADCs. Notably, both the area and power benefits

are largely independent of the number of cells per match line (B) and the number of match lines (N_{ML}), indicating favorable scalability with array size.

Beyond hardware overheads, Table 8.2 also reports the worst-case BER between the 0-match and 1-match levels, which represents the most challenging discrimination scenario under process variation. For ADC-based sensing, BER is fundamentally constrained by voltage-domain quantization. For arrays (B=16 and B=32), the inter-level voltage separation is significantly smaller than the effective noise floor of the 8-bit ADC, causing multiple match levels to collapse into a limited number of output codes. As a result, the worst-case BER remains high, at approximately 3.2×10^{-1} for B=16 and 8.5×10^{-2} for B=32. In contrast, the proposed TDC-based sensing consistently achieves lower BER across all configurations. By operating in the time domain and exploiting the separation in ML discharge times between adjacent match levels, the TDC preserves substantially larger effective decision margins than the voltage-domain approach. This leads to orders-of-magnitude improvements in reliability for different sizes of arrays and a sustained advantage over the ADC baseline for larger configurations. Overall, relative to the ADC-based design, the TDC reduces the worst-case BER by 8 orders of magnitude for B=32, and very small values are negligible for B=16. These results demonstrate that time-domain sensing substantially enhances robustness to process variation while simultaneously reducing area and power overhead.

8.7. Summary

This chapter presented a TDC-augmented sensing methodology for reliable ReCAM architectures. Instead of digitizing the residual ML voltage using an ADC, the proposed approach measures the ML threshold-crossing time and converts it into a digital code using a low-overhead TDC. This time-domain formulation reduces reliance on high-resolution ADCs and improves robustness to PV-induced voltage overlap.

The proposed sensing approach was evaluated through Gaussian modeling of discharge-time distributions and bit-error-rate analysis between adjacent match levels. The results showed that time-domain sensing provides improved match-level separability and enables reliable threshold selection under PV. Furthermore, the discharge-time span across all match levels was analyzed to determine the required TDC resolution and bit width, providing practical guidance for hardware implementation. Overall, this chapter demonstrated that TDC-based sensing is a scalable, area- and power-efficient, and variation-tolerant alternative to conventional ADC-based ReCAM readout. These results establish time-domain sensing as a promising solution for reliable similarity search in future ReCAM-based computation-in-memory architectures.

9. Runtime BIST for ReRAM-based CAM using Frequency-Domain Monitoring

Conventional testing methodologies for ReCAM primarily focus on manufacturing-time fault detection and often rely on area- and power-intensive sensing circuits, such as high-resolution ADCs. While effective for offline testing, these approaches are not well-suited to continuous runtime monitoring, where lightweight, low-overhead reliability mechanisms are required. Therefore, efficient runtime BIST techniques are essential for ensuring reliable ReCAM operation throughout the system's lifetime.

In this chapter, a lightweight runtime BIST framework for ReCAM is presented using frequency-domain monitoring. The proposed approach employs a VCRO-based sensing mechanism that converts analog voltage behavior into frequency signatures, enabling compact, scalable anomaly detection without relying on high-resolution ADCs. By monitoring frequency deviations associated with abnormal voltage behavior, the framework can effectively detect reliability degradations caused by process and runtime variations while introducing minimal hardware overhead.

Simulation results demonstrate that the proposed runtime monitoring framework achieves high fault coverage while maintaining negligible area and power overheads, making it a practical solution for reliable next-generation ReCAM-based CiM architectures.

9.1. Motivational Example

The reliability of ReCAM degrades over time due to aging-induced variations in the ReRAM devices and peripheral circuits. Such variations alter the ML behavior during similarity-search operations, which can eventually lead to incorrect match detection. Figure 9.1(a) and (b) show ML voltage decay under the impact of different aging parameters, including filament radius (r_f) and disc region length (l_d). Aging slightly alters the discharge slope, but the resulting voltage differences are small and often overlap across similarity levels, particularly at low similarity (e.g., 0-match). However, at high similarity (e.g., all-match), there is significantly greater sensitivity to aging effects, resulting in noticeable shifts in V_{ML}. Thereby requiring a high-resolution, power-hungry ADC or precise timing measurements for reliable detection.

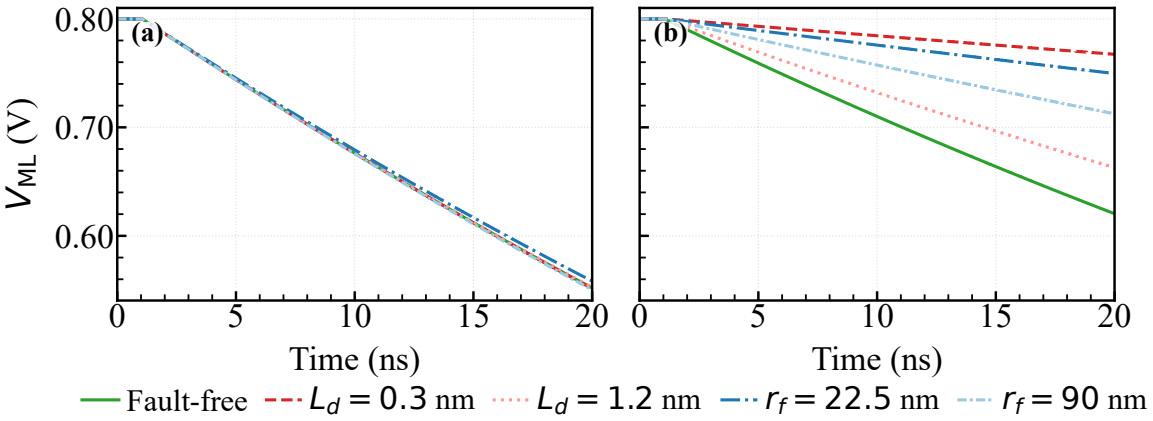


Figure 9.1.: V_{ML} behavior under different aging parameters. (a) 0-match condition showing relatively low sensitivity to filament radius (r_f). (b) all-match condition where aging effects produce larger deviations in V_{ML}.

9.2. Problem Statement

The preceding discussion shows that ReCAM reliability cannot be guaranteed solely through manufacturing testing or one-time calibration. During field operation, aging, resistance drift, and runtime variations can progressively alter the ML discharge behavior, reducing sensing margins and potentially causing similarity-search errors. However, repeatedly invoking full-array diagnosis is impractical due to its latency and energy overhead. Similarly, high-resolution ADC- or TDC-based monitoring increases the area and power cost of the sensing circuitry.

Therefore, the key problem addressed in this chapter is how to monitor runtime degradation in ReCAM arrays with minimal hardware overhead and without disrupting normal similarity-search operation. This motivates a lightweight frequency-domain monitoring approach that can track abnormal ML behavior during lifetime operation.

9.3. Contributions

To address the above problem, this chapter proposes a runtime BIST framework for ReCAM using frequency-domain monitoring. The main contributions are summarized as follows:

- A lightweight Voltage-Controlled Ring Oscillator (VCRO)-based sensing architecture is introduced to convert ML-dependent analog behavior into frequency signatures without requiring high-resolution ADCs.
- A runtime BIST methodology is developed to detect frequency deviations caused by aging- and variation-induced changes in ML behavior during ReCAM operation.
- A statistical decision mechanism is employed to suppress transient fluctuations and improve the robustness of runtime anomaly detection.

- Circuit-level simulation results demonstrate high test coverage with negligible area and power overhead, confirming the suitability of the proposed framework for reliable lifetime operation of ReCAM-based CiM architectures.

9.4. Runtime Reliability Monitoring Challenges in ReCAM

Prior research on reliability in ReRAM arrays has primarily focused on post-manufacturing testing and calibration techniques, including defect modeling and March-like test algorithms, to detect stuck-at faults, deep-HRS/LRS defects, and other parametric variations [23], [135], [136]. While these approaches are effective for manufacturing-time screening and yield improvement, they are not designed to monitor gradual runtime degradation mechanisms such as resistance drift, aging, and time-dependent variability during field operation. Furthermore, repeated full-array diagnosis introduces considerable latency and energy overhead, limiting its applicability for continuous runtime reliability monitoring.

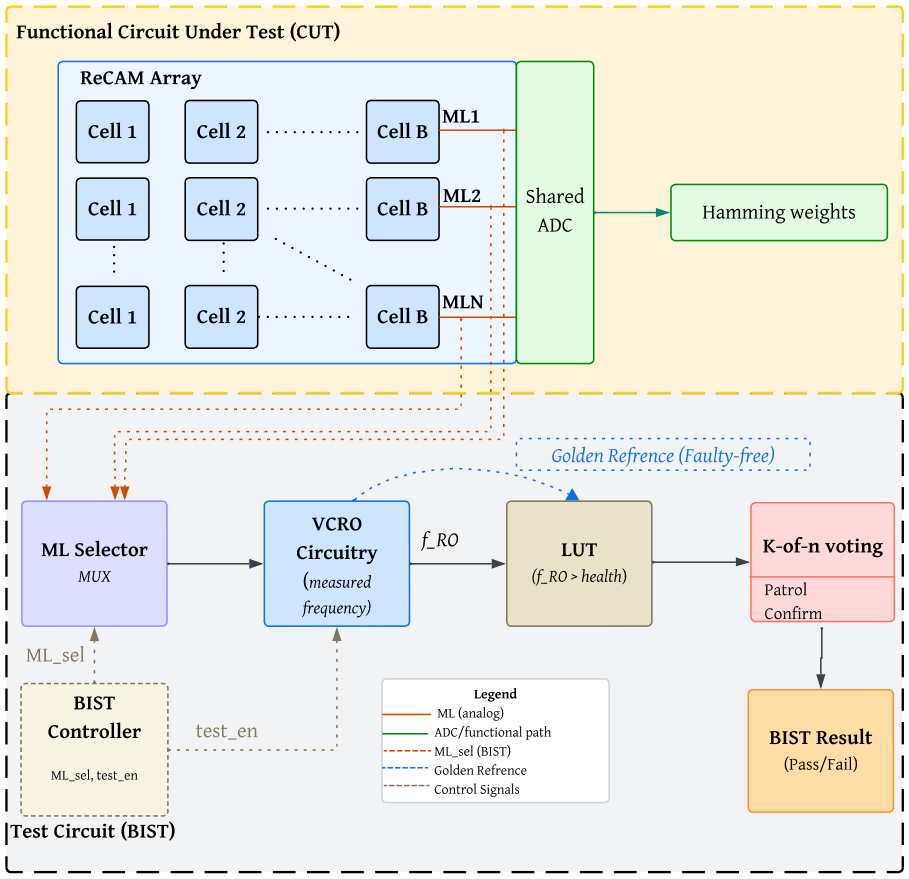
BIST techniques are widely adopted in conventional memory and logic circuits; however, extending them to ReRAM arrays remains challenging due to device stochasticity, analog sensing behavior, and time-varying resistance states [137]. Most existing ReRAM-oriented BIST approaches mainly target characterization, margin analysis, or manufacturing-time defect detection, making them less suitable for lightweight runtime monitoring during normal system operation [135], [136], [137].

Time-domain monitoring techniques have also been explored to capture circuit degradation through delay variations [138], [139], [140]. In such approaches, degradation effects are translated into timing signatures that can be digitized using TDC or delay sensors. However, directly applying delay-based monitoring to ReCAM architectures is challenging because similarity evaluation depends on the collective discharge behavior of multiple cells connected to a shared match line (ML). As a result, the timing differences between adjacent similarity levels are often extremely small, requiring high-resolution TDCs and precise delay measurement circuitry. Such requirements introduce significant hardware overhead and may interfere with normal array operation [129].

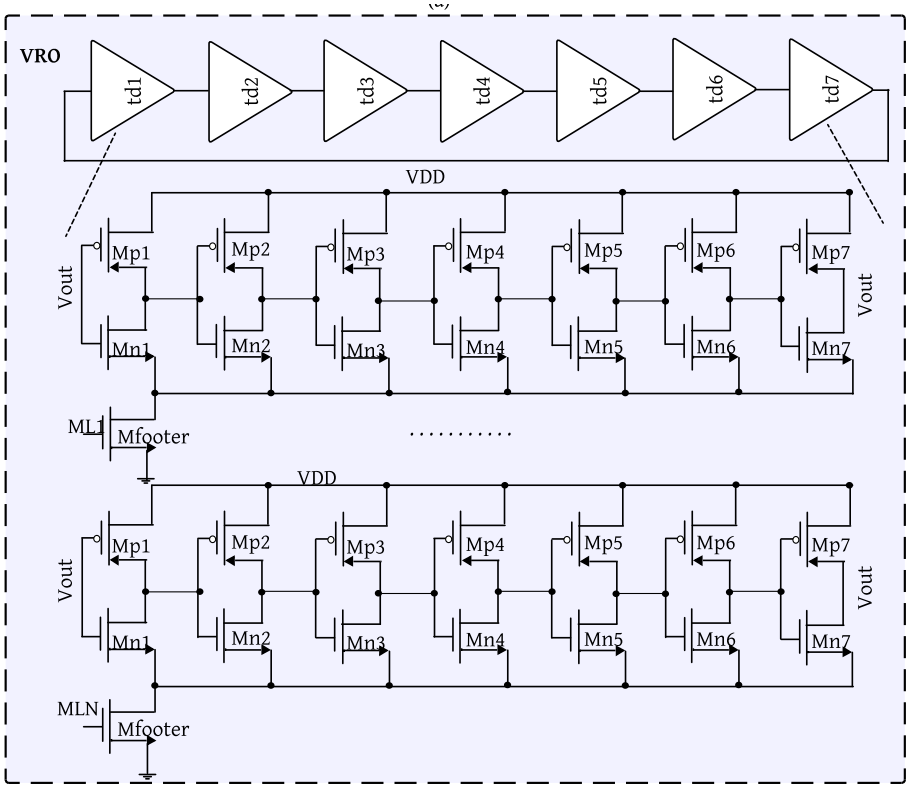
In parallel, Ring Oscillator (RO)-based sensors have been extensively used in CMOS circuits to monitor aging and delay degradation through frequency shifts [140], [141]. These approaches provide compact and low-overhead indicators of gradual circuit degradation. However, RO-based monitoring techniques have not been explored in ReCAM architectures, particularly for tracking ML degradation during similarity search operations.

9.5. Proposed BIST for ReCAM Architecture

The ML discharge behavior in ReCAM arrays is strongly affected by device aging and runtime variations, which modify the effective pull-down strength of the mismatch paths. As aging mechanisms alter the resistance characteristics of the ReRAM filament, and also the threshold voltage of CMOS transistors, the ML voltage decay profile changes



(a)



(b)

Figure 9.2.: Proposed runtime BIST: (a) architecture with the circuit under test (CUT) and VCRO-based monitoring; (b) VCRO readout circuitry converting ML voltage to oscillation frequency f_{RO}.

accordingly. However, these variations result in only minuscule shifts in the ML discharge behavior, making them difficult to reliably distinguish with direct voltage-domain sensing.

To address the challenge shown in Figure 9.1, the proposed BIST architecture converts the ML discharge voltage to the frequency domain using a compact VCRO. The ML voltage controls the current-starved stage of the VCRO, converting the ML discharge dynamics into an oscillation frequency f_{RO}. Small variations in the ML decay, therefore, manifest as measurable shifts in frequency. Finally, a lightweight LUT maps the measured frequency to the corresponding similarity level as shown in Fig.9.2.

9.5.1. Frequency-Domain Sensing Using VCRO

During evaluation, the ML voltage reflects the aggregate pull-down strength of the activated ReCAM cells as hamming weights across similarity levels. The proposed approach translates the ML discharge behavior into a frequency-domain observable. The ML voltage follows an exponential decay governed by the effective load capacitance C_L and the equivalent discharge resistance R_{eq}, as follows

$$V_{ML}(t) = V_{DD} \exp\left(-t/R_{eq}C_L\right) \tag{9.1}$$

The value of R_{eq} depends on the number of matching bits between the two patterns. Higher similarity increases R_{eq}, leading to slower discharge and a higher ML voltage within the sensing window. The ML node biases the current-starving device of VCRO. Consequently, the starve current becomes a function of V_{ML}.

Assuming a power-law current model

$$I_{st} = \beta(V_{ML} - V_{th,st})^{\alpha},$$

the oscillation frequency can be approximated as

$$f_{RO} \approx I_{st}/2NC_{inv}V_{sw},$$

where N denotes the number of stages, C_{inv} the effective inverter capacitance, and V_{sw} the switching swing. Substituting the ML decay expression yields

$$f_{RO}(t) \approx \frac{\beta}{2N\,C_{inv}\,V_{sw}} \left(V_{DD} \exp\left(-\frac{t}{R_{eq}C_L}\right) - V_{th,st}\right)^{\alpha}. \tag{9.2}$$

Equation (9.2) establishes a direct relationship between the oscillation frequency and the similarity levels depending on the discharge behavior of the ML. Larger R_{eq} values, corresponding to higher similarity levels, maintain a higher ML voltage during sensing and therefore produce a higher oscillation frequency. This preserves a monotonic ordering between similarity levels in the frequency domain. Figure 9.3 (a) and (b) confirms this behavior. The measured f_{RO} increases consistently with match similarity, while the corresponding oscillation period decreases according to $T_{RO} = 1/f_{RO}$. The dual representation highlights that the similarity representation is fully captured through the frequency signature.

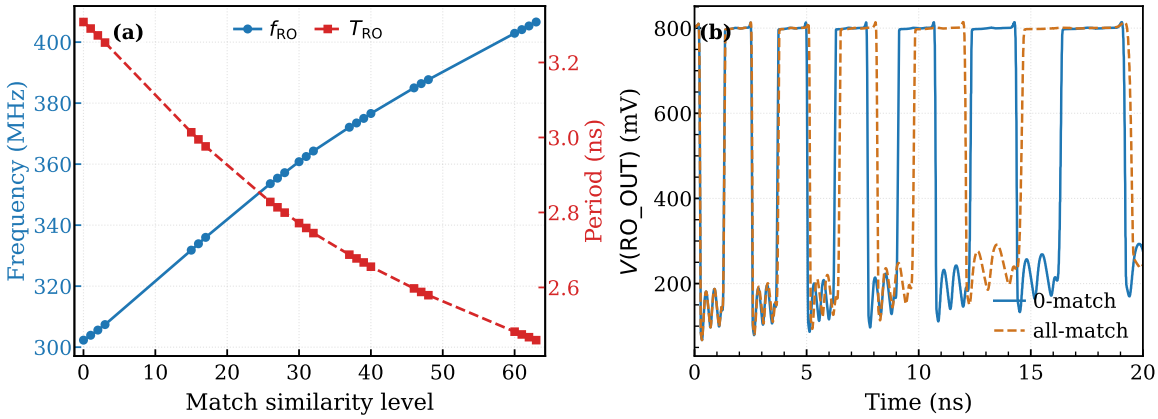


Figure 9.3.: (a) RO frequency and period versus match similarity and (b) RO output waveforms showing decreasing oscillation period with increasing similarity.

9.5.2. Analysis of aging and Drift Detection

Let X denote the per-read test statistic (RO frequency for each similarity level), and let τ be the decision threshold. A single read declares "anomaly" if $X \geq \tau$ (the opposite inequality can be used depending on the direction of the shift, without loss of generality). The per-read false-alarm and detection probabilities are defined as

$$P_{\mathrm{FA}}(\tau) \triangleq \Pr[X \geq \tau \mid \mathcal{H}_0], \tag{9.3}$$

$$P_{\mathrm{det}}(\tau) \triangleq \Pr[X \geq \tau \mid \mathcal{H}_1], \tag{9.4}$$

where $\mathcal{H}_0$ denotes the baseline (non-aged) distribution and $\mathcal{H}_1$ denotes the aged distribution. Sweeping τ produces the receiver operating characteristic (ROC) curve P_{det} versus P_{FA}.

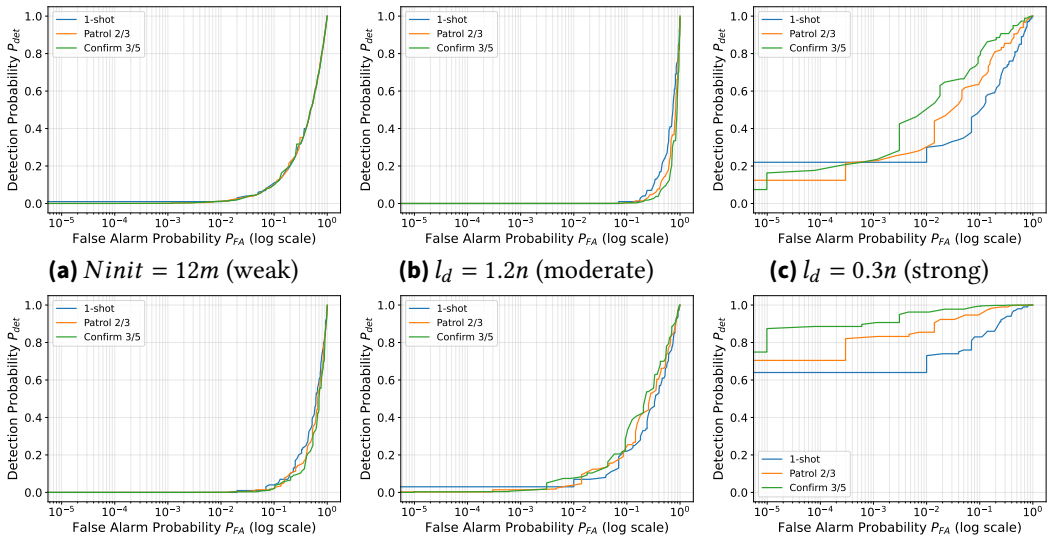
For multi-read k-of-n voting, an anomaly is declared if at least k out of n independent reads exceed the threshold. Let $\alpha \triangleq P_{\mathrm{FA}}(\tau)$ be the per-read false-alarm probability and $p \triangleq P_{\mathrm{det}}(\tau)$ be the per-read detection probability. The aggregated probabilities follow:

$$P_{\mathrm{FA}}^{(k/n)}(\tau) = \sum_{i=k}^{n} \binom{n}{i} \alpha^i (1-\alpha)^{n-i}, \tag{9.5}$$

$$P_{\mathrm{det}}^{(k/n)}(\tau) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}. \tag{9.6}$$

9.5.3. Fault Localization and X-Mask-Based Tolerance

When cumulative aging and drift variations significantly distort the ML discharge behavior and lead to incorrect similarity evaluations, a targeted localization procedure is triggered to identify the affected ML and its corresponding cell region. To further isolate the faulty cell, controlled diagnostic patterns are selectively applied to the identified ML while keeping other rows inactive, such as [111]. By observing variations in the frequency signature under carefully chosen match and mismatch conditions, the defective cell can be localized



(a) $Ninit = 12m$ (weak) **(b)** $l_d = 1.2n$ (moderate) **(c)** $l_d = 0.3n$ (strong)

(d) $r_f = 38n$ (weak) **(e)** $r_f = 51n$ (moderate) **(f)** $r_f = 90n$ (strong)

Figure 9.4.: ROC curves for VCRO-frequency drift detection under process variation, grouped by model parameter family ($Ninit$, r_f, l_d). Each subplot compares one-shot detection with two-stage k-of-n voting (Patrol: 2/3, Confirm: 3/5).

based on its contribution to the equivalent discharge resistance R_{eq}. This approach avoids full-array re-testing and limits the diagnosis overhead to the suspicious region only.

After localization, fault tolerance is achieved using an X-mask mechanism. The defective cell is logically masked during similarity evaluation so that its contribution to the ML discharge is neutralized. In practice, this is implemented by forcing the corresponding comparison outcome to a “don’t-care” state, preventing the faulty device from influencing the similarity result.

9.6. Evaluation and Results

9.6.1. Experiment Setup

Electrical-level SPICE simulations are performed to evaluate the impact of aging and resistance drift under the existence of process variations on the proposed VCRO-based sensing framework. A detailed circuit-level model of the ReCAM array is implemented, including the ML, access transistors, ReRAM devices, interconnect parasitics, and the VCRO sensing chain, as summarized in Table 9.1. Monte Carlo (MC) simulations are used to capture device- and interconnect-level variability and to quantify its impact on ML discharge behavior and the resulting RO frequency.

Input test patterns spanning the full similarity range from all-mismatch (0-match) to all-match (64-match) are applied to characterize ML discharge dynamics across different match levels. Since the separation between adjacent similarity levels is small, particularly at low match levels, the sensing margin is limited by the minimum inter-level spacing.

To improve resolvability, a parametric sweep of the ML load capacitance C_{L} is performed, as C_{L} directly determines the ML time constant and the resulting frequency separation

Table 9.1.: Simulation setup and parameters.

Category	Parameter Details
Technology-Level Parameters	
ReRAM Model [23]	JART VCM with read variability
Filament Radius (r_f)	45 nm
Disc Region Length (l_d)	0.6 nm
Initial Vacancy Concentration (N_{init})	HRS: 9 m; LRS: 3×10^{26} m^{-3}
Circuit-Level Parameters	
Simulation Tool	Cadence Virtuoso (SPICE)
CMOS Technology	GlobalFoundries 22FDX
Supply Voltage (V_{DD})	800 mV
Temperature	27°C
Interconnect per ReCAM Cell (22 nm) [47]	ML: $R = 11.95\,\Omega$, $C = 16.63$ aF
Match-Line Length	64-bit
Match-Line Load Capacitance (C_{L})	10 pF
VCRO Sensing Parameters	
Inverter Sizing (M_p/M_n)	$W_p/L_p = 320/20$, $W_n/L_n = 160/20$
Footer Device (M_{footer})	$W_f/L_f = 200/120$
Number of RO Stages (N)	7
RO Frequency Range	302–407 MHz
Performance and Area Parameters	
ReCAM Cell Area [108]	0.0242 μm^2
LUT Area	22.5 μm^2
ReCAM Cell Power	0.149 mW
LUT Power	2.95 mW

between similarity levels. This sweep allows balancing frequency, dynamic range, and inter-level spacing to maintain distinguishable signatures across the similarity range.

9.6.2. Aging and Drift Sensitivity Results

Beyond similarity dependence, (9.2) also captures the impact of device degradation on the sensing behavior. In advanced CMOS technologies, aging mechanisms such as bias temperature instability and hot-carrier injection cause threshold-voltage shifts and mobility degradation. In the VCRO sensing path, these effects increase the threshold voltage $V_{\text{th,st}}$ and/or reduce the gain factor β of the current-starving device, thereby lowering the starve current and the resulting oscillation frequency.

ReRAM degradation is modeled through variations in technology parameters such as the filament radius (r_f), disc region length (l_d), and minimum disc concentration (N_{discmin}),

Table 9.2.: Patrol and confirm voting results for the all-match (fault-free) baseline. Test coverage (TC) equals the probability of detection.

Aging parameter	**Patrol (2/3)**		**Confirm (3/5)**	
	P_{FA} (%)	TC (%)	P_{FA} (%)	TC (%)
Filament radius (r_f)				
$r_f = 51\,\text{nm}$	1.04	3.97	1.12	7.44
$r_f = 90\,\text{nm}$	1.04	85.48	1.12	96.25
Disc region length (l_d)				
$l_d = 0.15\,\text{nm}$	1.04	90.54	1.12	97.80
$l_d = 0.30\,\text{nm}$	1.04	32.35	1.12	53.75
Initial vacancy concentration (N_{init})				
$N_{\text{init}} = 12,\ 20,\ 90\,m$	1.04	1.40	1.12	1.43

P_{FA}: false-alarm probability. TC: test coverage (= P_{det}).

which modify filament conduction and the equivalent discharge resistance R_{eq}. These variations alter the ML discharge behavior and consequently shift the oscillation frequency.

As a result, aging and drift appear as systematic shifts in f_{RO} across similarity levels, while preserving the monotonic ordering between them. This property enables reliable detection of long-term performance drift without modifying the similarity evaluation mechanism.

Figure 9.5 illustrates RO frequency distributions for selected match levels (0-, 32-, and all-match) under baseline and degraded conditions ($r_f = 22.5\,\text{nm}$ and $90\,\text{nm}$). While moderate degradation causes only a small frequency shift, severe degradation produces a larger displacement, particularly at high similarity levels where the sensitivity to R_{eq} is stronger. This increased separation between baseline and degraded distributions explains the improved detection capability observed in the ROC analysis.

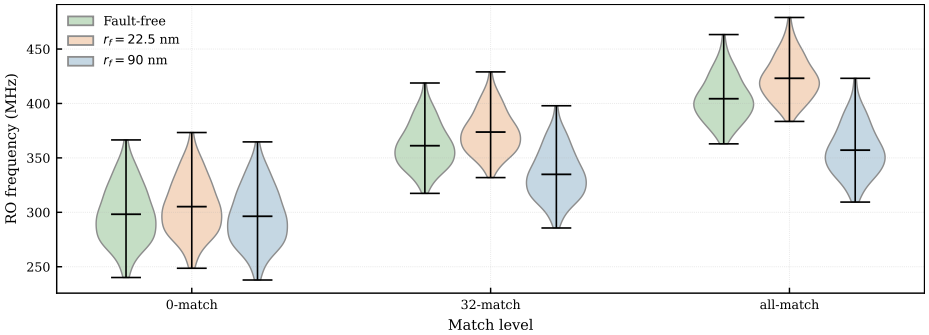


Figure 9.5.: RO frequency distributions across selected match levels under baseline and aged conditions ($r_f = 22.5\,\text{nm}$ and $90\,\text{nm}$).

9.6.3. Multi-Tier k-of-n Voting Strategy

Figure 9.4 presents the ROC curves for VCRO-frequency–based drift detection under different degradation scenarios. The horizontal axis shows the false-alarm probability P_{FA} measured on the baseline distribution, while the vertical axis shows the detection probability P_{det} evaluated on the degraded distribution. The curves are obtained by sweeping the decision threshold of the RO frequency signature. Three decision schemes are compared: single-shot detection, patrol (2/3), and confirm (3/5). As predicted by (9.5), the k-of-n aggregation suppresses sporadic outliers while reinforcing persistent deviations.

For weak parameter shifts (e.g., N_{init} variations), the ROC curves remain close to the diagonal, indicating strong overlap between the baseline and degraded distributions and, therefore, limited detection capability. As the magnitude of degradation increases (e.g., $r_f = 51\,\mathrm{nm}$), a clearer separation between the two populations emerges, and the ROC curves bend toward the upper-left region. In this regime, multi-read voting provides noticeable improvement, with the confirm (3/5) scheme achieving higher P_{det} at the same P_{FA}. For strong degradation (e.g., $r_f = 90\,\mathrm{nm}$ or $l_d = 0.15\,\mathrm{nm}$), the frequency distributions become clearly distinguishable, and the ROC curves approach the ideal upper-left corner, enabling near-unity detection probability even at very low ($P_{\mathrm{FA}} < 10^{-3}$).

Table 9.2 reports representative the test coverage values near $P_{\mathrm{FA}} \approx 10^{-2}$ for the all-match condition. The results confirm the ROC trends: as the degradation-induced frequency shift increases, the separation between baseline and aged populations becomes more pronounced, leading to significantly higher detection probabilities. In particular, strong degradation ($l_d = 0.15\,\mathrm{nm}$) achieves detection probabilities exceeding 90% for Patrol and 97.8%.

9.6.4. Overhead Analysis of BIST Hardware

Table 9.3 summarizes the hardware overhead of the proposed VCRO-based BIST relative to the functional ReCAM array (CUT) across different array configurations. The results show that the additional monitoring circuitry introduces negligible overhead compared with the array itself. Specifically, the VCRO-based BIST occupies only about 0.025%–0.030% of the ReCAM area, while the corresponding power overhead ranges from approximately 1.5% to 0.44%, depending on the number of MLs.

In contrast, conventional delay-based monitoring approaches typically rely on high-resolution ADCs or precise timing-measurement circuits to detect small degradation-induced variations caused by aging and resistance drift. In practice, this requires a significantly higher-resolution ADC (e.g., at least 12-bit), which substantially increases hardware overhead. Based on the performance parameters of such a 12-bit ADC [142], we find the area required by delay-based BIST is several orders of magnitude larger than that of the proposed VCRO-based approach.

Table 9.3.: Hardware overhead comparison between the functional ReCAM array (CUT) and the proposed VCRO-based BIST

Config.		Area (μm^2)			Power (mW)		
B	N_{ML}	CUT	VCRO	VCRO/CUT	CUT	VCRO	VCRO/CUT
32	16	0.0950	2.39×10^{-5}	0.025%	205.8	3.12	1.52%
64	16	0.0950	2.39×10^{-5}	0.025%	205.6	3.12	1.51%
32	64	0.0951	2.83×10^{-5}	0.030%	823.2	3.61	0.439%
64	64	0.0951	2.83×10^{-5}	0.030%	825.2	3.61	0.438%

*B: number of cells per match line; N_{ML}: number of match lines.

9.7. Summary

This chapter presented a lightweight runtime BIST framework for the reliable operation of ReCAM under aging and runtime variations. The chapter first highlighted the limitations of conventional manufacturing testing and calibration techniques, showing that gradual degradation mechanisms such as resistance drift and aging can progressively distort the ML behavior during field operation. A motivational example demonstrated that aging effects become more pronounced at higher similarity levels, motivating the need for continuous runtime monitoring.

To address these challenges, a frequency-domain monitoring framework based on a VCRO was introduced. The proposed approach converts ML-dependent analog behavior into frequency signatures, enabling compact runtime anomaly detection. In addition, a two-stage statistical voting mechanism was employed to improve robustness against transient variations while maintaining low false-alarm probability.

The proposed framework was evaluated under different aging conditions and ReCAM configurations. Simulation results demonstrated high detection coverage for severe degradation scenarios while introducing negligible area and power overhead. Furthermore, selective fault masking using an X-mask mechanism enables continued functionality without requiring costly full-array re-testing.

Overall, this chapter demonstrated that frequency-domain sensing provides an efficient and scalable solution for runtime reliability monitoring in ReCAM architectures, enabling reliable long-term operation with minimal hardware budget overhead.

10. Conclusion and Perspectives

10.1. Conclusion

In this dissertation, we investigated the reliability, testing, and sensing challenges of emerging analog CiM architectures, with a primary focus on ReCAM. We demonstrated that the scalability, manufacturability, and energy efficiency of analog CiM architectures can be fully realized only when process variation, sensing uncertainty, fault masking, and runtime degradation are carefully addressed through reliability- and quality-aware methodologies.

Chapter 3 presented a comprehensive reliability analysis and mitigation framework for analog NVM-based CiM architectures spanning from technology-level variability to circuit- and application-level reliability implications. We demonstrated how device-level non-idealities and sensing uncertainties propagate through the computation flow and affect the robustness of analog in-memory computation. Furthermore, we investigated mitigation methodologies across multiple abstraction levels to improve the reliability of analog CiM operations.

Chapter 4 focused on fault modeling and testing challenges in ReCAM arrays. We analyzed the limitations of conventional testing approaches for analog ReCAM architectures and introduced dedicated fault-modeling and testing methodologies that capture the unique behavior of analog similarity-search operations. The presented analysis highlighted the growing complexity of testing analog ReCAM arrays under device variability and sensing uncertainty.

Chapter 5 investigated the impact of process variation on diagnosis resolution in analog ReCAM arrays. We demonstrated how process variation significantly reduces the distinguishability between neighboring fault locations due to overlapping analog signatures. To address this challenge, we proposed a PVT-variation-aware fault diagnosis framework that leverages adaptive amplification and multi-step diagnosis to improve fault localization while maintaining low hardware overhead.

Chapter 6 investigated the limitations of conventional March tests under process variation and showed how variation-induced shifts can mask defects and lead to test escapes. We demonstrated that traditional testing methodologies are insufficient in the presence of process variation and proposed process-variation-aware testing methodologies that reliably detect both physical and variation-induced faults in ReCAM arrays.

Chapter 7 introduced a post-manufacturing calibration methodology for ReCAM arrays. The proposed calibration framework compensates for global variation effects by generating per-chip lookup tables, thereby improving post-manufacturing yield while preserving distinguishability between similarity levels within each chip. The proposed methodology

Table 10.1.: Summary of the main contributions of this dissertation.

Chapter	Problem	Contribution	Phase
3	Decision failure	Cross-level reliability analysis and mitigation	Analysis
4	Manufacturing defects	Fault modeling and March-like testing methodology	Testing
5	Fault localization	DfT-assisted diagnosis framework for ReCAM	Diagnosis
6	Process variation masking	PV-aware testing methodology	Testing
7	Yield loss due to variations	Post-manufacturing calibration using LUT-based mapping	Calibration
8	Similarity-level overlap and sensing limitations	Time-domain sensing architecture (TDsReCAM)	Sensing
9	Aging and drift during operation	VCRO-based runtime BIST and monitoring	Runtime Monitoring

further demonstrated the importance of calibration-aware reliability enhancement for analog ReCAM architectures.

Chapter 8 proposed a fundamentally different sensing paradigm for reliable ReCAM architectures through time-domain sensing. Instead of relying on conventional voltage-domain digitization using high-resolution ADCs, the proposed architecture converts match-line discharge behavior into temporal information using a compact TDC. The proposed time-domain sensing methodology improves robustness against process variation while significantly reducing sensing complexity, hardware overhead, and scalability limitations associated with conventional analog sensing approaches.

Finally, Chapter 9 focused on runtime reliability monitoring for ReCAM arrays under aging and temporal degradation. We proposed a lightweight runtime BIST framework based on frequency-domain monitoring using VCRO to continuously track match-line behavior through frequency signatures. The proposed methodology enables low-overhead online monitoring, runtime anomaly detection, and reliability tracking for aging-aware ReCAM operation.

This dissertation presents a comprehensive reliability and quality assurance framework for analog CiM architectures by jointly addressing reliability analysis, fault modeling, testing, diagnosis, calibration, runtime monitoring, and sensing, as illustrated in Table 10.1. The proposed approaches collectively improve the reliability, testability, scalability, and manufacturability of next-generation in-memory computing architectures.

10.2. Future Perspectives

The work carried out in this dissertation has provided insights into reliability- and quality-aware methodologies for analog CiM architectures, with a particular focus on testing, diagnosis, calibration, runtime monitoring, and sensing optimization for ReCAM arrays. The presented methodologies demonstrated the importance of cross-level co-optimization across device, circuit, architecture, and test layers to improve reliability while maintaining low hardware overhead, scalability, and energy efficiency.

Although this dissertation primarily focused on ReCAM architectures, the proposed reliability-aware methodologies can potentially be extended to other emerging analog and in-memory computing paradigms. In particular, future CiM architectures based on alternative non-volatile memory technologies, such as FeFET, PCM, or hybrid memory technologies, may also experience similar reliability challenges related to process variation, sensing uncertainties, runtime degradation, and defect masking [143], [144]. Consequently, the testing, diagnosis, and calibration methodologies presented in this dissertation can serve as a foundation for reliability- and quality-aware design in future in-memory computing architectures.

Furthermore, as analog CiM architectures continue to scale toward higher-density arrays and lower operating voltages, process variation and sensing resolution limitations are expected to become increasingly critical. Future research can therefore investigate more adaptive and self-correcting sensing methodologies capable of dynamically compensating for variation, aging, and environmental fluctuations during runtime operation. The integration of machine-learning-assisted calibration and adaptive sensing frameworks also represents a promising research direction for improving long-term reliability and yield [145], [146].

In addition, while this dissertation explored time-domain sensing as an alternative to conventional voltage-domain sensing, further research can investigate fully time-based or hybrid sensing architectures for analog CiM arrays. Future work can investigate adaptive, low-overhead time-domain sensing methodologies to improve scalability while reducing the complexity and power consumption of conventional high-resolution sensing circuits [147], [148].

Another important future research direction involves runtime reliability monitoring and predictive failure analysis. Although this dissertation introduced a lightweight runtime BIST framework based on frequency-domain monitoring, future work may investigate more advanced online monitoring techniques that can predict gradual degradation and aging effects before functional failure occurs. Combining runtime sensing with dynamic calibration and self-recovery mechanisms can further improve the reliability and lifetime of future ReCAM architectures [149], [150].

Finally, enhancing existing Electronic Design Automation (EDA) tools to better support analog CiM architectures while incorporating reliability, testing, and manufacturability objectives during the design process remains an important open challenge. Developing reliability-aware design automation frameworks that jointly optimize performance, energy efficiency, sensing adaptivity, and testability can significantly accelerate the practical deployment of next-generation analog in-memory computing architectures [151], [152].

Part III.

Appendix

Bibliography

[1] M. Mayahinia, H. G. Hezayyin, and M. B. Tahoori, "Reliable Analog Computation-in-Memory: From Technology to Application", in *Proceedings of the IEEE VLSI Test Symposium (VTS)*, 2024.

[2] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Fault Modeling and Testing of ReRAM-based CAM Array", in *Proceedings of the IEEE VLSI Test Symposium (VTS)*, 2025.

[3] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "PV-ReCAM: Process Variation–Aware Testing for ReRAM-based CAM", in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2025.

[4] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Design-for-Testability for ReRAM-based CAM", in *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS)*, 2025.

[5] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Tdsrecam: Time-domain sensing for reliable reram-based content-addressable memory", in *Proceedings of the IEEE European Test Symposium (ETS)*, accepted, 2026.

[6] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "PVTVAD: PVT-Variation–Aware Fault Diagnosis in ReRAM-based CAM Arrays", *IEEE Transactions on Device and Materials Reliability (TDMR)*, 2026, submitted.

[7] H. G. Hezayyin and M. B. Tahoori, "Runtime BIST for ReRAM-based CAM using Frequency-Domain Monitoring", in *Proceedings of the IEEE International On-Line Testing Symposium (IOLTS)*, accepted for publication, 2026.

[8] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Testing ReRAM-based TCAM for Computation-in-Memory Applications", in *Proceedings of the IEEE International Conference on Design and Technology of Integrated Systems in Nanoscale Era (DTIS)*, 2024.

[9] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "A Post-Manufacturing Calibration Method for ReRAM-based CAM", in *Proceedings of the IEEE VLSI Test Symposium (VTS)*, accepted for publication, 2026.

[10] W. A. Wulf and S. A. McKee, "Hitting the memory wall: Implications of the obvious", *ACM SIGARCH Computer Architecture News*, vol. 23, no. 1, pp. 20–24, 1995. DOI: 10.1145/216585.216588.

[11] V. Sze, Y.-H. Chen, T.-J. Yang, and J. Emer, *Efficient Processing of Deep Neural Networks*. Morgan & Claypool, 2017.

[12] A. Rahimi, P. Kanerva, and J. M. Rabaey, "A robust and energy-efficient classifier using brain-inspired hyperdimensional computing", *Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 64–69, 2016. DOI: 10.1145/2934583.2934624.

[13] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences", *Bioinformatics*, vol. 34, no. 18, pp. 3094–3100, 2018. DOI: 10.1093/bioinformatics/bty191.

[14] M. Mitzenmacher, "Compressed bloom filters", *IEEE/ACM Transactions on Networking*, vol. 10, no. 5, pp. 604–612, 2002. DOI: 10.1109/TNET.2002.803864.

[15] A. e. a. Sebastian, "Memory devices and applications for in-memory computing", *Nature Nanotechnology*, vol. 15, pp. 529–544, 2020.

[16] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices", *Nature Electronics*, vol. 1, no. 6, pp. 333–343, 2018.

[17] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (cam) circuits and architectures: A tutorial and survey", *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, 2006. DOI: 10.1109/JSSC.2005.864128.

[18] H.-S. P. Wong, S. Raoux, S. Kim, et al., "Metal-oxide rram", *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012. DOI: 10.1109/JPROC.2012.2190369.

[19] G. W. Burr, R. M. Shelby, S. Sidler, et al., "Neuromorphic computing using non-volatile memory", *Advances in Physics: X*, vol. 2, no. 1, pp. 89–124, 2017. DOI: 10.1080/23746149.2016.1259585.

[20] D. Ielmini, "Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling", *Semiconductor Science and Technology*, vol. 31, no. 6, p. 063 002, 2016. DOI: 10.1088/0268-1242/31/6/063002.

[21] H. Li, Q. Li, et al., "Recam: Resistive content addressable memory for in-memory computing", in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, 2016, pp. 721–724. DOI: 10.1109/ISCAS.2016.7527314.

[22] X. Guo, E. Ipek, and T. Soyata, "Resistive content addressable memory systems: Architecture, performance and scalability", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 2, pp. 687–698, 2017. DOI: 10.1109/TVLSI.2016.2618869.

[23] C. Bengel, A. Siemon, et al., "Reliable conductive bridge ram based on current-controlled set operation", *IEEE Transactions on Electron Devices*, vol. 67, no. 6, pp. 2613–2618, 2020. DOI: 10.1109/TED.2020.2986158.

[24] Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing", *Nature Materials*, vol. 18, no. 4, pp. 309–323, 2019. DOI: 10.1038/s41563-019-0291-x.

[25] S. Ambrogio, S. Balatti, et al., "Equivalent-accuracy accelerated neural-network training using analogue memory", *Nature*, vol. 558, no. 7708, pp. 60–67, 2018. DOI: 10.1038/s41586-018-0180-5.

[26] A. J. van de Goor, *Testing Semiconductor Memories: Theory and Practice*. John Wiley & Sons, 1991.

[27] ITRS, *International technology roadmap for semiconductors*, 2013.

[28] K. E. Petersen, "Silicon as a mechanical material", *Proceedings of the IEEE*, vol. 70, no. 5, pp. 420–457, 1982. DOI: 10.1109/PROC.1982.12331.

[29] M. A. Alam et al., "3d nand flash memory: A survey", *IEEE Transactions on Device and Materials Reliability*, 2017.

[30] A. S. Spinelli, C. M. Compagnoni, and A. L. Lacaita, "Reliability of nand flash memories: Planar cells and emerging issues in 3d devices", *Computers*, vol. 6, no. 2, p. 16, 2017. DOI: 10.3390/computers6020016.

[31] S. Raoux, W. Welnic, and D. Ielmini, "Phase change materials and their application to nonvolatile memories", *Chemical Reviews*, vol. 110, no. 1, pp. 240–267, 2010. DOI: 10.1021/cr900040x.

[32] J.-P. Wang et al., "Prospects of spin-transfer torque mram", *IEEE Transactions on Magnetics*, 2012.

[33] K. Garello et al., "Manufacturable 300mm platform solution for field induced magnetic random access memory", in *IEEE International Electron Devices Meeting (IEDM)*, 2009.

[34] T. Mikolajick, S. Slesazeck, M. H. Park, and U. Schroeder, "Ferroelectric hafnium oxide for ferroelectric random-access memories and ferroelectric field-effect transistors", *MRS Bulletin*, vol. 43, no. 5, pp. 340–346, 2018.

[35] M. H. Park et al., "Ferroelectric field-effect transistor memory based on hfo2 thin films", *MRS Bulletin*, 2013.

[36] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices", *Nature Electronics*, vol. 1, no. 6, pp. 333–343, 2018. DOI: 10.1038/s41928-018-0092-2.

[37] S. Menzel, "Rram—from material to application", *IEEE Transactions on Device and Materials Reliability*, 2015.

[38] K.-M. Kim, B. J. Choi, and C. S. Hwang, "Materials and mechanisms of resistive random access memory", *Journal of Applied Physics*, 2004.

[39] 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. 4125–4133, 2020.

[40] M.-F. Chang, J.-J. Wu, T.-F. Chien, Y.-C. Liu, T.-C. Yang, W.-C. Shen, Y.-C. King, C.-J. Lin, K.-F. Lin, Y.-D. Chih, S. Natarajan, and J. Chang, "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.

[41] S. T. Ahmed, M. Hefenbrock, C. Münch, and M. B. Tahoori, "Neuroscrub: Mitigating retention failures using approximate scrubbing in neuromorphic fabric based on resistive memories", in *2021 IEEE European Test Symposium (ETS)*, 2021, pp. 1–6. DOI: 10.1109/ETS50041.2021.9465446.

[42] M. Mayahinia, C. Münch, and M. B. Tahoori, "Analyzing and mitigating sensing failures in spintronic-based computing in memory", in *2021 IEEE International Test Conference (ITC)*, 2021, pp. 268–277. DOI: 10.1109/ITC50571.2021.00036.

[43] M. Mayahinia, A. Jafari, and M. B. Tahoori, "Voltage tuning for reliable computation in emerging resistive memories", in *2022 IEEE 40th VLSI Test Symposium (VTS)*, 2022, pp. 1–7. DOI: 10.1109/VTS52500.2021.9794233.

[44] P. Chi, S. Li, C. Xu, T.-J. Liu, J. Zhang, et al., "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory", *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 27–39, 2016. DOI: 10.1145/3007787.3001140.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] W. Guo et al., "Reram-based content addressable memory for approximate matching", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017.

[52] H. Li et al., "Tcam and reram-based associative computing architectures", *IEEE Micro*, 2016.

[53] B. Murmann, "A/d converter trends: Power dissipation, scaling and digitally assisted architectures", *IEEE Custom Integrated Circuits Conference*, 2008.

[54] R. H. Walden, "Analog-to-digital converter survey and analysis", *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, 1999.

[55] T. S. Copetti, T. Gemmeke, and L. M. B. Poehls, "A dft strategy for detecting emerging faults in rrams", in *IFIP/IEEE International Conference on Very Large Scale Integration-System on a Chip*, Springer, 2021, pp. 93–111.

[56] C. Wang, H. Wu, B. Gao, T. Zhang, Y. Yang, and H. Qian, "Conduction mechanisms, dynamics and stability in rerams", *Microelectronic Engineering*, vol. 187-188, pp. 121–133, 2018, ISSN: 0167-9317. DOI: https://doi.org/10.1016/j.mee.2017.11.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167931717303672.

[57] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of mos transistors", *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[58] S. R. Nassif, "Modeling and analysis of manufacturing variations", *IEEE Custom Integrated Circuits Conference*, 2001.

[59] 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.

[60] Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing", *Nature Materials*, vol. 18, no. 4, pp. 309–323, 2019.

[61] Y. Cai, E. Haratsch, O. Mutlu, and K. Mai, "Error characterization, mitigation, and recovery in flash-memory-based solid-state drives", *Proceedings of the IEEE*, vol. 105, no. 9, pp. 1666–1704, 2017.

[62] A. e. a. Shafiee, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars", in *ISCA*, 2016, pp. 14–26.

[63] P. e. a. Chi, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory", in *ISCA*, 2016.

[64] D. Ielmini, "Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling", *Semiconductor Science and Technology*, vol. 31, no. 6, p. 063 002, 2016.

[65] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys", *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[66] M. L. Bushnell and V. D. Agrawal, *Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits*. Springer, 2000.

[67] M. Abramovici, M. A. Breuer, and A. D. Friedman, *Digital Systems Testing and Testable Design*, 1st. IEEE Press, 1990.

[68] A. J. van de Goor and Z. Al-Ars, *Functional Memory Testing: From Bits to High-Density Memories*, 2nd. Springer, 2018.

[69] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, "Matching properties of mos transistors", *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

[70] 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.

[71] L. Xie, H. Du Nguyen, J. Yu, A. Kaichouhi, M. Taouil, M. AlFailakawi, and S. Hamdioui, "Scouting logic: A novel memristor-based logic design for resistive computing", in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2017, pp. 176–181. DOI: 10.1109/ISVLSI.2017.39.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] M. Mayahinia, M. Tahoori, M. P. Komalan, H. Zahedmanesh, K. Croes, T. Marinelli, J. I. G. Pérez, T. Evenblij, G. S. Kar, and F. Catthoor, "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.

[78] 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.

[79] J. Song, Y. Park, C. Lim, Y. Choi, S. Ahn, S. Park, and C. Kim, "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.

[80] Y.-X. Chen and J.-F. Li, "Fault modeling and testing of 1t1r memristor memories", *Advanced Reliable Systems (ARES) Lab. Department of Electrical Engineering, National Central University, Taoyuan, Taiwan*, 2020.

[81] K.-J. Lin and C.-W. Wu, "Testing content-addressable memories using functional fault models and march-like algorithms", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 5, pp. 577–588, 2000. DOI: 10.1109/43.845082.

[82] Y. .-. Yang and J.-F. Li, "Fault modeling and testing of rram-based computing-in memories", in *2022 IEEE International Test Conference in Asia (ITC-Asia)*, 2022, pp. 7–12. DOI: 10.1109/ITCAsia55616.2022.00012.

[83] P. Liu, Z. You, J. Kuang, Z. Hu, H. Duan, and W. Wang, "Efficient march test algorithm for 1t1r cross-bar with complete fault coverage", *Electronics Letters*, vol. 52, no. 18, pp. 1520–1522, 2016. DOI: https://doi.org/10.1049/el.2016.1693. eprint: https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/el.2016.1693. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/el.2016.1693.

[84] H. Sribhuvaneshwari and K. Suthendran, "A novel march c2rr algorithm for nano-electronic resistive random access memory (rram) testing", in *VLSI Design and Test*, S. Rajaram, N. Balamurugan, D. Gracia Nirmala Rani, and V. Singh, Eds., Singapore: Springer Singapore, 2019, pp. 578–589, ISBN: 978-981-13-5950-7.

[85] Y.-X. Chen and J.-F. Li, "Fault modeling and testing of 1t1r memristor memories", in *2015 IEEE 33rd VLSI Test Symposium (VTS)*, IEEE, 2015, pp. 1–6.

[86] P. Liu, Z. You, J. Wu, B. Liu, Y. Han, and K. Chakrabarty, "Fault modeling and efficient testing of memristor-based memory", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 11, pp. 4444–4455, 2021.

[87] L. Sun, N. Zheng, T. Zhang, and P. Mazumder, "Fault modeling and parallel testing for 1t1m memory array", *IEEE Transactions on Nanotechnology*, vol. 17, no. 3, pp. 437–451, 2018.

[88] S. Kannan, J. Rajendran, R. Karri, and O. Sinanoglu, "Sneak-path testing of memristor-based memories", in *2013 26th International Conference on VLSI Design and 2013 12th International Conference on Embedded Systems*, 2013, pp. 386–391. DOI: 10.1109/VLSID.2013.219.

[89] M. Fieback, F. Bradarić, M. Taouil, and S. Hamdioui, "Online fault detection and diagnosis in rram", in *2023 IEEE European Test Symposium (ETS)*, 2023, pp. 1–6. DOI: 10.1109/ETS56758.2023.10174113.

[90] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Red survey of reram-based architectures for processing-in- memory ox-based resistive switching memories–nanoionic mechanisms, prospects, and challenges", *Advanced Materials*, vol. 21, no. 25-26, pp. 2632–2663, 2009. DOI: 10.1002/adma.200900375.

[91] H. Karami, R. Zarandi, and R. Zarandi, "Testing of sram-based content addressable memory", *IEEE Transactions on Computers*, vol. 49, no. 10, pp. 1054–1058, 2000.

[92] P. Sidorowicz and J. Brzozowski, "An approach to modeling and testing memories and its application to cams", in *Proceedings. 16th IEEE VLSI Test Symposium (Cat. No.98TB100231)*, 1998, pp. 411–416. DOI: 10.1109/VTEST.1998.670899.

[93] J.-F. Li, R.-S. Tzeng, and C.-W. Wu, "Testing and diagnosing embedded content addressable memories", in *Proceedings 20th IEEE VLSI Test Symposium (VTS 2002)*, 2002, pp. 389–394. DOI: 10.1109/VTS.2002.1011169.

[94] T. Copetti, M. Fieback, T. Gemmeke, and et al., "A dft strategy for guaranteeing reram's quality after manufacturing", *J Electron Test*, vol. 40, pp. 245–257, 2024. DOI: 10.1007/s10836-024-06108-8.

[95] M. Zhang, Y. Zhu, C.-H. Chan, and R. P. Martins, "An 8-bit 10-gs/s 16× interpolation-based time-domain adc with <1.5-ps uncalibrated quantization steps", *IEEE Journal of Solid-State Circuits*, vol. 55, no. 12, pp. 3225–3235, 2020. DOI: 10.1109/JSSC.2020.3012776.

[96] C. Li, C. E. Graves, X. Sheng, D. Miller, M. Foltin, G. Pedretti, and J. P. Strachan, "Analog content-addressable memories with memristors", *Nature Communications*, vol. 11, no. 1, p. 1638, Apr. 2020. DOI: 10.1038/s41467-020-15254-4. [Online]. Available: https://doi.org/10.1038/s41467-020-15254-4.

[97] Y. Halawani, B. Mohammad, M. Abu Lebdeh, M. Al-Qutayri, and S. F. Al-Sarawi, "Reram-based in-memory computing for search engine and neural network applications", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 2, pp. 388–397, 2019. DOI: 10.1109/JETCAS.2019.2909317.

[98] R. Sharma, N. S. Dhakad, G. S. Reddy, V. Sharma, and S. K. Vishvakarma, "Recam: Resistive ram digital content addressable memory using novel 3t1r bitcell", in *2024 8th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*, 2024, pp. 1–3. DOI: 10.1109/EDTM58488.2024.10511788.

[99] G. Pedretti, C. Graves, T. Vaerenbergh, S. Serebryakov, M. Foltin, X. Sheng, R. Mao, C. Li, and J. P. Strachan, "Differentiable content addressable memory with memristors", *Advanced Electronic Materials*, vol. 8, Mar. 2022. DOI: 10.1002/aelm.202101198.

[100] C. Münch, J. Yun, M. Keim, and M. B. Tahoori, "Mbist-based trim-search test time reduction for stt-mram", in *2022 IEEE 40th VLSI Test Symposium (VTS)*, 2022, pp. 1–7. DOI: 10.1109/VTS52500.2021.9794178.

[101] S. B. Mamaghani, J. Yun, M. Keim, and M. Tahoori, "Multi-level reference for test coverage enhancement of resistive-based nvm", in *2024 IEEE 42nd VLSI Test Symposium (VTS)*, 2024, pp. 1–7. DOI: 10.1109/VTS60656.2024.10538585.

[102] A. Singh, M. Fieback, R. Bishnoi, F. Bradarić, A. Gebregiorgis, R. V. Joshi, and S. Hamdioui, "Accelerating rram testing with a low-cost computation-in-memory based dft", in *2022 IEEE International Test Conference (ITC)*, 2022, pp. 400–409. DOI: 10.1109/ITC50671.2022.00085.

[103] M. Fieback, G. C. Medeiros, L. Wu, H. Aziza, R. Bishnoi, M. Taouil, and S. Hamdioui, "Defects, fault modeling, and test development framework for rrams", *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 18, no. 3, pp. 1–26, 2022. DOI: 10.1145/3510851. [Online]. Available: https://doi.org/10.1145/3510851.

[104] S. Martirosyan and G. Harutyunyan, "An efficient fault detection and diagnosis methodology for volatile and non-volatile memories", in *2019 Computer Science and Information Technologies (CSIT)*, 2019, pp. 47–51. DOI: 10.1109/CSITechnol.2019.8895189.

[105] H. Aziza, J. Plantier, and J.-M. Portal, "Non volatile memory signatures extraction for defects diagnosis purpose", in *5th International Conference on Design & Technology of Integrated Systems in Nanoscale Era*, 2010, pp. 1–5. DOI: 10.1109/DTIS.2010.5487556.

[106] J.-F. Li, R.-S. Tzeng, and C.-W. Wu, "Testing and diagnosis methodologies for embedded content addressable memories", *Journal of Electronic Testing*, vol. 19, no. 2, pp. 207–215, 2003, ISSN: 1573-0727. DOI: 10.1023/A:1022858128485. [Online]. Available: https://doi.org/10.1023/A:1022858128485.

[107] J.-F. Li, "Testing and diagnosing comparison faults of tcams with asymmetric cells", *IEEE Transactions on Computers*, vol. 61, no. 11, pp. 1576–1587, 2012. DOI: 10.1109/TC.2011.196.

[108] J. Li, R. K. Montoye, M. Ishii, and L. Chang, "1 mb 0.41 μm² 2t-2r cell nonvolatile tcam with two-bit encoding and clocked self-referenced sensing", *IEEE Journal of Solid-State Circuits*, vol. 49, pp. 896–907, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:9995754.

[109] Y.-C. Shih et al., "Logic process compatible 40nm 16mb, embedded perpendicular-mram with hybrid-resistance reference, sub-μa sensing resolution, and 17.5ns read access time", in *2018 IEEE Symposium on VLSI Circuits*, Jun. 2018, pp. 79–80. DOI: 10.1109/VLSIC.2018.8502260.

[110] D. Chang et al., "Algorithmic read resistance trim for improving yield and reducing test time in mram", in *2023 IEEE International Test Conference (ITC)*, Oct. 2023, pp. 87–92. DOI: 10.1109/ITC51656.2023.00022.

[111] H. G. Hezayyin, M. Mayahinia, and M. Tahoori, *Fault diagnosis in recam arrays*, 2025. DOI: 10.1109/IOLTS65288.2025.11117144.

[112] C. S. Vatti, R. M B, P. R. T. Reddy, and A. Acharyya, "A hierarchical fault-tolerant and cost effective framework for rram based neural computing systems", *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 3, pp. 684–688, 2022. DOI: 10.1109/TCSII.2022.3144193.

[113] L. Xia, W. Huangfu, T. Tang, X. Yin, K. Chakrabarty, Y. Xie, Y. Wang, and H. Yang, "Stuck-at fault tolerance in rram computing systems", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 36, no. 10, pp. 1673–1685, 2017.

[114] L. Lavoine, R. Auger, and Y. Savaria, "Fault-tolerant design techniques for resistive ram memories", *ACM Journal on Emerging Technologies in Computing Systems*, vol. 13, no. 1, pp. 1–27, 2017.

[115] H. G. Hezayyin, M. Mayahinia, and M. Tahoori, "Testing reram-based tcam for computation-in-memory applications", in *2024 IEEE International Conference on Design, Test and Technology of Integrated Systems (DTTIS)*, 2024, pp. 1–4. DOI: 10.1109/DTTIS62212.2024.10780399.

[116] S.-H. Chang, S.-Y. Chen, C.-W. Yang, H.-W. Huang, Y.-C. Yang, W.-L. Chen, C.-N. Liu, and H.-M. Chen, "Mitigating power and process variation for analog cim design migration", in *2024 20th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, 2024, pp. 1–4. DOI: 10.1109/SMACD61181.2024.10745463.

[117] D. Yi, S. Lee, I. Choi, G. Yun, E. J. Choi, J. Park, J. Kwak, S.-J. Jang, S. Ha, I.-J. Chang, and M. Je, "Skew-cim: Process-variation-resilient and energy-efficient computation-in-memory design technique with skewed weights", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 71, no. 5, pp. 2067–2078, 2024. DOI: 10.1109/TCSI.2024.3372392.

[118] T.-L. Tsai, J.-F. Li, C.-L. Hsu, and C.-T. Sun, "Testing of in-memory-computing memories with 8 t srams", *Microelectronics Reliability*, vol. 123, p. 114 215, 2021, ISSN: 0026-2714. DOI: https://doi.org/10.1016/j.microrel.2021.114215. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0026271421001815.

[119] L. Ammoura, M.-L. Flottes, P. Girard, J.-P. Noel, and A. Virazel, "A novel march test algorithm for testing 8t sram-based imc architectures", in *DATE 2024 - 27th Design, Automation & Test in Europe Conference & Exhibition*, Valence, Spain, Mar. 2024, pp. 1–6. DOI: 10.23919/DATE58400.2024.10546583. [Online]. Available: https://hal.archives-ouvertes.fr/hal-04658584.

[120] D. Kim, J. Park, C. Shin, J. Jung, K. Shin, S. Baek, S. Heo, W. Kim, I. Jeong, J. Cho, and J. Park, "GLOVA: Global and Local Variation-Aware Analog Circuit Design with Risk-Sensitive Reinforcement Learning", in *Proceedings of the 62nd ACM/IEEE Design Automation Conference (DAC)*, Accepted for publication, San Francisco, CA, USA, Jul. 2025. DOI: 10.48550/arXiv.2505.11208. arXiv: 2505.11208 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2505.11208.

[121] D. Ielmini, "Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling", *Semiconductor Science and Technology*, vol. 31, no. 6, p. 063 002, 2016. DOI: 10.1088/0268-1242/31/6/063002.

[122] C. Bengel, A. Siemon, S. Menzel, R. Waser, V. Rana, S. Hoffmann-Eifert, T. Breuer, N. v. d. Driesch, B. Temesvári, F. Lentz, 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.

[123] O. Mutlu, "Processing data where it makes sense in modern computing systems: Enabling in-memory computation", in *2018 Mediterranean Conference on Embedded Computing (MECO)*, IEEE, 2018, pp. 1–9.

[124] M. Fieback, F. Bradaric, P.-E. Gaillardon, F. Sebastiano, and E. I. Vatajelu, "Online fault detection and diagnosis in rram", in *2023 IEEE European Test Symposium (ETS)*, IEEE, 2023, pp. 1–6.

[125] M. Mayahinia, S. Thomann, P. R. Genssler, C. Münch, H. Amrouch, and M. B. Tahoori, "Algorithm to technology co-optimization for cim-based hyperdimensional computing", in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2024, pp. 1–6. DOI: 10.23919/DATE58400.2024.10546842.

[126] L. Zhao, L. Buonanno, R. M. Roth, S. Serebryakov, A. Gajjar, J. Moon, J. Ignowski, and G. Pedretti, "Race-it: A reconfigurable analog cam crossbar engine for in-memory transformer acceleration", *arXiv preprint arXiv:2312.06532*, 2023. [Online]. Available: https://arxiv.org/abs/2312.06532.

[127] Y. Xi et al., "Reliability of memristive devices for high-performance neuromorphic computing: (invited paper)", in *Proceedings of the 2023 IEEE International Reliability Physics Symposium (IRPS)*, Monterey, CA, USA: IEEE, 2023, pp. 1–7. DOI: 10.1109/IRPS48203.2023.10118214.

[128] F. Liu, W. Zhao, Z. Wang, X. Yang, and L. Jiang, "Simsnn: A weight-agnostic reram-based search-in-memory engine for snn acceleration", in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2023, pp. 1–2. DOI: 10.23919/DATE56975.2023.10136973.

[129] Z. Li, S. Liu, J. Zhang, R. Xu, X. Zhu, H. Yang, K. Zhu, P. Wang, H.-S. Wong, and S. Yu, "Rram-based cam combined with time-domain circuits for hyperdimensional computing", *Nature Electronics*, vol. 4, no. 3, pp. 312–322, 2021.

[130] Y. Liu, X. Hu, Y. Liu, and J. Lu, "Logic operation-based design for testability method and parallel test algorithm for 1t1r crossbar memory", *Electronics Letters*, vol. 53, no. 13, pp. 877–879, 2017. DOI: 10.1049/el.2016.1693.

[131] J. Lou, F. Freye, C. Lanius, and T. Gemmeke, "A 22nm 96.83-tops/w time-domain compute-in-memory engine utilizing mixed-fidelity for edge-ai applications", in *Proceedings of the Great Lakes Symposium on VLSI (GLSVLSI)*, 2025, pp. 1–6. DOI: 10.1145/3716368.3735161.

[132] F. Moradi, H. Li, et al., "Time-domain computing in memory using spintronic devices", *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 537–550, 2018.

[133] A. Ankit, I. E. Hajj, S. Chalamalasetti, G. Ndu, M. Foltin, R. S. Williams, and P. Faraboschi, "Puma: A programmable neural network accelerator based on memristor arrays", in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2019, pp. 1–15.

[134] A. Quinn, S. Xie, A. Apreysan, F. Fahim, S. Los, C. Peña, C. E. P. Lara, T. Zimmerman, and D. Braga, "A sub-5 ps jitter time-to-digital converter asic with back-gate delay tuning in 22 nm cmos", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 1080, p. 170 784, 2025, ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2025.

170784. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900225005856.

[135] C.-Y. Chen, H.-C. Shih, C.-W. Wu, C.-H. Lin, P.-F. Chiu, S.-S. Sheu, and F. T. Chen, "Rram defect modeling and failure analysis based on march test and a novel squeeze-search scheme", *IEEE Transactions on Computers*, vol. 64, 2015. DOI: 10.1109/TC.2014.12.

[136] H. G. Hezayyin, M. Mayahinia, and M. B. Tahoori, "Fault modeling and testing of reram-based cam array", in *Proc. IEEE VLSI Test Symposium (VTS)*, 2025.

[137] H. Aziza et al., "Built-in self-test structure (bist) for resistive rams characterization and variability evaluation", in *Proc. International Symposium on Device Research and Simulation (ISDRS)*, 2013.

[138] D. Sengupta and S. S. Sapatnekar, "Estimating circuit aging due to bti and hci using ring-oscillator-based sensors", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017. DOI: 10.1109/TCAD.2017.2648840.

[139] J. Keane, T.-H. Kim, and C. H. Kim, "On-chip reliability monitors for measuring circuit degradation", *Microelectronics Reliability*, 2010.

[140] L. R. Juracy et al., "A survey of aging monitors and reconfiguration techniques", *arXiv preprint*, 2020. eprint: 2007.07829.

[141] J. Díaz-Fortuny et al., "A ring-oscillator-based degradation monitor concept", 2022.

[142] H. Karrari, P. Andreani, and S. Tan, "A 12-bit high-speed time-interleaved pipelined asynchronous successive-approximation adc in 22-nm fdsoi cmos", *IEEE Access*, vol. 12, pp. 44 115–44 124, 2024. DOI: 10.1109/ACCESS.2024.3381037.

[143] K. Ni et al., "Ferroelectric fet-based in-memory computing", *Nature Electronics*, vol. 5, no. 12, pp. 782–793, 2022.

[144] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing", *Nature Nanotechnology*, vol. 15, no. 7, pp. 529–544, 2020.

[145] X. Chen et al., "Machine-learning-assisted variation-aware calibration for reliable memory systems", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 9, pp. 2871–2884, 2023.

[146] Y. Wang et al., "Adaptive sensing and calibration techniques for reliable in-memory computing", *IEEE Transactions on Circuits and Systems I*, vol. 69, no. 11, pp. 4602–4615, 2022.

[147] J. Kim et al., "Time-domain computing-in-memory using tdc-based sensing", *IEEE Journal of Solid-State Circuits*, vol. 58, no. 8, pp. 2145–2158, 2023.

[148] S. Lee et al., "Asynchronous time-based sensing for energy-efficient in-memory computing", *IEEE Transactions on Circuits and Systems I*, vol. 69, no. 6, pp. 2514–2527, 2022.

[149] M. Rahman et al., "Self-healing techniques for reliable memory architectures", *IEEE Design & Test*, vol. 38, no. 5, pp. 74–83, 2021.

[150] H. Zhang et al., “Runtime reliability monitoring and predictive failure analysis for emerging memory systems”, *IEEE Transactions on Device and Materials Reliability*, vol. 22, no. 4, pp. 561–572, 2022.

[151] H. Xu et al., “Design automation challenges and opportunities for computing-in-memory architectures”, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 5, pp. 1450–1463, 2023.

[152] M. Shafique et al., “Reliability-aware design of emerging computing systems”, *Proceedings of the IEEE*, vol. 109, no. 5, pp. 752–774, 2021.

List of Figures

List of Tables

Acronyms

CiM Computation-in-Memory

NVM Non-volatile Resistive Memories

NVM-CiM Computation in Non-volatile Resistive Memories

MAC Multiply-Accumulate

AI Artificial Intelligence

BEoL Back-End-of-the-Line

ReCAM ReRAM-based content-addressable memory

STT-MRAM Spin-Transfer Torque Magnetic RAM

ReRAM Redox-based RAM

PCM Phase Change Memory

FeFET Ferroelectric Field-Effect Transistor

SRAM Static Random Access Memory

DRAM Dynamic RAM

PV Process Variation

PVT Process, Voltage, and Temperature Variations

DNA Deoxyribonucleic Acid analysis

BF Bloom filters

DfT Design-for-Testability

BER bit-error-rate

VCRO Voltage-Controlled Ring Oscillator

LRS Low Resistive State

HRS High Resistive State

ML match-line

CAM Content Addressable Memory

NVM-CAM NVM-based CAM

SRAM-CAM SRAM-based CAM

OpAmp Operational Amplifier

ADC Analog-to-Digital Converter

DAC Digital-to-Analog Converter

LUT Look-Up Table

TDC Time-to-Digital Converter

BIST Built-In Self-Test

VCRO voltage-controlled ring-oscillator

D2D Device-to-device

C2C Cycle-to-cycle

FSM Finite State Machine

FF Fault-Free

HDC Hyper-Dimensional Computing

EDA Electronic Design Automation