Self-Healing Integrated Circuits/Systems in Semiconductor Nanometer Technologies

A Dissertation

submitted to the designated

by the General Assembly

of the Department of Computer Science and Engineering

Examination Committee

by

Eleni-Maria Dounavi

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

University of Ioannina

February 2020

Advisory Committee:

- **Yiorgos Tsiatouhas**, Professor at the Department of Computer Science and Engineering, University of Ioannina (Supervisor)
- **Chrysovalantis Kavousianos**, Professor at the Department of Computer Science and Engineering, University of Ioannina
- **Aristeidis Efthymiou**, Assistant Professor at the Department of Computer Science and Engineering, University of Ioannina

Examination Committee:

- **Yiorgos Tsiatouhas**, Professor at the Department of Computer Science and Engineering, University of Ioannina
- **Chrysovalantis Kavousianos**, Professor at the Department of Computer Science and Engineering, University of Ioannina
- **Aristeidis Efthymiou**, Assistant Professor at the Department of Computer Science and Engineering, University of Ioannina
- **Vasileios Tenentes**, Assistant Professor at the Department of Computer Science and Engineering, University of Ioannina
- **Dimitrios Nikolos**, Professor at the Department of Computer Engineering and Informatics, University of Patras
- **Alkiviadis Chatzopoulos**, Professor at the School of Electrical & Computer Engineering, Aristotle University of Thessaloniki
- **Spyridon Nicolaides**, Professor at the Department of Physics, Aristotle University of Thessaloniki

DEDICATION

To my husband,

for his endless love and support that always inspires me

to keep conquering my dreams.

ACKNOWLEDGEMENTS

First of all, I would like to thank my parents, for their love and support and for always seeing the best in me.

I would like to express my gratitude to my supervisor, Prof. Y. Tsiatouhas for his valuable contribution in every effort of mine during all the years of my education. I would like to thank him for his unreserved support and for believing in me under all circumstances. His guidance and kindness always illuminated my path.

I would like to specially thank the other members of the Advisory Committee, Prof. C. Kavousianos and A. Efthymiou, for their encouragement and collaboration we had all these years.

I would also like to thank Prof. D. Nikolos, A. Chatzopoulos, S. Nicolaides and V. Tenentes for serving as members of the Examination Committee.

I would also like to express my appreciation to all my colleagues and friends, for their continuous support and for all the cheerful hours we spent together giving meaning to my life.

Last but not least, I would like to thank from my heart my husband Simos, for lighting up my life and for being my shelter even during the toughest times. His patience, understanding and strong heart, have always encouraged me to keep moving forward. I could not have succeeded if it weren't for him standing by my side. After all, my life is full of... CMOS!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Eleni-Maria N. Dounavi,

Ph.D., Department of Computer Science and Engineering, University of Ioannina, Greece, January 2020.

Title of Dissertation: Self-Healing Integrated Circuits/Systems in Semiconductor Nanometer Technologies.

Thesis Supervisor: Yiorgos Tsiatouhas, Professor.

The evolution of CMOS technology over the years has allowed the presence of billions of transistors in an integrated circuit, with state-of-the-art technology platforms to be Systems-on-a-Chip (SoC), Networks-on-a-Chip (NoCs) and 3D-chips. The most significant challenge for the further development of technology is the problem of reliability both at the transistor level and at the circuits and systems level.

As the size of the transistors scales, important reliability issues have emerged, due to aging phenomena such as bias-temperature instability (BTI), hot-carrier injection (HCI), dielectric breakdown and electromigration. Also, as consumer needs change, systems are required to operate under ever-changing conditions, and therefore the reliability of integrated circuits is crucially influenced by extrinsic factors (such as radiation, noise, temperature variations, voltage supply variations etc.) under which they are called upon to maintain their proper functionality.

One promising approach to cope with reliability issues is the development of self-healing circuits and systems. That is, circuits and systems that will "sense" their aging status as well as changes that the environment will set and react appropriately to continue operating reliably under any conditions.

The present thesis focuses on the analysis of the mechanisms that influence the reliable operation of very large scale integration (VLSI) circuits and systems in order to determine their characteristics and then to develop innovative methods and techniques that can offer them self-healing so that they function reliably and uninterrupted throughout their useful life.

The target of this dissertation is to develop embedded techniques in the field of SRAM memories for early, on-line, excess aging prediction and oncoming failure diagnosis, aiming to maintain, by proper actions, the reliable operation and prolong the lifetime of the integrated circuit/system where the SRAM belongs. These techniques will allow the sensing of the SRAM status and thus, the prediction of upcoming failures on its different parts (memory cell, sense amplifier, decoders). Next, after over-aging prediction, our goal is to present SRAM operation adjusting techniques or repairing options for the self-healing of the corresponding sub-circuits in order to maintain the memory's reliable operation.

Two new self-healing methods are proposed for SRAM Memory Cells and Sense Amplifiers (SAs) and one new method for SRAM Decoders. Firstly, a scheme that addresses individually the SAs and the Memory Cells for aging monitoring of their transistors is presented and corresponding self-healing options that can be applied during their operation are suggested in order to maintain the reliability of the SRAM. The monitoring scheme is based on the use of a small Differential Ring Oscillator (DRO), which is placed at every bit-line in the memory array. The duty cycle of the DRO signal is used for the discrimination of aged memory cells or SAs. Next, a unified approach of the above scheme is presented for monitoring the

age of both memory cells and SAs. Simulations performed reveal that with a periodic application of the proposed technique, this scheme is capable of efficiently detecting overaged cells and SAs before failures appearance in the SRAM and with the exploitation of the proposed repairing mechanisms, the SRAM is self-healed and its reliable operation is retained throughout its lifetime. The proposed scheme can be also applied during the manufacturing testing operations.

An alternative approach of the previous technique is also proposed for the needs of the self-healing on SRAM SAs. The idea is once more based on the aging monitoring with the use of a low cost DRO considering however some alterations in the structure of the circuitry. For the detection of over-aged SAs, the frequency ratio of two signals is used and the DRO turns to be reconfigurable (rDRO) with the addition of a pair of switches. Self-healing is achieved with the detection of over-aged SAs and the further application of a repairing methodology on the ones detected. The efficiency of the scheme is again validated with simulations.

Finally, an aging monitoring technique for the SRAM Decoders, along with an adjusting technique to achieve self-healing are presented. The circuit proposed for monitoring the performance degradation of the Decoders' transistors aims to diagnose over-aged Decoders early and properly react in order to prolong the lifetime of the SRAM. It suggests the addition of a simple, low cost embedded circuit consisting of two extra word-line slices along with a Counter, a MUX, a simple Test Controller and a Comparator. The response signal of the Comparator, is used for the discrimination of over-aged Decoder slices. In the direction of the SRAM's self-healing, upon detection of an over-aged Decoder, a repairing methodology is applied and the SRAM's reliability is ensured. The simulation results performed on the proposed scheme validate its ability for early (before the presence of failures) self-healing of over-aged Decoders, while offering low cost in silicon area as well as the ability to avoid aging of the Decoder when the SRAM is operating in normal mode.

Εκτεταμενη Περιληψη

Ελένη-Μαρία Νικολάου Δούναβη,
Δ.Δ., Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
Ιανουάριος 2020
Τίτλος Διατριβής: Αυτό-ιασώμενα Ολοκληρωμένα Κυκλώματα/Συστήματα σε Νανομετρικές Τεχνολογίες Ημιαγωγών.
Επιβλέπων: Γεώργιος Τσιατούχας, Καθηγητής.

Η εξέλιξη της τεχνολογίας CMOS στο πέρασμα των χρόνων, έχει επιτρέψει την παρουσία δισεκατομμυρίων τρανζίστορ σε ένα ολοκληρωμένο κύκλωμα, με τεχνολογικές πλατφόρμες αιχμής τα συστήματα σε ένα ολοκληρωμένο κύκλωμα (Systems-on-a-Chip – SoC), τα δίκτυα σε ένα ολοκληρωμένο κύκλωμα (networks-on-a-Chip – NoCs) και τα τρισδιάστατα ολοκληρωμένα κυκλώματα (3D-chips). Η σημαντικότερη πρόκληση για την περαιτέρω εξέλιξη της τεχνολογίας είναι το πρόβλημα της αξιοπιστίας τόσο στο επίπεδο των τρανζίστορ, όσο και σε εκείνο των κυκλωμάτων και των συστημάτων.

Με την κλιμάκωση του μεγέθους των τρανζίστορ, έχουν έρθει στην επιφάνεια σημαντικά θέματα αξιοπιστίας εξ αιτίας φαινομένων γήρανσης όπως η αστάθεια πόλωσης-θερμοκρασίας (bias-temperature instability - BTI), η έγχυση θερμών φορέων (hot-carrier injection - HCI)], η διάτρηση του διηλεκτρικού (dielectric breakdown) και η ηλεκτρομετανάστευση (electromigration). Επίσης, καθώς οι ανάγκες των καταναλωτών μεταβάλλονται, τα συστήματα καλούνται να

λειτουργήσουν υπό συνεχώς μεταβαλλόμενες συνθήκες και κατ' επέκταση η αξιοπιστία των ολοκληρωμένων κυκλωμάτων επηρεάζεται καθοριστικά από εξωγενείς παράγοντες (όπως οι ακτινοβολίες, ο θόρυβος, οι διακυμάνσεις της θερμοκρασίας, οι διαταραχές της τροφοδοσίας κ.α.) κάτω από τους οποίους αυτά καλούνται να διατηρήσουν την ορθή λειτουργία τους.

Μια πολλά υποσχόμενη προσέγγιση για την αντιμετώπιση προβλημάτων αξιοπιστίας είναι η ανάπτυξη αυτό-ιασόμενων κυκλωμάτων και συστημάτων. Δηλαδή κυκλωμάτων και συστημάτων που θα "αισθάνονται" την κατάσταση λειτουργίας τους και τις περιβαλλοντικές μεταβολές και θα αντιδρούν κατάλληλα ώστε να συνεχίσουν να λειτουργούν απρόσκοπτα κάτω από οποιεσδήποτε συνθήκες.

Η παρούσα διατριβή εστιάζει στην ανάλυση των μηχανισμών που επηρεάζουν την αξιόπιστη λειτουργία κυκλωμάτων και συστημάτων πολύ υψηλής κλίμακας ολοκλήρωσης (very large scale integration – VLSI) ώστε να καθοριστούν τα ιδιαίτερα χαρακτηριστικά τους και εν συνεχεία στην ανάπτυξη καινοτόμων μεθόδων και τεχνικών που μπορούν προσφέρουν σε αυτά αυτο-ίαση ώστε να λειτουργούν αξιόπιστα και αδιάλειπτα καθ' όλη τη διάρκεια της ωφέλιμης ζωής τους.

Σε αυτή την κατεύθυνση, ο αρχικός στόχος της έρευνας είναι η ανάπτυξη ενσωματωμένων τεχνικών στον τομέα των Στατικών Μνημών Τυχαίας Προσπέλασης (SRAMs) για έγκαιρη και εν λειτουργία πρόγνωση και εντοπισμό (διάγνωση) φαινομένων έντονης γήρανσης, με στόχο την παράταση της διάρκειας ζωής του ολοκληρωμένου κυκλώματος/συστήματος όπου ανήκει η SRAM. Ο επιπρόσθετος στόχος είναι να παρουσιαστούν μεθοδολογίες αυτο-ίασης των κυκλωμάτων σε περίπτωση υπερβολικής υποβάθμισης των επιδόσεων των κυκλωματικών στοιχείων της SRAM προκειμένου να διατηρηθεί η αξιόπιστη λειτουργία της μνήμης.

Δύο νέες μέθοδοι αυτο-ίασης προτείνονται για τα κελιά και τους αισθητήρες σήματος μνημών SRAM και μία νέα μέθοδος για τους αποκωδικοποιητές μνημών SRAM. Αρχικά, παρουσιάζεται ένα σχήμα που απευθύνεται ξεχωριστά στους αισθητήρες σήματος και στα κελιά μιας μνήμης SRAM για την ανίχνευση παρουσίας γήρανσης στα τρανζίστορ τους και προτείνονται αντίστοιχες μεθοδολογίες αυτο-ίασης που μπορούν να εφαρμοστούν κατά τη διάρκεια λειτουργίας τους προκειμένου να διατηρηθεί η αξιοπιστία της SRAM. Το κύκλωμα ελέγχου βασίζεται στη χρήση ενός μικρού διαφορικού ταλαντωτή (DRO) που τοποθετείται σε κάθε γραμμή ψηφίου στη συστοιχία μνήμης. Ο κύκλος λειτουργίας (duty cycle) του σήματος του DRO χρησιμοποιείται για τη διάκριση γερασμένων κελιών μνήμης ή αισθητήρων σήματος. Στη συνέχεια, παρουσιάζεται μια ενοποιημένη προσέγγιση του παραπάνω κυκλώματος για την ανίχνευση της γήρανσης τόσο στα κελιά μνήμης όσο και στους αισθητήρες σήματος από κοινού. Οι προσομοιώσεις που πραγματοποιήθηκαν επιβεβαιώνουν ότι με την περιοδική εφαρμογή της προτεινόμενης τεχνικής, δίδεται η δυνατότητα αποτελεσματικής ανίχνευσης γερασμένων κελιών και αισθητήρων σήματος πριν από την εμφάνιση αστοχιών στην SRAM και με την αξιοποίηση των προτεινόμενων μηχανισμών διόρθωσης, η SRAM αυτό-ιάσεται και διατηρείται η αξιόπιστη λειτουργία της καθ' όλη τη διάρκεια ζωής της. Το προτεινόμενο κύκλωμα μπορεί επίσης να επαναχρησιμοποιηθεί κατά τη διάρκεια των κατασκευαστικών δοκιμών (manufacturing testing).

Προτείνεται επίσης μια εναλλακτική προσέγγιση της προηγούμενης τεχνικής για τις ανάγκες της αυτό-ίασης στους αισθητήρες σήματος μιας SRAM. Η ιδέα βασίζεται και πάλι στην ανίχνευση της γήρανσης με τη χρήση ενός χαμηλού κόστους DRO λαμβάνοντας υπόψη ωστόσο ορισμένες αλλαγές στη δομή του κυκλώματος. Για την ανίχνευση γερασμένων αισθητήρων σήματος, χρησιμοποιείται ο λόγος συχνοτήτων δύο σημάτων και ο DRO μετατρέπεται σε επαναρυθμισμένο κύκλωμα (reconfigurable DRO – rDRO) με την προσθήκη ενός

ζεύγους διακοπτών. Η αυτό-ίαση επιτυγχάνεται με την ανίχνευση γερασμένων αισθητήρων σήματος και την περαιτέρω εφαρμογή μιας μεθοδολογίας διόρθωσης σε αυτούς. Η αποτελεσματικότητα του κυκλώματος επιβεβαιώνεται και πάλι με τη χρήση προσομοιώσεων.

Τέλος, παρουσιάζεται μια τεχνική ανίχνευσης γήρανσης για τους αποκωδικοποιητές μιας SRAM, μαζί με την προτεινόμενη διαδικασία αυτο-ίασης για την επιτυχή αυτο-ίαση. Το κύκλωμα που προτείνεται για την ανίχνευση της υποβάθμισης των επιδόσεων των τρανζίστορ των αποκωδικοποιητών στοχεύει στην έγκαιρη διάγνωση γερασμένων αποκωδικοποιητών και στην κατάλληλη αντίδραση προκειμένου να παραταθεί η διάρκεια ζωής της SRAM. Προτείνει την προσθήκη ενός απλού, χαμηλού κόστους ενσωματωμένου κυκλώματος αποτελούμενου από δύο πρόσθετες γραμμές λέξης μαζί με έναν μετρητή, έναν πολυπλέκτη, έναν απλό ελεγκτή και έναν συγκριτή. Το σήμα απόκρισης του συγκριτή χρησιμοποιείται για τον προσδιορισμό των γερασμένων γραμμών του αποκωδικοποιητή. Στην κατεύθυνση της αυτο-ίασης της SRAM, ύστερα από την ανίχνευση ενός γερασμένου αποκωδικοποιητή, εφαρμόζεται μια διαδικασία αυτό-ίασης ώστε να εξασφαλιστεί η αξιόπιστη λειτουργία της SRAM. Τα αποτελέσματα των προσομοιώσεων που εκτελέστηκαν στο προτεινόμενο κύκλωμα επικυρώνουν την ικανότητά του για έγκαιρη αυτο-ίαση γερασμένων αποκωδικοποιητών (πριν της παρουσίας αστοχιών), προσφέροντας χαμηλό κόστος στην επιφάνεια του πυριτίου καθώς και την ικανότητα της αποφυγής γήρανσης του κυκλώματος ανίχνευσης γήρανσης όταν η SRAM λειτουργεί σε κανονική λειτουργία.

CHAPTER 1

INTRODUCTION

1.1. Prologue

For over four decades technology evolves rapidly with utility devices becoming smaller in size, thereby satisfying consumers' need for space economy. In the production of advanced generation electronic devices, the tendency to integrate more processors, memory and other elements onto a single chip is typically described by the term System on Chip (SoC) [1].

Such a rapid development in microelectronics technology is due in large part to the ability of integrating billions of transistors onto a small surface of semiconductor material. The use of large and complex chips requires an enormous number of vectors to test their proper operation, with the expected increase in testing time and energy consumption during integrated circuit testing, which can lead to a significant increase in cost. The reduction in transistor size has already dropped the scale from micrometers to nanometers, and the reduction seems to continue day by day. This conclusion is based on a prediction made by Gordon

Moore in the 1960s [2], based on which the number of transistors that can be integrated on a single silicon chip will double every two years. This prediction, later called Moore's Law, has been verified with astonishing accuracy.

This dramatic shrinkage has created a significant impact on both the circuits design and their testing for faults and consequently their diagnosis. Most of the designs found on a chip consist of many hundreds of millions to billions of transistors, operating at frequencies in the order of gigahertz. These designs may include digital, analog, optical circuits, microelectromechanical systems (MEMS), as well as circuits operating at radio frequencies.

The increased design complexity, as well as the high operating speeds of modern integrated circuits have dramatically contributed to the occurrence with increasing frequency and probability of errors during the operation of integrated circuits. As market needs shift to increasingly sophisticated electronic systems that work not only properly, but also at maximum speed and with minimal power consumption, researchers' attention is urged to devise methods to minimize the probability of errors occurrence not only during manufacturing but equally important, during the operation of such systems. Many techniques based on both software and hardware have been proposed to address these critical issues.

But what about the physical aging of integrated circuits that inevitably leads to errors and also reduces operating speed? How can a problem that follows the laws of physics be addressed?

Every manufactured integrated circuit has certain specifications for both its functions and its life span. The proper operation of an integrated circuit throughout its lifetime is an important factor that is seriously threatened by errors occurrence, the presence of which can have a huge financial impact. For example, on August 6, 1999, eBay's servers crashed and this resulted in more than eight

hours of downtime. The cause was attributed to a hardware error. Other high profile sites like Amazon and PayPal experienced similar interruptions due to hardware errors. Such errors are in many cases due to the aging of integrated circuits, which makes them insufficient to handle large amounts of data at the high speeds required by today's market. Aging effects, frequently occurring when an integrated circuit is exposed to excess voltage or temperature stress, lead to a degradation in the performance and reliability of an electronic system, hence limiting its expected lifetime. Subsequently, the problem of systems reliability and the need to find innovative solutions that may effectively handle similar issues arises more than ever.

Designing robust systems to ensure the required hardware reliability, although not trivial, is nowadays feasible but at a high cost. The ability to detect errors during the operation of a system is an extremely important aspect of these systems. Some errors are easy to detect such as, for example, accessing illegal memory locations or dividing by 0. Troubleshooting would be a simple process if all errors occurred in such ways. More subtle errors, however, can occur - errors that produce incorrect results or incorrect control flow in a program for example - that may not be detectable by such simple tests. Such errors, if not detected, can cause serious consequences, ranging from data loss to economic and productive losses or even loss of life. More specifically, when addressing aging effects, even a careful testing cannot reduce the risk of error occurrence that may potentially lead to an early system failure. Aging aware design thus needs a reliable prediction of the effects of aging in order to optimize the system's performance throughout its lifetime.

Several design techniques have been suggested in the open literature, but they are generally expensive. The most important challenge is to achieve acceptable levels of robustness with minimal cost at system level (power, performance, surface area and design complexity). In addition, many error detection techniques often

consider a single root of cause and further assume that failures are independent. Such assumptions, however, cease to be sufficient as demands increase.

Such a significant problem is added to the fact that as technology advances and evolves in all areas of life, the same electronic devices for which we had once focused on solving these issues now come to bring to the surface another factor leading to insufficient operation. Most electronic devices are no longer intended to operate under constant and pre-known environmental conditions. On the contrary, every consumer expects the electronic device he buys to operate to its full potential and under constantly changing situations. Thus, one device may be required to operate at one time under very high temperatures, the other in the presence of high humidity, or even to operate in a high-radiation environment capable of causing malfunctions in an electronic device. Such changes however, may well degrade the systems' performance and reliability and may lead to the presence of aging phenomena.

Exposure of integrated circuits to highly variable or at times extreme environmental conditions renders them vulnerable and often leads to failures in their operation [3]. The industry often focuses on producing devices capable of meeting specific conditions while still guaranteeing a limited life span, which constantly seems to decrease without the presence of faults, whether temporal or permanent.

So as we see the time during which a circuit is guaranteed to function properly be so limited, the ability to quickly detect and immediately manage errors of such complexity is a major challenge if not a serious issue requiring research. External environmental conditions cannot be known in advance and aging caused by either serious environmental changes or by natural wearout cannot be avoided. It is therefore extremely important that a designer can introduce solutions to detect errors randomly occurring in the integrated circuit and then provide feasible

solutions to tolerate them in order to achieve correct operation of the integrated circuits and ensure their integrity.

The present thesis in order to address all the problems presented above, proposes the design of embedded techniques on integrated circuits capable of self-sensing the potentially catastrophic changes in the operation of their internal structures when exposed to serious voltage or temperature stress that are the main aging factors and properly react to allow either their correction or their mitigation (e.g. their self-healing). In this direction, we aim to maintain the reliable operation of the integrated circuits that are seriously affected by aging and thus contribute to maintaining the reliability of the overall system, by monitoring the system's status and when necessary, to adjust the system operation or repair it by applying self-repair mechanisms e.g. by replacing the faulty components with other spare ones if these are available. [4]-[7]. Interestingly, both the sensing procedure and the self-healing actions on the integrated circuit are suggested to be performed periodically in the field of operation. It is considered particularly important an electronic device to adapt its operation according to the aging condition of its devices in order to maintain a reliable and efficient functionality under any circumstances as well as to extend its life span.

1.2. Dissertation Scopes

The target of this dissertation is to develop embedded techniques in the field of SRAM memories for early, on-line, excess aging prediction and oncoming failure diagnosis, aiming to maintain, by proper actions, the reliable operation and prolong the lifetime of the integrated circuit/system where the SRAM belongs. These techniques will allow the sensing of the SRAM status and thus, the prediction of upcoming failures on its different parts (memory cell, sense amplifier, decoders).

Next, after over-aging prediction, our goal is to present SRAM operation adjusting techniques or repairing options for the self-healing of the corresponding sub-circuits in order to maintain the memory's reliable operation.

1.3. Dissertation Structure

The Dissertation consists of 6 Chapters. Chapter 1 is a brief introduction, while Chapter 2 outlines the basic preliminaries on design for testability, reliability degradation and aging mechanisms of integrated circuits, as well as the self-healing concept and the general categories of aging monitoring techniques. Chapter 3 presents fundamental issues on SRAM memory design and operation, the device aging mechanisms under consideration and their influence on the SRAM's reliability and also discusses the state of the art on SRAM aging monitoring techniques. Chapters 4 and 5 present the new techniques proposed for aging monitoring and repairing on the different sub-circuits of an SRAM. Finally, in Chapter 6 the conclusions are drawn.

CHAPTER 2

PRELIMINARIES

This chapter introduces the relevant background of the following Chapters 3, 4 and 5 respectively.

2.1. Testing and Design for Testability

The testing of a circuit includes all those actions that are taken to ensure that the circuit delivered to the market is operating properly under all conditions [8]. This is a process that is not as easy as it may seem at first glance. At the design stage of a circuit, each designer has unlimited access to all nodes of the network and is able to fully analyze its operation. He is also given the freedom to apply a variety of input combinations and to observe the response that comes to any node selected. Such a testing for the correct operation of the integrated circuit cannot be

performed after the circuit is fabricated. In this case, the only access to the circuit is through imposing values on the input probes and observing their responses at the output probes. Not to mention that the total number of these probes is significantly limited especially with respect to the number of transistors in a design.

A complex unit, such as a microprocessor, consists of tens or hundreds of millions to billions of transistors and contains countless possible states. It is therefore, a very long process - if not impossible - to set such a unit in a particular state and observe the resulting circuit response, through the limited bandwidth provided by the I/O probes. In addition, hardware test equipment tends to be very expensive and every second spent by the tester during a test operation of a circuit increases the depreciation cost.

It is therefore important to bear in mind the issue of testing the circuits early in the design process to ensure the integrity of the integrated circuit and to eliminate the possibility of its being given to the market in the event of malfunction. Some minor modifications to the circuit can help make the absence of errors easier to validate. This design approach is called design for testability (DFT). Although often overlooked by designers who prefer to focus on more interesting design issues such as design optimization, this approach is an integral and important part of the design process and should be taken into account as early as possible in the design phases.

DFT techniques are intended to incorporate assistive mechanisms to test the proper functioning of integrated circuits in order to ensure detection of possible faults in them [9].

A DFT strategy comprises two parts [10]:

- Implementation of the necessary circuits so that the test process can be fast and extensive.
- Provision of the necessary test patterns or excitation vectors to be used during the testing process. For cost reasons, it is desirable that the sequence of testing combinations is as small as possible and covers all possible faults.

Testing the correct operation of an integrated circuit can be done either by applying off-chip testing vectors or by using integrated on chip testing techniques. Testing operation can be performed while the circuit is in use (on line testing) either concurrently or periodically or when the circuit is not in use (off line testing) [11].
The initial correct design of an integrated circuit does not guarantee that the constructed circuit will be operational. There may be manufacturing defects or damage (for example, deposition of material on the silicon crystal) or deviations in the manufacturing process leading to significant deviations from the expected function. Defects can also occur during stress tests performed after manufacture. These tests expose a circuit to high temperatures, mechanical pressure, etc., to ensure its operation over a wide range of operating conditions. Typically such defects are the short circuits between the wires and the ruptured interconnects (open circuits). This is translated to nodes that are short-circuited or nodes that are "floating". The impact of the defects is becoming more and more important as technology evolves and the transition to nanotechnology era.

In Figure 2.1 below, mechanisms for defect generation are presented. Such defects are rapidly increasing as the technology scales down.

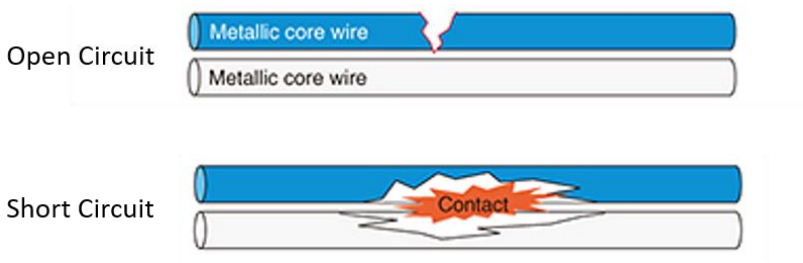

Figure 2.1 Defect Example.

It is important at this point to make clear that it is different when talking about a defect, a fault or an error in a circuit. Each of these terms will be explained by giving the relevant definitions.

Defects in an integrated circuit are the manufacturing imperfections and permanent damage that occur during its manufacturing process (e.g. short circuit, cut line, etc.).

Faults are the modeling of the defects effect on the behavior of integrated circuits (e.g. a line with a constant value of 1 or 0, etc.).

Errors are the incorrect logic responses of integrated circuits in the presence of faults.

Integrated circuits are therefore often subject to failures, the main causes of which are: specification errors, manufacturing errors, external disturbances, faulty devices, poor circuit management, as well as natural wearout.

Specification errors include incorrect algorithms, architectural or design specifications in both hardware and software. For example, if the time constraints of all the individual circuits in a design are not set correctly, the final circuit will not function properly.

Errors in the implementation of circuits are also a common cause of failures. For example, if technology design rules such as minimum metal distance are not met, short circuits may be created. This category also includes design errors that are not highlighted during the verification process.

When referring to external disturbances, we refer, for example, to the extreme conditions under which integrated circuits are often called upon to operate. For example, if a circuit is subjected to extreme temperature fluctuations its responses may be incorrect. Furthermore, failures can be caused by the effects of radiation or electromagnetic interference (EMI). In addition, electronic systems are usually susceptible to electrostatic sources such as lightning or other environmental related causes.

Improper management of the systems by the users can also cause serious problems in the operation of the circuits and this may cause failures and eventually reduced or incorrect operation of the circuits.

The most common cause of failures is the existence of defects. As mentioned before, such may be possible imperfections in the process of manufacturing integrated circuits and accidental manufacturing effects such as extrusion of exogenous particles, perforation of insulating material and cut or short-circuited metals. Manufacturing defects are not limited to die only but include defects in the integrated circuit packaging such as disconnection of pins or corrosion of metals [12], [13].

The deterioration of individual components of a system due to aging is also an important cause of failures. The operation of circuits under extreme stress conditions (for example temperature) results in the physical wear of the circuitry causing the integrated circuit responses to deteriorate over time.

2.2. Yield and Fault Coverage

An important metric that needs to be analyzed is the yield (Y) which is used to grade the quality of an ICs manufacturing process. The quality of a manufacturing process is important because it is linked to the profit margin. The yield depends on the technology, the area of the integrated circuit and its physical layout. A percentage of manufactured integrated circuits are expected to be defective when given to the market due to manufacturing defects. Images taken with electronic microscope of manufacturing defects are presented in Figure 2.2. Figures 2.2a and 2.2b present two sites where particles trapped in the silicon during the manufacturing process caused non-conductive and conductive defects respectively. Figure 2.2c presents a thin interconnection, probably caused by process variations, which may cause the poor performance of the interconnection because of its higher resistivity.

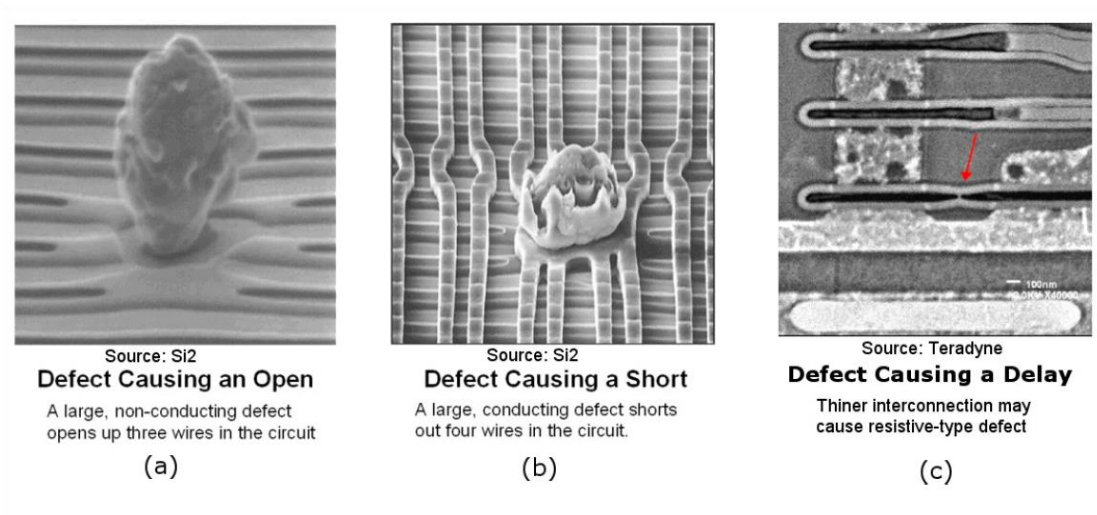

Figure 2.2 Silicon Defects.

The yield of the manufacturing process is defined as the ratio of the number of accepted wafer circuits to the total number of integrated circuits on the wafer. It is therefore important to ensure a high value for manufacturing yield so that the

sum of properly constructed integrated circuits reaches the sum of all integrated circuits.

Yield loss is divided into two types: catastrophic and parametric. The catastrophic yield loss is due to accidental defects during manufacturing, while the parametric is due to process variations. **Process variations** are deviations of parameters from their desired values due to imperfect nature of fabrication process. Such parameters include transistor channel length variation, transistor threshold voltage variation, metal interconnect thickness variation or intermetal layer dielectric thickness variation and have a big impact on circuits' performance. In general, the effect of process variations shows up first in the most critical paths in the design, those with near maximum delays. Automation and improvements in the manufacturing process of integrated circuits drastically reduce the density of particles that generate random manufacturing defects. As a result, the parametric changes that result from the deviations in the manufacturing process, as technology scales down, are a crucial cause of yield loss.

The methods used to reduce the effect of deviations during the manufacturing process and thereby increase efficiency are referred to as design for yield (DFY). Circuit fabrication methods to avoid accidental manufacturing imperfections are referred to as design for manufacturability (DFM). In general, any DFM method helps to increase the manufacturing efficiency and can therefore be considered as a DFY method. Manufacturing efficiency is related to the percentage of defects created. The following Figure 2.3 presents the failure diagram of a typical system showing how accidental failures and failures at the beginning and end of the circuit life contribute to the overall failure rate of the system.

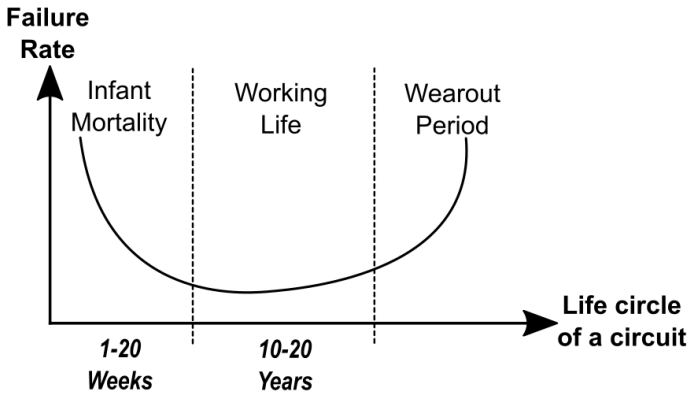

Figure 2.3 Failure Rate Diagram of an Integrated Circuit Over Time.

As shown in Figure 2.3, the rate of failures in relation to the life of an integrated circuit outlines a three-phase curve [1]. At the initial stage of manufacturing the failure rate is high (infant mortality period), as it relates to the period where the defects are in the process of being detected at the factory and the possible poor design of the circuit presents faults that need correction. The next period refers to the working life of the circuits and defines the beneficial life of the product. At this stage the failure rate is constant as the failures are the result of random events. The wearout period indicates the final stage of product life, where the failures rate increases. The failures in this period are due to the aging of the circuits and their continuous use. Some commercial electronic products usually do not enter this period as they are replaced by their newer versions due to technological developments. Nevertheless, in systems where reliability is an important design specification, there should be mechanisms to address the increased rate of failures.

Despite performing the proper manufacturing tests, however, some circuits that pass the test may contain defects or failures that have not been detected, while other circuits may fail the test process without actually being defective. As a result, even if all products pass acceptance test, some faulty devices can still be found in the field. When these faulty devices are returned to the IC manufacturer, they

undergo Failure Mode Analysis (FMA) for possible improvements to the ICs designing and manufacturing processes. The ratio of the number of integrated circuits that although defective, they succeed to pass the test procedure to the total number of integrated circuits is defined as the reject rate or is referred to as the defect level (DL).

The reject rate provides an indication of the overall quality of the manufacturing testing process [14]. Generally speaking, a reject rate of 500 parts per million (PPM) chips may be considered to be acceptable, while 100PPM or lower represents high quality. The goal of six sigma manufacturing, also referred to as zero defects, is 3.4PPM or less.

The ultimate target of any ICs test mechanism is to test the chips for all possible defects, or in other words, to achieve complete defect coverage. However, such a goal is not realistic, and thus fault models are adopted. Fault models save time and improve test efficiency, as a limited number of test patterns that target specific faults, related to structural testing. Any input pattern (test stimuli), that produces a different output response in a faulty circuit from that of the fault-free circuit is a test vector that is capable of detecting faults. Any set of test vectors is called a test set. The goal of Automatic Test Patterns Generation (ATPG) tools is to find an efficient test set that detect as many defects as possible for a given CUT and a given fault model. These tools provide a quantitative measure of the fault-detection capabilities of a given test set for a targeted fault model. This measure is called fault coverage and is defined as:

$$\text{fault coverage} = \frac{\#\text{number of detected faults}}{\#\text{ total number of faults}}$$

Fault coverage is linked to the quality of a manufacturing process, which is expressed by the yield, and the quality of the testing process, which is expressed by the reject rate, by the following relation [15]:

reject rate = 1 - yield (1-fault coverage).

2.3. Permanent and Intermittent Faults

A fault in an integrated circuit is defined as modeling the effect of defects during its operation. Faults are divided into two major general categories [14], [16]. The so-called permanent faults belong to one category. These remain active and permanently affect the circuit from the moment they appear until the cause of the fault is corrected (if the latter is possible). Faults in this category are usually due to defects with the main ones being the following:

1. **Stuck-at fault:** It is a logical (logic-level) fault model. A stuck-at fault transforms the correct value on the faulty signal line to appear to be stuck at a constant logic value. When this value is logical zero, then the fault is called stack-at-0, while when the value is logical one, the fault is called stack-at-1.
2. **Transistor stuck-on fault:** A transistor is constantly in a conductive state.
3. **Transistor stuck-open fault:** A transistor is constantly in a non-conductive state.
4. **Bridging fault:** A short circuit has been created between two elements. These elements can be transistor terminals or connections between transistors and gates.
5. **Delay fault:** The delay observed when transmitting a signal to a circuit in one or more paths.

The other category is intermittent faults. They are caused by internal damage of the circuit and lead to incorrect responses when the circuit is in a certain state.

The corresponding malfunction, although appearing periodically over a period of time, contributes to the progressive degradation of the circuit which can lead to permanent damage [17]-[19]. Two categories of intermittent faults are distinguished:

1. **Transient faults:** They occur in the form of an electrical pulse on a circuit node, usually of short duration, due to an accidental change of an environmental factor or noise, or a fluctuation in power supply, or cosmic radiation and bring about momentum effects on the operation of a circuit. Transient faults related to timing issues are the cause of timing errors and are due to various known mechanisms, such as crosstalk interference, noise or ground bounce.
2. **Temporary faults:** They occur irregularly over certain periods of time and are due to defective or aged elements and extrinsic factors such as temperature, humidity or vibration. In the case of aged data, as a rule, we have the transition of intermittent faults to permanent over time. The main way in which temporary faults occur is in the form of timing errors during a circuit's operation.

2.4. Reliability Degradation and Aging Mechanisms of Integrated Circuits

The ability of a system to maintain its proper functionality, even under unexpected conditions, during its lifetime is called reliability. For semiconductors, reliability is significantly influenced by the aging of a circuit's components and the consequences it has on the operation of the device. As technology continually scales to the nanometer sizes, aging mechanisms for integrated circuits play a significantly increasing role in the gradual degradation of their reliability constantly earlier in their useful life. The most important mechanisms that may lead to serious reliability degradation are described below.

2.4.1. Voltage and Temperature Variations

Voltage and temperature variations have been proven to be factors with serious negative impacts on the performance of a manufactured chip when they present excess variation from their reference value. They are considered to be dynamic variations since they change as a function of time and environment, and therefore cannot be compensated by static silicon tuning.

Voltage variations are considered as there are unexpected voltage drops in the power supply networks and also variations in the supply voltage generator itself. Voltage drops result from abrupt changes in the switching activity, inducing large current transients in the power delivery system, and contain high-frequency and low-frequency components which occur locally as well as globally across a die [20]. On the other hand, temperature is a factor which also affects the performance of ICs. Temperature variations are due to temperature fluctuations of the integrated circuits as well as the altering environmental conditions under which the circuits are requested to operate. When a circuit is forced to operate under such variations, propagation delays are most likely to occur leading to timing errors and faulty output results. As it is obvious, the performance degradation of a circuit is seriously threatened by the presence of these variations. Designers commonly use conservative guard-bands into the operating frequency and voltage to handle these variations and ensure error-free operation within the presence of worse case dynamic variations over a circuit's lifetime [21] that leads to loss of operational efficiency.

As the effect of voltage and temperature variations constantly increases due to the aggressive fabrication scaling of technology, accurate analysis coupled with new efficient design techniques, capable of sensing the reliability degradation caused and effective self-healing methods, are required to overcome the variability challenge.

2.4.2. Dielectric Breakdown

Gate oxide degradation and breakdown are critical problems for main VLSI technologies. Dielectric breakdown is the failure of a dielectric to withstand the electrical stress (electric field) that is applied on it. When electric fields induced on the oxide and silicon bulk become high enough, a large number of carriers pass through it. As the applied voltage increases, the currents due to these carriers increase also rapidly causing some portion of the dielectric to switch from being an electrical insulator to a partial conductor. This phenomenon can occur abruptly across the surface or through the dielectric, or it can occur as a series of small discharges that, over time, progressively damage the dielectric to the point where it eventually fails catastrophically. Electrical breakdown may be a momentary event (as in an electrostatic discharge), or may lead to a continuous arc if the system fails to interrupt the current in a power circuit [22].

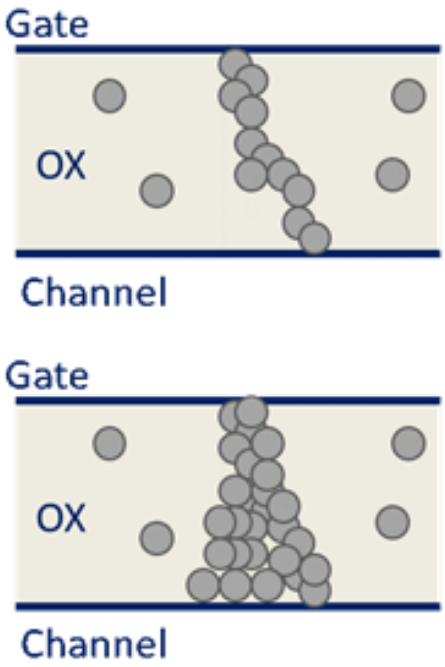

Figure 2.4 Dielectric Gate Breakdown.

Two types of dielectric breakdown are commonly observed in silicon layouts:

- the **abrupt** at which the value of the current is increased by several orders of magnitude for a certain value of the applied field or at a certain time duration and
- the **gradual** at which the current gradually increases until it exceeds a certain value, which is considered as the limit for the non-conducting properties of the oxide.

The physical mechanisms behind the dielectric breakdown are quite complex and have been the subject of extensive study to date. The most important of these mechanisms are: polarization of the dielectric, injection of carriers, trapping of holes and electrons in the oxide and interaction of all of them.

2.4.3. Electromigration

The phenomenon of electromigration relates to interconnections within an integrated circuit and occurs at the atoms level. It is the transfer of material caused by the gradual movement of ions in a conductor, due to the collisions between the moving (conductive) electrons (in the presence of high currents) on the one hand and the atoms of the metal on the other, so the latter move from their positions [23].

When a high current is applied to the surface of a metal conductor, the electrons collide with the conductor atoms. Constant electron collisions cause conductor atoms to move in the direction of flow, thereby increasing the resistance of the conductor. Thus, after some time accumulation of metal in one part of the line is observed while another part weakens and may at some point become discontinuous thereafter. When weakened enough or open-circuited, a high resistance occurs that limits the current or in the extreme case the current is eliminated. An image taken with electronic microscope is presented in Figure 2.5.

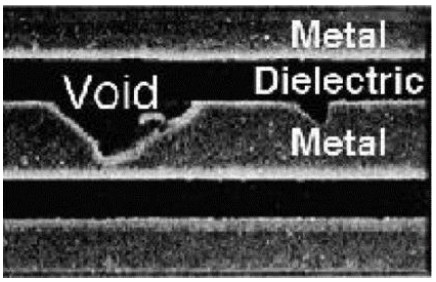

Figure 2.5 Electromigration Phenomenon.

This phenomenon is important in applications where high density continuous current is used, such as in microelectronics. As the size of conductors in ICs decreases, the practical importance of this effect is constantly increasing. In today's systems where processors consist of such metal - mainly copper - conductors, it is almost inevitable that this effect will occur [24].

The phenomenon was discovered by the French scientist Gerardin more than 100 years ago and became increasingly widespread with the first appearance of integrated circuits in the late 1960s. At that time, metal interconnections in integrated circuits were still about 10 microns. Today, interconnections are only a few hundred to ten nanometers wide, making research in the field of electromigration more and more important.

The significance of this phenomenon is that it substantially reduces the reliability of integrated circuits. It can cause the loss of connections and the total failure of a circuit.

The relation between electromigration and the corresponding mean time to failure (MTTF) [25] is presented next:

$$MTTF \sim A \cdot j^{-2} \cdot e^{Ea/kT}$$

where J is the current density in the cable, A is an empirically determined scaling factor and Ea is the activation energy of the electromigration mechanism which depends on the metal material. According to this relation, an increase of the wire's width or a lower current density contributes to increasing the mean time to failure [26].

Electromigration occurs during the operation of the circuit in the field of application and ultimately results in increased delay in the propagation of signals in the circuit. At the early stages of the phenomenon, various techniques dealing with timing errors detection appear to be effective in order to detect possible violations of the signals' timing. This allows the detection of any propagation delays and the application of appropriate repairing mechanisms.

2.4.4. Bias Temperature Instability - BTI

Another mechanism that contributes to integrity reduction of integrated circuits is the so-called Bias Temperature Instability mechanism. It is a complex electrothermal phenomenon that affects nanometer transistors and occurs at high gate to source voltage stress conditions and high temperatures causing over time threshold voltage changes. Its effect relates to both pMOS and nMOS transistors. As the gate-source polarization of the pMOS transistors is negative the mechanism is referred to as negative bias-temperature instability (NBTI) whereas in the case of nMOS transistors the polarization is positive it is referred to as positive bias temperature instability (PBTI). In previous CMOS technology generations the impact of the NBTI phenomenon was a much more fundamental issue of reliability compared to the PBTI phenomenon which was neglected due to its small effect on NMOS transistors. However, in today high-κ metal-gate nanometer technologies PBTI seems to be a comparable to NBTI reliability reduction mechanism [27], [28].

The BTI mechanism manifests itself as an increase to the absolute value of the transistors' threshold voltage and a subsequent decrease in their drain current. The change in threshold voltage in transistors due to the bias-temperature stress applied to them, leads them to a reduced conductivity [29], [30]. Gradually, this degradation becomes logarithmically dependent on time, with the result that over time transistors logarithmically degrade in reliability [31], [32].

Negative Bias Temperature Instability – NBTI

Negative Bias Temperature Instability phenomenon is considered as one of the most important aging reliability issue in integrated circuits, as a result of technology scaling and the increase of device operating temperatures.

NBTI manifests as a degradation of linear/saturation electrical parameters of a MOS transistor under a negative V_{GS} (for a pMOS transistor) with a high-temperature dependence. The dynamic of this phenomenon has usually a time power dependence, leading to an overall degradation that is monotonous over time. This degradation becomes worse when the temperature increases, but it also depends on the type of oxide (SiO2, SiON, HfO2, HfSiON) the transistor is made of and its thickness [33, 34]. At device level, this degradation caused by the continuous stress applied on the transistor, is usually quantified as an important increase of the threshold voltage (V_{th}) and a current reduction.

The phenomenon can be reversed if stress conditions cease and the circuit is restored to idle state with the threshold voltage returning to its original value. Typically, stress on a transistor is not permanent but increases and decreases with time (AC stress). However, as a full recovery is not feasible, there is a permanent gradual degradation in transistor performance, but at a slower pace.

As the thickness of the gate insulator deviates around 1nm in modern nanotechnologies (even with decreasing voltage levels), the vertical electric field in the insulator, up to a few MV / cm2, intensifies the BTI phenomenon. The problems related to the BTI mechanism is further exacerbated with each new generation of integrated circuits, as the transistors' characteristic size keeps decreasing and as the number of transistors integrated on a chip and their frequency of operation increases, leading to a greater statistical deviation of transistor performance degradation [30].

NBTI can produce more than 20% timing degradation in the worst-case operating conditions [35], which means that the performance degradation, especially of the synchronous circuit systems, is becoming more intense.

Extensive research has been made on the NBTI phenomenon. The two main models that explain both PBTI and NBTI mechanisms are:

- Reaction-Diffusion Model (R-D model)
- Trapping-Detrapping (T-D model)

The following explanation of both the models is based on the NBTI effect on pMOS transistors. However, the same models can be used to describe the PBTI mechanism in nMOS transistors.

Reaction-Diffusion Model

In order to predict the behavior of an IC due to NBTI effects, the Reaction-Diffusion (R-D) model is the most widely used for the estimation of the V_{th} degradation. [15, 36].

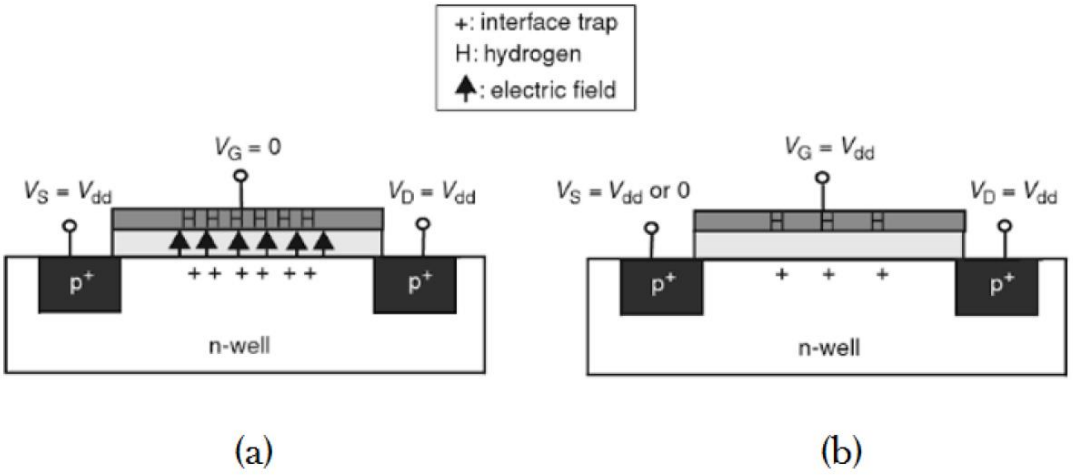

Figure 2.6 R-D model of NBTI in pMOS transistor (a) Stress Phase (b) Recovery Phase [37].

When pMOS transistors are manufactured, due to certain inaccuracies in the processes, some of the silicon atoms may bond with hydrogen atoms at the interfacial layer (layer between the n-well and the gate insulator). During the transistor operation, a negative gate-to-source voltage is applied so it turns ON (V_G = 0, V_S = V_{DD}). Negative V_{GS} repels the electrons in the n-well so as to attract holes below the oxide substrate, forming a positive conducting channel between the source and drain, thereby, turing the transistor ON. At high temperatures positive charges enter the silicon-hydrogen bond and cause the hydrogen atoms to break free. Since the gate is negatively charged, the hydrogen atoms crowd near the gate, leaving holes (due to dangling silicon atoms) at the interfacial layer, as shown in Figure 2.6(a). The generation of these holes at the interface of the oxide leads to an increase in the threshold voltage of the transistor. This is called the **stress phase** of NBTI.

When the negative gate voltage is removed (V_G = V_{DD} and V_S = V_{DD}), the hydrogen atoms drift back toward the Si-Oxide interface, recombining with the silicon atoms to form Si-H bonds again. Therefore, the NBTI phenomenon is partially reversible. This phase is called the **recovery phase**, as shown in Figure

2.6(b). The recovery phase can be accelerated by applying a positive gate voltage ($V_G = V_{DD}$ and $V_S = 0$). However, as seen from Figure 2.6(b), the recovery is not complete, since a portion of hydrogen atoms diffuse away permanently. This leads to an accumulation of holes in the interfacial layer [16]. The stress-recovery combination is referred to as **dynamic NBTI**. On the other hand, if the transistor is constantly under stress, it is said to undergo **static NBTI**. Dynamic NBTI is more accurate in representing actual transistor operation because a transistor does not remain always ON.

Post the recovery phase, the few holes at the interfacial layer that are not recovered, are filled by electrons from the n-well. Thus, when the transistor is turned ON the next time, higher voltage needs to be applied to repel those electrons in order to form the conducting channel below the substrate. Specifically, this voltage increase essentially translates to an elevated threshold voltage. As the transistor is in continuous operation, the threshold voltage undergoes degradation, which is elevated at high temperatures. Ultimately, this worsens the switching speed of the transistor [16].

As the number of silicon-hydrogen bonds in a transistor is finite, breaking and reconnecting them presents a significant statistical deviation throughout the degradation process [38]. This statistical deviation of the BTI mechanism (dynamic BTI phenomenon) results in an additional random deviation of the threshold voltage, upon its expected degradation due to the permanent BTI mechanism (static BTI phenomenon).

Trapping-Detrapping Model

The trapping-detrapping (T-D) model is proposed as an alternative to explain the Bias Temperature Instability phenomenon. One of the primary advantages of using the T-D based model is that, it exhibits a logarithmic dependence on the time evolution of BTI, threshold voltage and gate delay shifts can be determined

and thus a realistic aging rate can be specified, whereas the R-D model may overestimate aging effects as it tends to predict a very low degradation when the supply voltage is changed from a higher V_{DD} to lower $V_{DD.}$ [39].

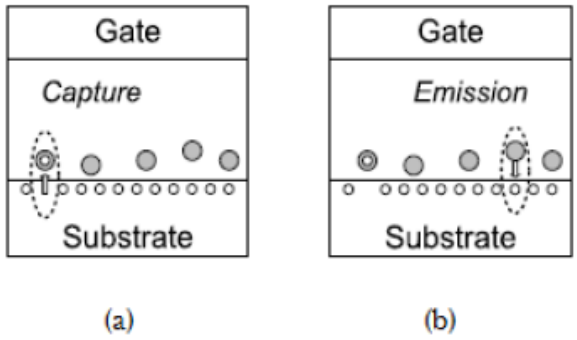

Figure 2.7 (a) Charge trapping component and (b) Charge detrapping component of T-D model [39].

Interface traps are electrically active defects that are present along the interface between the gate dielectric and the substrate. According to trapping-detrapping theory, traps located in the gate dielectric or at the silicon interface capture and re-emit some of the charge carriers responsible (Figure 2.7), for the current flowing between source and drain of a MOSFET [15]. When a trap captures a charged carrier, the transistor's threshold voltage increases, which constitutes the stress phase [40]. On the other hand, when the trapped carriers are released due to positive V_{GS}, it results in recovery and leads to a decrease in the number of occupied traps as seen in Figure 2.7(b) [40]. Previous research has revealed that the statistical probability of trapping depends on the time constants, the number of traps, the location of traps and the trap energies. Faster traps (i.e., shorter capture time constants) have a higher probability of getting filled compared to the slower traps. Furthermore, the trap occupation probability increases with gate bias and temperature [40].

However, much effort is needed to develop an accurate, compact aging model based on this mechanism. Thus, the R-D model is mostly used for modelling

purposes. Even though the R-D model suited the observations of the early technologies well, measurements using more recent technologies' transistors showed significant deviations. The R-D models seems failing to explain the time constants, observed in BTI for small devices, compared to the T-D model which sufficiently succeeds to explain all observed NBTI and PBTI data.

Positive Bias Temperature Instability – PBTI

When a positive gate-to-source voltage is applied to the nMOS transistor, it is in inversion and experiences PBTI. Similar to NBTI, PBTI effect can also be approached with the R-D model. PBTI was not very significant in earlier technology nodes. With the recent use of high-$\varkappa$ metal gates, it causes almost the same amount of V_{th} degradation as NBTI.

2.4.5. Hot Carrier Injection - HCI

If the electric field (voltage) is high enough at the transistor gate and drain, then electrons or holes can gain enough energy from the field to reach the interface, to overcome the energy barrier between silicon and oxide, to enter into the oxide and get trapped there as presented in Figure 2.8. High energy electron injection is more likely than hole injection because (a) the electrons have a lower active mass, so they can easily gain energy from the field than the holes, and (b) the energy barrier on the Si-SiO_2/HfO_2 interface is larger for holes than for electrons [41].

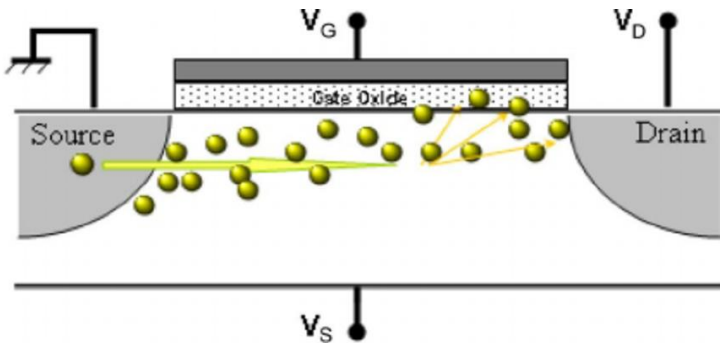

Figure 2.8 Bias Temperature Instability Phenomenon.

Hot Carriers Injection contributes to the gradual increase in the transistor absolute V_{th} threshold voltage as well as to the reduction in the mobility of the carriers, thereby reducing the performance of integrated circuits (operating speed). This phenomenon has a significant impact on current technologies due to the increasing electric fields per unit area in the transistors used [36].

As the electrons are warmer than the holes, the effect of the HCI effect has been found to be more significant in the nMOS transistors than in the pMOS transistors [36]. The elimination of HCI stress partially reduces the intensity of the phenomenon, but the recovery is not as significant as in the BTI mechanism.

Countering the HCI phenomenon is particularly important as it leads to the occurrence of intermittent faults in the form of timing faults that cause timing errors. The presence of the HCI phenomenon for a prolonged period of time has shown the conversion of intermittent faults to permanent ones, which makes the problem even more crucial.

2.4.6. Single event effects – SEEs

Simple event effects occur when high-energy particles (e.g. alpha particles or cosmic ray induced neutrons) hit the integrated circuit, transferring enough energy to alter the polarity of a node in the circuit (Figure 2.9). SEEs can be destructive or transient, depending on the energy of the particles and where they hit.

An important consequence of SEEs in circuit logic is the induction of Single Event Transients (SETs). The state of a circuit node changes and a pulse may propagate up to the primary outputs of the affected circuit. The minimum amount of collected charge that generates a SET on the affected circuit node is usually referred to as *critical charge*. In combinational logic, the generated voltage glitch may propagate through the downstream logic and be captured by a memory

element, thus resulting in a soft error. When the generated voltage glitch affects directly a memory element of the circuit, a Single Event Upset (SEU) occurs. SEUs result in the inversion in the stored value of a flip-flop, latch or memory cell due to the radiation induced charge.

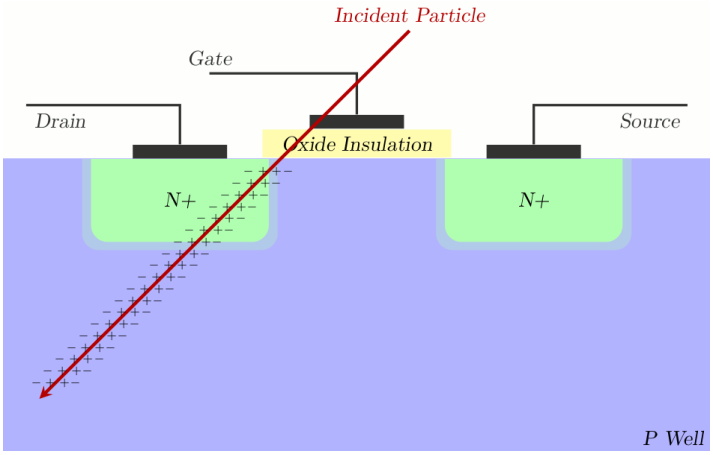

Figure 2.9 Single Event Effect generation mechanism.

2.5. Effect of Aging

Mechanisms such as BTI and HCI [42], [43] contribute to the aging of the pMOS and nMOS transistors and along with the process variations that occur during manufacturing (e.g. in the thickness of the gate oxide [44]), they lead to performance degradation of the integrated circuits.

Aging manifests differently at various levels of a circuit. At the lowest level of abstraction, aging causes deterioration in the transistor device parameters, such as threshold voltage, leakage current etc. The extent of change depends on several factors, such as temperature, voltage, and activity factor. Gates that contain such degraded transistors experience an increase in the threshold voltage and the gate propagation delay. The gate delay against the threshold voltage for a healthy and an aged gate is shown in Figure 2.10.

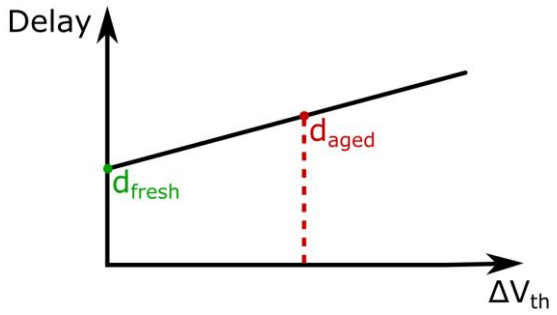

Figure 2.10 Gate delay degradation as a linear function of V_{th}.

Figure 2.10 shows that as transistors experience aging, their threshold voltage shows a gradual increase, which causes a significant impact on the propagation delay of the gate where they belong. Similar to combinational gates, sequential elements in the circuit (e.g. flip-flops) also have increased propagation delay due to aging. At the circuit level, the critical path of the circuit dictates the timing requirements, i.e., the critical path delay determines the operating frequency of the circuit. As long as the delay of the critical path does not exceed the maximum permissible delay, the circuit's reliability is retained. As a result of aging, the critical path delay may increase over time, resulting in timing violations and circuit failure.

A more eminent reliability issue due to aging is that it introduces new critical paths over time. A circuit may also contain near-critical paths, which exhibit a delay closer to the critical path delay. Sometimes, the critical path of the circuit may age slowly when compared to the near-critical paths. In such cases, timing violations may arise from near-critical paths, and thus, designers must take those paths into careful consideration too. Therefore, as time progresses, aging can turn a non-critical path into a potentially critical one.

The combination of the above aging factors exaggerates the adverse effects of the reliability degradation of circuits, such as the ability to provide correct responses in

certain time frames. Subsequently, the IC industry requires design solutions that provide tolerance to the presence of aging in order to achieve correct operation of the integrated circuits and ensure their integrity throughout their lifetime.

Various techniques have been introduced in the international literature to address the aging phenomenon in the field of application. However IC industry also requires the development of techniques that will be capable of ensuring not only reliability, but also low cost on silicon area and energy consumption.

2.6. The Self-Healing Concept

Self-healing is increasingly becoming a promising approach to designing reliable systems and it refers to the ability of a system to detect faults or failures and fix them through healing or repairing with the minimum impact on its performance. The role of self-healing is to recover the fault in the system and to keep it working with the highest possible performance and for a long time as shown in [45]-[47]. Modern systems with architecture for self-healing are expected to compensate faults. The idea behind the self-healing architecture is shown in Figure 2.11. Self-monitor mode monitors the system and gives an indication of irregular events. The next step is that the system switches to a self-diagnosis mode that identifies the fault, and information will be extracted with respect to the problem cause, symptoms, and impact on the system. Once these are identified, the system will try to adapt itself on the way of generating candidate fixes, which are tested to find the best-expected state [48].

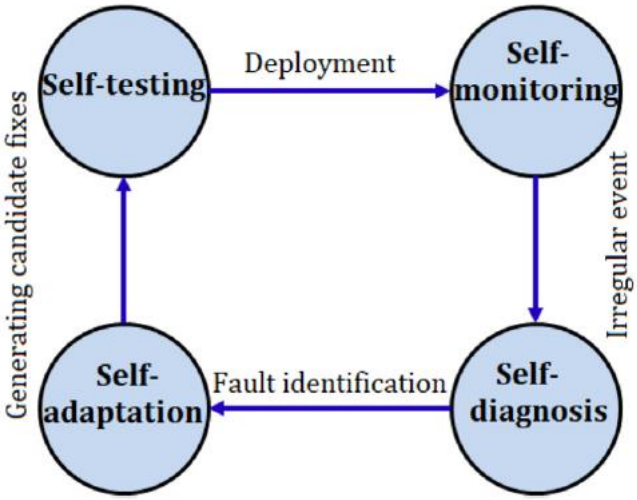

Figure 2.11 General Architecture of Self-Healing.

Some of the self-healing methods can be applied for off-line while others for on-line healing. In off-line healing, the system has to be off as the self-healing method requires time to isolate faulty parts and recover them. On the other hand, partially reconfiguration can be used to allow the system operation while fixing faults. Thus, real-time healing is challenging, especially in states which require a continuous running such as bio-medical and military applications.

2.7. General Categories of Aging Monitoring Techniques

Under the above general concept of self-healing and as aging effects have been proven to seriously affect modern integrated circuits and systems, various techniques have been proposed targeting the monitoring and repairing of aging. These techniques can be classified into three general categories:

Error detection and correction techniques

Error detection and correction techniques provide the ability of detecting a transient error occurrence (e.g. a bit in a memory that has flipped) during the operation of an integrated circuit and then reconstruct the original error-free data to overcome this issue. They are based on the monitoring of data path signals to

detect transitions that arrive after the clock triggering. In the available literature we identify techniques that suggest the use of stability checkers [49] to observe delayed transition arrivals, circuits to detect delay faults for self-checking applications [50], as well as signal resampling methods after detecting a signal delay [51]. One of the most common techniques proposed for on-line timing error detection is the Razor technique [52], which adopts the use of a shadow flip-flop to detect timing failures due to setup violations on the logic stage paths, using the double sampling technique. Also important are the techniques that suggest the addition of a sense amplifier to detect errors [53], as well as those that suggest triplication of the circuit and integration of voters to identify the correct value [54]. Finally, error correcting codes (ECCs) are widely used for the detection and correction of errors. By exploiting these codes failures can be avoided during a system operation [55].

Error prediction techniques

Error prediction techniques provide the ability of sensing a given propagation path delay violation and thus detect a potential fault before its occurrence. This is in contrast to classical error detection techniques where a failure is detected after errors appear in the system. They are based on monitoring the signals of a data path for a specific period of time before the clock edge. An error prediction technique, called the canary flip-flop [56], pads the data-path with a delay element and samples the delayed data-path signal in another flip-flop. A timing error is predicted when the value on the flip-flop of the data path differs from the value on the canary flip-flop. Another technique presented in [57] proposes for each critical path the creation of an identical one. The error prediction is based on duplicating the critical paths and using timing errors on the duplicated paths to predict a timing error on the original paths. The effectiveness of this approach is limited as the two paths may differ due to process variations and the critical paths may also change over time [58]. An important technique of this category is described in [58]. This BTI degradation estimation method detects over-aged cells

(near failure) and replaces them with spare ones using embedded repair mechanisms. According to this method, in NBTI testing operations the current strength of the pMOS transistors in the cell is used for the detection of aged devices, as it will be further described in Chapter 3.5.1.

As it is obvious, such techniques require an extra number of transistors as well as additional time dedicated both to prediction and recovery. Although today aging aware designs ensuring the reliability of an IC have become mandatory, the demand of speed and reduced cost should be also taken into great consideration during the design process. The present dissertation aims to develop an easy to implement on-line error prediction technique for SRAMs with low cost in silicon area and with reduced requirements in time.

Error masking techniques

As technology moves to areas where the supply voltage is very close to the transistor threshold voltage and there is a continuing trend to increase the degree of integration, there is a need for flexible techniques that will tolerate errors. That is, techniques that will help to adapt the circuit to function properly even in the presence of faults. Error masking techniques logically mask timing errors by adding extra logic.

The error masking techniques proposed in the literature can be divided into two categories:

Logical redundancy: They use redundant logic to calculate the correct value of the output with a small delay when critical paths are exercised.

Temporal redundancy: They are tolerant to errors through time-borrowing, delaying for example the arrival time of the clock edge to the next stage of a pipeline [59].

In [60], the authors propose a temporary error masking technique based on stalling the clock signal for a cycle after the detection of a timing error, in order to correct the condition of the system. This technique proposes the replacement of some suspicious flip flops with other sequential circuit elements having time-borrowing capability and corrects timing errors by delaying the arrival time of the correct data to the next logic level. It assumes that the latency errors detected by the replaced elements can be corrected during one cycle stall of the clock. In [61] an edge detector is proposed for detecting circuit timing violations near the positive edge of the clock. In this case, a delayed clock signal is used for the resampling and correction of the data values with time borrowing from the next stage of the pipeline. This technique assumes that this time is absorbed by a non-critical path of the next stage. This assumption may not be valid and it may lead to timing errors, especially in high performance implementations. Additionally, the edge detector circuit is based on precise delay values and the clock period may have to increase due to variations in the manufacturing process variations.

CHAPTER 3

AGING EFFECTS ON SRAMS

This chapter deals with aging mechanisms in SRAM memories and their influence on the memory's reliable operation. Initially, the most popular memory types are briefly presented. Next, the typical SRAM structure with its elements is analyzed. The effect of transistor aging on the performance characteristics of the SRAM memory cells, decoders and sense amplifiers is discussed. Finally, the state of the art aging monitoring techniques in the open literature are presented.

3.1. Memory Types

In general a Computing System is divided into three subsystems: i) the Central Processing Unit (CPU), ii) the input-output (I/O) devices and iii) the memory [10]. The memory subsystem usually consists of two memory categories [10], [62]:

- The mass storage devices (secondary memory), such as hard disk drives, compact disks etc. Their main characteristic is their large storage capacity with low cost/capacity ratio and low access speed. They are used for permanent data storage purposes.

- The main memory, the cache memory and the register file, which store data and programs during processing. They are characterized by high access speed and high cost/capacity ratio. The silicon area of the memory in a computer system is significantly high compared to the silicon area of the combinational logic. Therefore a failure in the memory subsystem is one of the main causes of a computing system failure.

Various memory types of the second category are used, depending on the requirements of each subsystem [10], [62]. The most important of them are as follows:

- Static Random Access Memory – SRAM

This memory type has the highest access and data transfer speed, due to its very low latency and very high bandwidth. Its great disadvantage is the low data storage density per silicon area which highly increases the cost/capacity ratio. It is mainly used as cache memory, where the main demand is the high performance. It is a volatile type of memory. Its lifetime is crucially impacted by aging phenomena.

- Dynamic Random Access Memory –DRAM

This memory type is characterized by high data storage density per silicon area and relatively high access and data transfer speed. In other words, it is cheap compared to SRAM and it is fast compared to a hard disk. For these reasons it is mainly used as the main memory of a system. Its main disadvantage is that a periodical refresh procedure is necessary in order to retain the data. It is a volatile

type of memory, since the data stored are lost when the power supply is disconnected.

- Read-Only Memory – ROM

These memories are non-volatile and the data stored at them are pre-stored by the manufacturer and cannot be altered, expanded or deleted by the user. Usually they store basic information that is needed by the microprocessors in order to perform basic operations such as interaction with keyboards, display, disks etc.

- Programmable Read-Only Memory – PROM

This is a variation of a ROM memory which is not pre-programmed by the manufacturer but it is programmed by the user. After it is programmed, the data stored can no longer be altered.

- Erasable Programmable Read-Only Memory – EPROM

This is a variation of the Programmable Read-Only Memory which can be programmed by the user more than once. In order to be reprogrammed, firstly it has to be removed from the system and its data must be erased using ultra-violet radiation. Its main disadvantages are that the programming procedure requires special equipment, it is sensitive to light, it has lower performance than ROM and its packaging is expensive.

- Electrically Erasable, Programmable, Read-Only Memory – EEPROM

Nowadays, they are widely used and are known under the name "Flash Memories". They are read-write memories with non-volatile capability and are used as a permanent storage. Despite the fact that their cost per bit is significantly high, their extremely small dimensions and weight make them ideal for the storage unit of cameras, video cameras and other portable devices. Note that the write operations cause gradual wear out to the memory cells and, thus, they can sustain a limited number of write operations before they start to fail. Moreover, their data retention time has limitations.

3.2. SRAM Architecture

Memory is one of the most universal cores in that almost all *system-on-chip* (SoC) devices contain some type of embedded memory [1]. Nowadays embedded static random-access memories (SRAMs) are widely used as high-speed cache memories inside microprocessors because of their high access speed.

In general memories consist of the memory cells where the data are stored (memory array) and a number of assisting circuits [62]-[66]. In the latter case, the most important assisting circuits are the *sense amplifiers (SA)*, the *precharge circuits*, the *I/O circuits* and the *address decoders*. The functionality of these circuits will be analyzed in the paragraphs that follow.

3.2.1. SRAM Memory Array

Memories consist of memory cells, which are arranged in matrix like structures called *memory arrays* and a number of peripheral circuits. In Figure 3.1 an SRAM memory array along with the basic assisting circuits is presented [62]. Each cell is a circuit that stores a single bit. The cell matrix of an SRAM memory has 2^m rows (*Word-Lines*) and 2^n columns (*Bit-Lines*), for a total storage capacity of 2^{m+n}. When a row address of "m" is given to the row decoder, it activates one of the 2^m word-lines of the array. When a column address of "n" is given to the column decoder, it activates one of the 2^n bit-lines of the array. A particular cell is selected for reading or writing by activating the word and its bit-line. A bit-line is a data transfer line to which one memory cell is connected and a word-line is the activation line for a read or write operation of a cell.

The row decoder, which activates one of 2^m word-lines, is a combinational logic circuit that raises the voltage of a particular word-line whose m-bit address is applied to the decoder input.

The sense amplifier is applied to every bit-line in order to read the small voltage signal provided by cells. The signal is then delivered to the column decoder, which selects one column based on bit address, causing the signal to appear on the chip I/O data line [67].

The precharge circuitry pre-charges and equalizes the bit-lines allowing a proper and easier detection of the cell's data by the sense amplifier. In this way, different blocks function together to facilitate the read/write operations.

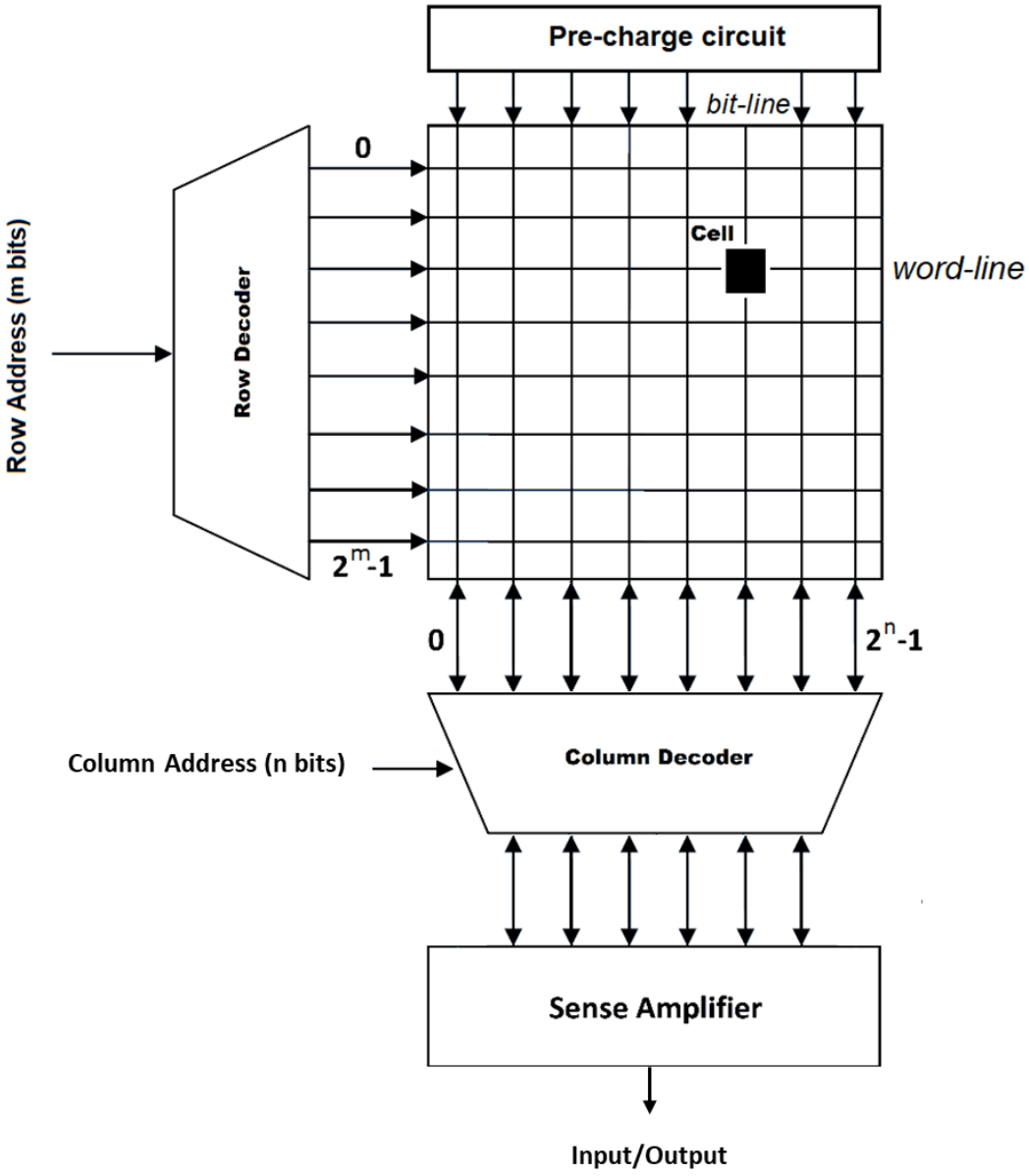

Figure 3.1 An SRAM array along with the peripheral circuits.

3.2.2. Precharge Circuit

In SRAM memories during normal mode, the voltage level of the bit-lines is unpredictable. Consequently, all bit-lines need to be precharged to a certain voltage level (V_{DD} in typical SRAM designs), in order to have all the array columns ready for a new read operation. The precharge and equalization circuit is used at the beginning of the read operations to precharge each column (bit-line).

It consists of three pMOS transistors, as shown in Figure 3.2. Usually all these transistors have the same L and W. One of them, which is called the *equalization transistor*, connects the two bit-lines with each other while the other two transistors connect each bit-line with the V_{DD} power supply. When the circuit is activated (*PRE* signal set to low) voltage of the two bit-lines starts to move towards V_{DD}.

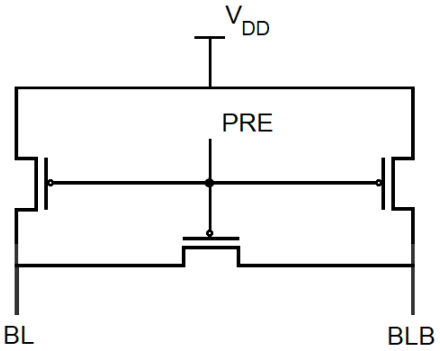

Figure 3.2 The precharge and equalization circuit.

3.2.3. Word-Line Driver

The word-line driver is the circuit responsible for activating and deactivating the word-line. Ideally it raises the word-line to V_{DD} and drops it to 0V within the appropriate time interval. However, deviations from the ideal behavior may result to a delay or even to a failure to set the appropriate voltage on the word-line, influencing the operations performed on the cells attached to it.

3.2.4. Address Decoder

The Address Decoder is responsible for accessing the appropriate cell according to a given address and is divided in two sub-circuits: the row decoder and the column decoder. The row decoder generates a word-line activation signal to access a specific row, while the column decoder generates a bit-line activation signal to access a specific column or columns of the memory array in order to obtain the data read or provide the data to be written. Thus, the address is partitioned into a *column address* and a *row address.*

Row Decoder

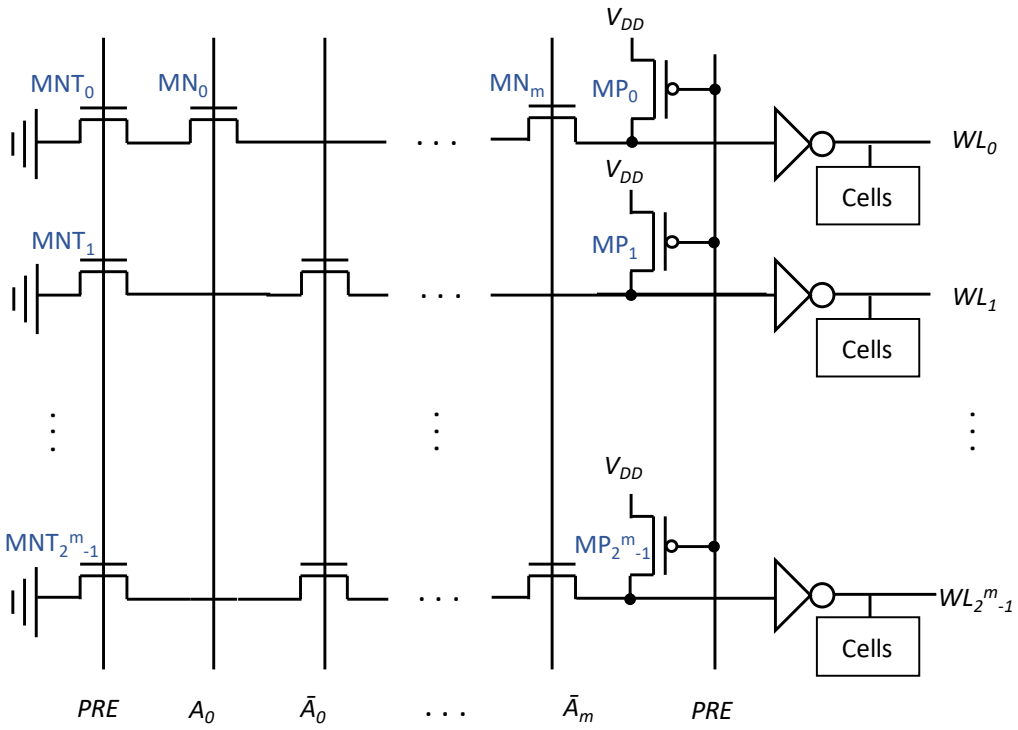

Figure 3.3 An m-to-2^m NAND Row Decoder design.

The topology of a typical m-to-2^m NAND type dynamic logic row Decoder (m input address bits to 2^m rows) is shown in Figure 3.3. This address decoder is made with NAND gates, and each NAND gate is connected with the appropriate

address, corresponding to a word-line. It is composed of 2^m slices, where each slice consists of m+1 nMOS transistors. One of them is the activation transistor and is driven by the PRE signal (as the precharge circuit) while each one of the rest m nMOS transistors is driven by a bit, or its complementary, of the row address. A precharge transistor along with an inverter are connected to each slice and the word-line activation signal (*WL*) is generated at the output of the inverter. Depending on the input address, a Decoder slice is activated and the corresponding word-line is turned to high.

The precharge signal *PRE* allows the initialization of all word-lines to low. The Decoder then operates in two phases. When the memory is not accessed, the signal *PRE* is set to low and the pMOS precharge transistors MPi are activated for the initialization of the word-lines to low. During a read or write operation, the *PRE* signal is turned to high so that the nMOS networks are active, the address is decoded and a single word-line is activated (the pertinent word-line signal *WL* is set to high). Consequently, only a single word-line is being pulled up after the precharge while all the rest remain low.

Column Decoder

The function of a Column Decoder is best described as a bidirectional 2^n multiplexer, where *n* stands for the size of the column address word. During the read operations, they have to provide the signal path from the cell to the sense amplifier. When performing a write operation to the memory array, they have to provide signal path from the write driver to the cell.

The typical topology of an n-to-2^n Column Decoder (n input column address bits to 2^n bit-lines) is shown in Figure 3.4. The multiplexer is implemented with pass transistors, and each of the 2^n bit-lines is connected to the sense amplifier through

an nMOS transistor of the decoder. A NOR type decoder can be used to select one of 2^n bit-lines.

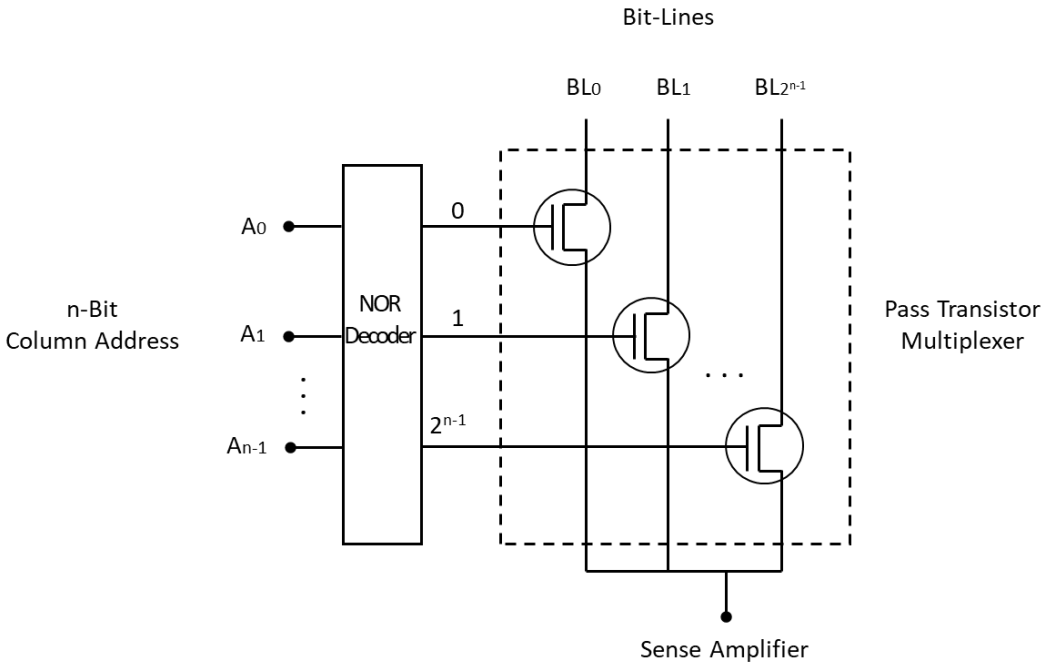

Figure 3.4 Pass-Transistor-Based Column Decoder design.

3.2.5. SRAM Memory Cell

The most common SRAM cell is the 6T SRAM memory cell which uses six MOSFET transistors as seen in Figure 3.5 It consists of a pair of inverters (MCP1, MCN1 and MCP2, MCN2) in a cross-coupled topology and two access transistors (MCN3 and MCN4) that are driven by the word-line (*WL*). The cross-coupled inverters which form a latch are connected to the bit-lines *BL* and *BLB* respectively through the access transistors. When the word-line turns high, the access transistors are connected with a bidirectional stream of current between the cell and the bit-lines.

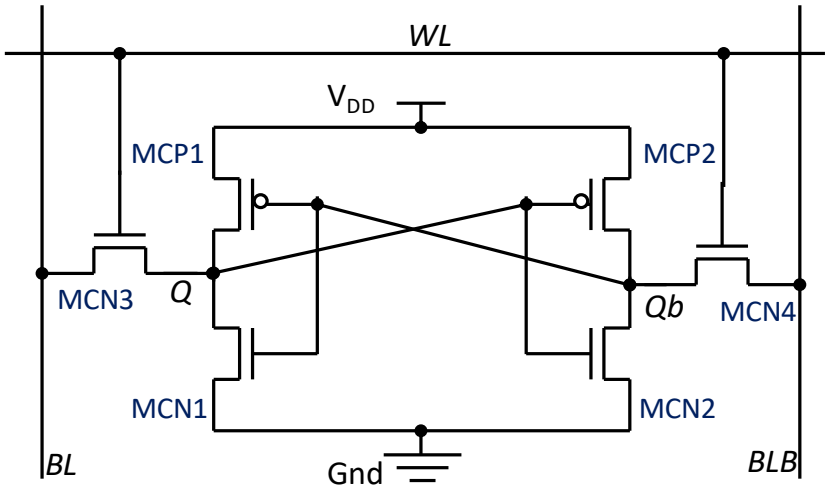

Figure 3.5 Typical topology of a 6T SRAM memory cell.

The cell has two stable states, logic '0' and logic '1'. Since the inverters function as long as power is supplied to them, SRAM keeps its value as long as there is power. The access transistors allow for writing and reading data to and from the bit cell respectively.

The SRAM cell should be sized as small as possible to achieve high memory densities. Reliable operation of the cell however, imposes some sizing constraints. To understand the operation of the memory cell, next we analyze the read, write and hold operations.

Read Operation

Aiming to read a data bit from a memory cell, the bit-lines *BL* and *BLB* are initially precharged to V_{DD} with the use of the precharge circuit. Then the precharge circuit is deactivated and the proper word-line is selected by turning the corresponding *WL* signal to high. When the word-line is selected ($WL=V_{DD}$), the memory cell is connected to the bit-lines. Considering the stored bit at Q node to be logic '0', the bit-line *BL* which is precharged to V_{DD} will start discharging through MCN3 and MCN1 in Fig. 3.5. This will gradually reduce the potential on

the bit-line *BL*. At the same time, since the stored bit at Qb is '1' and the bit-line *BLB* is high, the bit-line *BLB* will continue to maintain its potential at V_{DD}. This will create a voltage difference between the *BL* and *BLB* bit-lines which feeds the differential inputs of the sense amplifier so that the presence of logic '0' stored into the cell will be detected and appear at the output of the sense amplifier. The higher the sensitivity of the sense amplifier, the faster the speed of the read operation.

Figure 3.6 shows the setup of the cell at the beginning of a read operation.

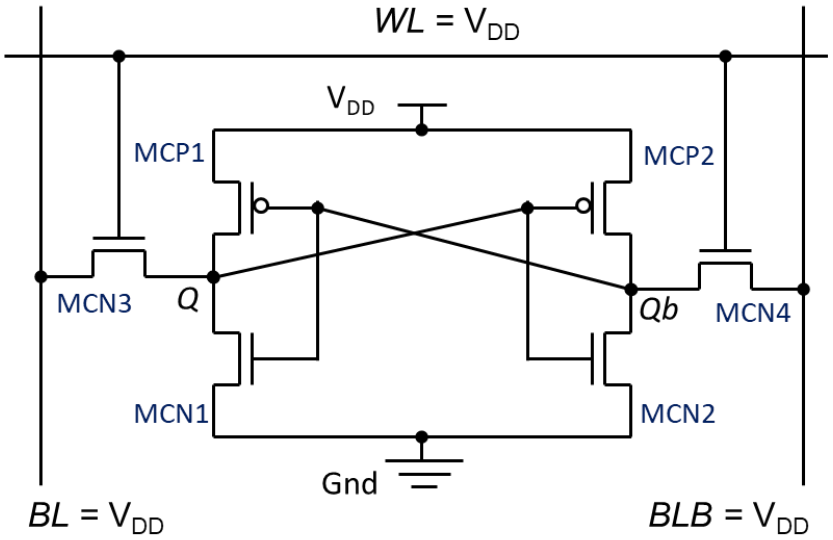

Figure 3.6 SRAM Cell Setup at the beginning of a Read Operation.

For a successful cell read operation, a design constraint called **Cell Ratio** (CR) [68] should be followed. Considering the case where node Q is at logic '0' and Qb is at '1', in order to read the stored value both bit-lines are set to V_{DD} and signal *WL* is activated. Transistors MCN3 and MCN4 turn ON to give bit-lines access to the inner cell. Since Qb is at logic '1', the potential on *BLB* is expected to remain at V_{DD}. Since Q is at logic '0', bit-line *BL* will start discharging through MCN3 and MCN1. While discharging, the active transistors MCN3 and MCN1 form a voltage divider. During the discharge, the potential at Q raises as transistor MCN1 acts as

a resistance. If this potential at Q raises above the switching threshold of inverter 'MCP2-MCN2', the pMOS transistor MCP2 will be turned OFF and the nMOS transistor MCN2 will be turned ON which will result in the reversal of data bits and thereby the data will be corrupted. In order to avoid this situation, the resistance of MCN1 should be less than that of MCN3 i.e., the size of MCN1 should be made larger than that of MCN3.

The Cell Ratio is defined as the ratio of (W/L) of the nMOS pull-down transistor MCN1 (MCN2) to the (W/L) of the nMOS access transistor MCN3 (MCN4).

$$\text{Cell Ratio} = \frac{W_{MCN1}/L_{MCN1}}{W_{MCN3}/L_{MCN3}} \quad \text{and} \quad \frac{W_{MCN2}/L_{MCN2}}{W_{MCN4}/L_{MCN4}} \qquad \text{Eq. 3.1}$$

Write Operation

To write a value on the SRAM cell, the column decoder selects the bit-line and the write driver injects the data value (logic '0' or '1') intended to store on the memory cell. Assuming that both bit-lines are precharged to V_{DD} and supposing that the cell's previously stored logic value is '1' and the value to be written is the logic '0', the *BL* is pulled down to ground and *BLB* remains at V_{DD}. This will cause the cell to change state if the transistors are sized properly. It is reasonable to assume that the gates of transistors MCN1 and MCP2 are at V_{DD} and gnd respectively, as long as the switching has not commenced. Then, the *WL* signal is enabled, raises to V_{DD}, activating the access transistors MCN3 and MCN4 and giving the bit-lines access to the inner bit cell.

Since *BL* is set to '0' and is given as input to the inverter 'MCP2-MCN2', transistor MCP2 is turned ON and transistor MCN2 is turned OFF raising the node Qb to V_{DD} (logic '1'). Since *BLB* is set to '1' and is given as input to inverter 'MCP1-MCN1', transistor MCP1 is turned OFF and transistor MCN1 is turned ON pulling the node Q to ground (logic '0').

Figure 3.7 shows the cell during the writing operation of a logic '0' in a memory cell, while Figure 3.8 shows the memory cell during the writing operation of logic '1'. In both cases it is assumed that the new cell value is written at node Q.

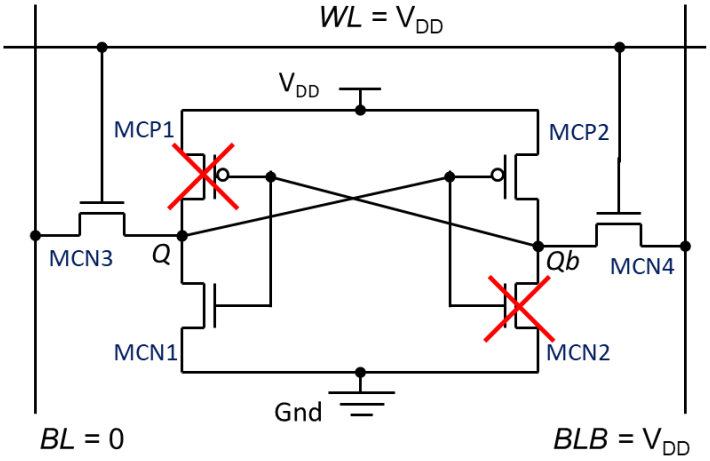

Figure 3.7 SRAM Cell Writing Operation of '0'.

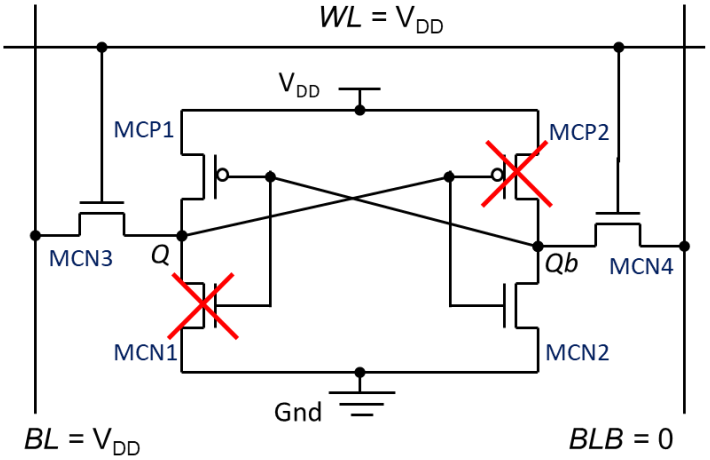

Figure 3.8 SRAM Cell Writing Operation of '1'.

For the successful writing of a value to the cell, a design constraint called **Pull-up Ratio** (PR) should be followed. Considering the case where node Q is at logic '0' and Qb is at '1' in order to change the state of the cell, logic '1' needs to be written to Q and '0' to Qb. According to the procedure, the *BL* is raised to V_{DD}

and *BLB* is pulled down to gnd. Next, the *WL* is activated. In this case, the pMOS transistor MCP2 and the nMOS access transistor MCN4 are in ON state. Since *BLB* is pulled to gnd, node Qb will start discharging through MCN4 and at the same time it will be charged through MCP2. Since both MCN4 and MCP2 are ON, they will act as a voltage divider. In the case where the potential at Qb is not brought below the switching threshold voltage of the inverter 'MCP1-MCN1', the state of the bits will not be able to change. In order to achieve a successful write operation, the potential at Qb should be below the switching threshold of the inverter 'MCP1-MCN1' i.e., the resistance of transistor MCN4 should be less than that of MCP2 which in-turn means that the width of MCN4 should be greater than that of MCP2.

Pull-up Ratio [68] is defined as the ratio of (W/L) of the pMOS pull-up transistor MCP1 (MCP2) to the (W/L) of the nMOS access transistor MCN3 (MCN4).

$$\text{Pull-up Ratio} = \frac{W_{MCP1}/L_{MCP1}}{W_{MCN3}/L_{MCN3}} \quad \text{and} \quad \frac{W_{MCP2}/L_{MCP2}}{W_{MCN4}/L_{MCN4}} \qquad \text{Eq. 3.2}$$

Lowering the Pull-up Ratio ensures a more successful writing operation of the cell. In general, for a standard 6T cell the PR is kept to 1 while the CR is varied from 1.25 to 2.5 for a functional cell, in order to have a minimum sized cell for high density SRAM arrays. Therefore, in high density and high performance standard 6T cell, the recommended value for CR and PR are 2 and 1, respectively [69].

In conclusion, the bit-line write-drivers need to be stronger than the relatively weak transistors in the cell itself so that the previous state of the cross-coupled inverters can be easily overridden. In addition, the nMOS access transistors (MCN3, MCN4) have to be stronger than the top pMOS (MCP1, MCP2) transistors, while the bottom nMOS (MCN1, MCN2) must be stronger than the nMOS access

transistors (MCN3, MCN4). Consequently, when the pair of transistors on the one side of the cell is only slightly overridden by the writing operation and starts switching, the other side eventually follows, engaging the positive feedback. Thus, cross-coupled inverters magnify the writing process.

Hold Operation

The state where the data is neither written to the cell nor read from the cell is called hold/standby state. During this state, signal *WL* remains low so that the access transistors MCN3 and MCN4 are turned OFF, which disables the bit-lines' access to the inner cell. During this stage, the cell is unaffected by the voltages on the bit-lines. The two cross-coupled inverters formed by MCN1 – MCP2 continue to reinforce each other as long as they are connected to the power supply.

3.2.6. Sense Amplifier

Sense amplifiers play a major role in the functionality, performance and reliability of memory circuits. In particular, they perform the following functions:

- *Amplification* – In certain memory structures, such as the SRAMs, amplification is required for proper functionality by detecting small voltage differences on the bit-lines, thus achieving reduced power dissipation and delay [10].

- *Delay Reduction* – The amplifier compensates for the restricted fan-out driving capability of the memory cell by accelerating the bit-line transition, or by detecting and amplifying small transitions on the bit-line to large signal output swings.

- *Power Reduction* – Reducing the signal swing on the bit-lines can eliminate a substantial part of the read power dissipation.

The topology of the sense amplifier is a strong function of the type of memory device, the voltage levels and the overall memory architecture. SRAM memories most usually utilize *Differential Voltage Sensing Amplifiers* for signal sensing.

It is generally known that a differential approach presents numerous advantages over its single-ended counterpart – one of the most important being the *common-mode rejection.* That is, such an amplifier rejects noise that affects both inputs. This is especially attractive in memories where the exact value of the bit-line signal varies from die to die and even for different locations on a single die. In other words, the absolute value of a '1' or '0' signal is not exactly known and might vary over quite a large range. The phenomenon is further complicated by the presence of multiple noise sources, such as switching spikes on the supply voltages and capacitive cross talk between word and bit-lines. The impact of those signals can be substantial, especially when the amplitude of the signal to be sensed is generally small. The effectiveness of a differential amplifier is characterized by its ability to reject the common noise and amplify the true difference between the signals [10].

As previously stated, the sense amplifier is the circuit that senses the voltage difference between the bit-lines after the word-line is activated, enhances and finally maximizes this voltage difference, setting the one bit-line to V_{DD} and the other to 0V. This module is important for the proper operation of an SRAM memory as by amplifying small voltage differences, the read operation is accelerated.

In its typical implementation, the differential latch type sense amplifier consists of two cross-coupled inverters made by two pairs of transistors (MSPI-MSN1 and MSP1-MSP2) as presented in Figure 3.9 [70]. The p-net and n-net of the sense amplifier are not constantly connected to V_{DD} and 0V respectively; instead they are connected through two activation transistors (MSP3 and MSN3). The sense

amplifier is activated by setting the *SA_EN* signal to high; in the complementary situation the sense amplifier is inactive. Two access pMOS transistors MSPL and MSPR act as switches and allow only when it is needed, the connection of the sense amplifier to the bit-lines *BL* and *BLB*.

As seen in Fig. 3.9, the *NL* and *NR* nodes of the sense amplifier are connected to the bit-lines through the access pMOS transistors, and the sense amplifier is able to detect small-signal differential inputs (i.e. the bit-line voltage differences) and amplify them towards the memory output. This signal input differences can range between 30mV and 50mV, and the sense amplifier will respond with a full swing (0 to V_{DD}) signal to the output terminals. This type of sense amplifier employs positive feedback providing fast response and a straightforward design. As it is differential, it is usually applied to SRAM memories where the memory cells provide a true differential output.

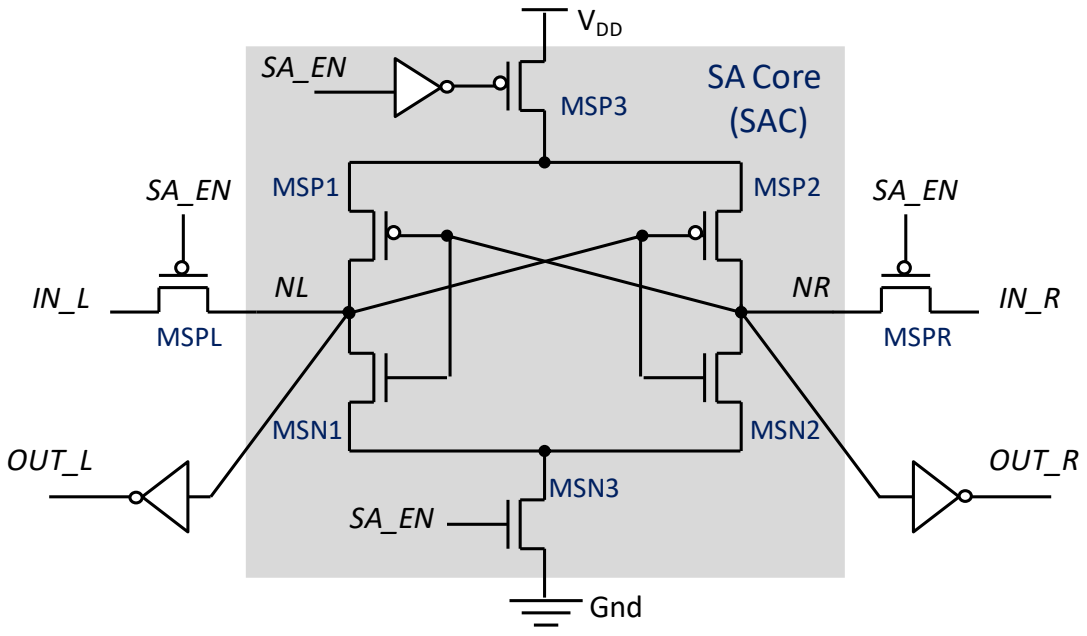

Figure 3.9 The latch type sense amplifier (SA).

The operation of the sense amplifier in Fig. 3.9 is divided into two phases. In the first phase, the *SA_EN* signal is low and the sense amplifier is inactive; the access transistors MSPL and MSPR are on so that the initial voltage difference on the bit-lines (which is generated by the activated memory cell) is passed to the internal nodes *NL* and *NR* of the sense amplifier core (SAC). Next, in the second phase the *SA_EN* signal turns to high and the sense amplifier core is isolated from the bit-lines and connected to the power supplies through the MSP3 and MSN3 transistors. The cross-coupled pair amplifies the initial voltage difference and the read value appears at the outputs *OUT_L* and *OUT_R* of the sense amplifier.

Another frequently used sense amplifier is the pMOS cross-coupled amplifier with equalization (PCCEQ) presented in Figure 3.10 [71]. It is also a differential latch type sense amplifier consisting of a two cross-coupled pMOS transistors (MSP1 and MSP2) and two nMOS transistors (MSN1 and MSN2) driven by the bit-lines *BL* and *BLB* that are attached at the sense amplifier inputs *IN_L* and *IN_R* respectively. The advantage of this SA is that it greatly reduces the DC current after amplification and latching, because it provides a nearly full supply voltage swing with positive feedback towards the outputs of the SA. Moreover, this positive feedback effect gives a high sensing speed. The sense amplifier is again activated by setting the *SA_EN* signal to high; in the complementary situation the sense amplifier is inactive. To obtain correct and fast operation, an equalization pMOS transistor MSP3 is connected between the output terminals and is turned on when the *SA_EN* signal is low to equalize the output voltage levels when the sense amplifier is inactive.

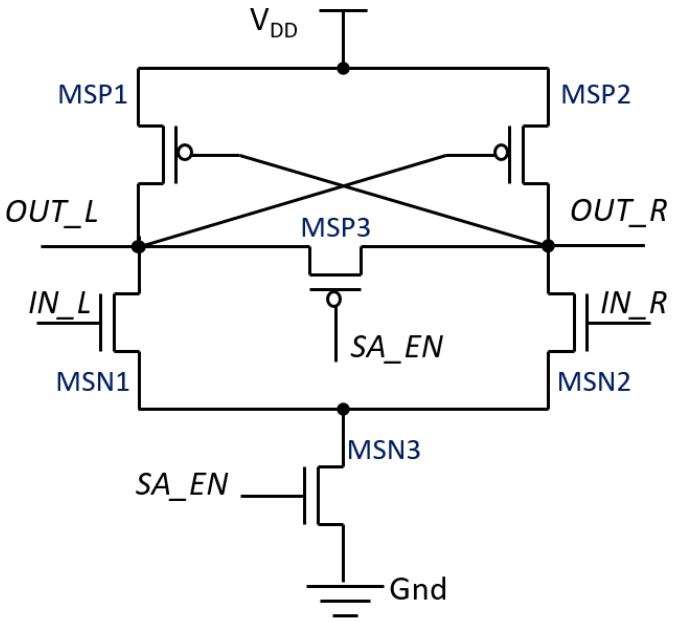

Figure 3.10 The pMOS cross-coupled sense amplifier with equalizer (PCCEQ SA).

The PCCEQ sense amplifier operates as follows. Initially the *SA_EN* signal is low and the internal nodes of the SA are equalized to an intermediate level through the activated pMOS transistor MSP3. Both bit-lines are pre-charged to V_{DD} and the nMOS transistors MSN1 and MSN2 are active. Then, one of the bit-lines starts falling to lower voltage, creating a small voltage difference, causing one of the nMOS transistors to conduct smaller current. The *SA_EN* signal is turned to high and the nMOS transistor MSN3 is activated. As the path from the source of the nMOS transistor MSN3 to the gnd is activated, the corresponding internal node of the SA (drain of the MSN1 or MSN2 respectively) that has started discharging, is quickly set to gnd. One of the pMOS transistors MSP2 or MSP1 respectively is thus activated while the opposite side of the SA is charged to V_{DD}. Thus, the cross-coupled pair amplifies the initial voltage difference and the read value appears at the outputs *OUT_L* and *OUT_R* of the sense amplifier.

3.3. Aging Effects on the SRAM Operation

SRAM occupies a large portion of silicon area in today's integrated circuits and plays a major role in their performance characteristics [72]. Thus, SRAM's reliability is a key issue for the reliable operation of modern systems. As it has been previously illustrated, in nanometer technologies the SRAM's reliability is seriously threatened by the increased process variations setting the transistors' performance away from the expected level (time-zero variability effect), as well as by aging mechanisms like the *Bias Temperature Instability* (*BTI*) [73] and *Hot Carrier Injection* (*HCI*) [74] phenomena which may lead to a significant increase of the absolute value of transistors' threshold voltage (V_t) and thus to gradual aging. Aging due to BTI or HCI is accelerated when the transistors are under excessive stress conditions (high gate-to-source voltage levels and high temperatures of operation). However, it should be noted that when a transistor faces DC (permanent) BTI stress its V_t degradation tends to become much higher with respect to AC BTI stress, since the latter involves recovery cycles that alleviate the influence on V_t (Figure 3.11) [75]. Thus, in BTI the switching activity determines the transistors' aging rate.

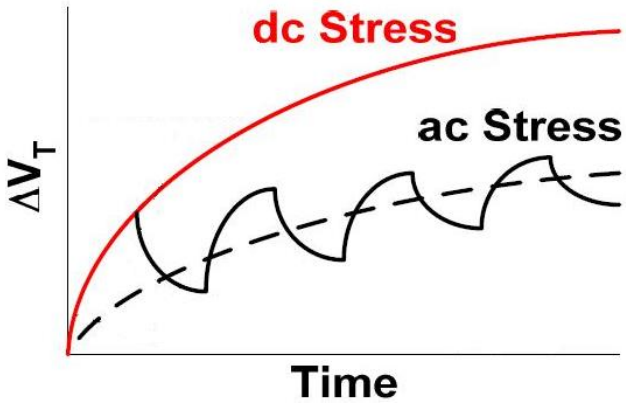

Figure 3.11 DC stress Vs AC stress over time.

Aging phenomena significantly impact the performance characteristics of SRAMs since they affect among others speed, operating voltages, memory cells' noise margins, sense amplifiers' input offset voltage and decoding delays. Excessive performance degradation due to aging in an SRAM will lead to failures generation, a phenomenon that raises the reliability as a key issue.

3.3.1. Aging effect on the SRAM Memory Cell Operation

The vulnerability of a memory cell to transistor aging has been extensively studied in literature [58], [76]-[82]. According to these studies aging substantially affects the operation of a memory cell and consequently influences the SRAM performance characteristics, like the read and hold (retention) Static Noise Margins (SNMs), the write margin, the access time and the minimum operating voltage. The SNMs are defined as the minimum noise voltage level that is capable to flip a memory cell. The write margin is the minimum voltage on the bit-lines in order to perform a transition write operation on a cell. Finally, the access time is the maximum time duration for a read operation.

Initially, we should mention that since most of the time during the SRAM operation the access transistors (transistors MCN3 and MCN4 in Figure 3.5) are off, almost no degradation is expected on these devices [14]. In addition, when a pMOS transistor in the cross-coupled pair (e.g. transistor MCP1) is under stress (Qb=low) then the nMOS transistor of the other inverter (MCN2) will be also under stress (Q=high). Obviously, in that case transistors MCP2 and MCN1 are not stressed (they are in a recovery state). Consequently, a memory cell that does not alter its memory state is expected to become increasingly skewed due to an asymmetric transistor aging. However, skew problems are also reported even if a memory cell has a more frequent switching activity. Thus, the performance characteristics of a memory cell will be degraded over time and will lead to

failures generation, putting at risk the reliable operation of the whole SRAM during its lifetime.

3.3.2. Aging effect on the SRAM Sense Amplifier Operation

Recently, the impact of transistor aging on the sense amplifier (SA) operation has been also studied [72], [83]. According to these studies, under aging conditions both the sensing delay and the input offset voltage of the sense amplifier are negatively affected (increased). The sensing delay is related to the speed performance of the sense amplifier (read response time), which is degraded. The input offset voltage is also an important performance characteristic of the sense amplifier and it is defined as the differential input voltage that results in a differential output voltage equal to zero. A non-zero input offset voltage has a negative impact since it alters (increases) the minimum bit-line voltage difference that is necessary for a successful read operation. For an ideal sense amplifier the input offset voltage is zero since the corresponding transistors are perfectly matched. However, in practice, due to local process variations, transistor mismatches are always present and various input offset voltage levels are observed even in "fresh" sense amplifiers (time-zero variability just after fabrication), which may change over time due to aging (time-dependent variability) [83]-[85].

According to the experimental results in [83], it is reported that under common and typical memory workloads aging will induce transistor degradation in such a way that the sense amplifier will become increasingly skewed, as in the case of the memory cell. This stems from the fact that under these workloads a pair of transistors (e.g. MSP1 and MSN2) is under successive stress while the complementary pair of transistors (MSP2 and MSN1) is always relaxed (in a recovery state), causing an asymmetric aging. Thus, the V_t degradation of the first pair of transistors increases over time so that the mismatches between transistors MSP1 and MSP2 as well as between MSN1 and MSN2 will continuously increase

and the same stands for the sense amplifier input offset voltage. As a result, over time, failures are expected to appear during the read operations. Designers, as it will be further presented, only recently have started working towards elimination of the presence of such failures in order to ensure the reliable operation of the SRAM at all times.

It should be noted that in both cross-coupled pairs either of the SRAM memory cell or the sense amplifier, the nMOS V_t degradation does not cancel the effects of pMOS V_t degradation and vice-versa [58], [76].

3.3.3. Aging effect on the SRAM Address Decoder Operation

As all nanometer transistors are prone to BTI and HCI aging mechanisms [43], Address Decoders are also threatened by these aging phenomena. This is due to the fact that BTI and HCI aging leads to the gradual increase of the transistors' absolute threshold voltage (V_t). In the Address Decoders, an increase of the transistors' threshold voltage is translated to increased decoding delays and thus to timing violations that may cause memory failures over time.

A few works presented in the open literature analyze the impact of aging phenomena on the transistors of Address Decoders [86]-[89] and demonstrate their strong impact on the induced delays in the word-line activation and de-activation. Both NBTI and PBTI phenomena impact the delays uniquely. That is, the NBTI phenomenon in pMOS transistors affect the de-activation delay while the PBTI phenomenon in the nMOS transistors affect the activation delay. In addition, different access patterns (or workloads) stress the transistors differently, resulting in different delays. Experimental results in [88] showed that on average the induced delays in the word-line activation and de-activation are 13% (due to PBTI) and 7.50% (due to NBTI) respectively. Hence, the BTI and HCI impact on the word-line activation appears to be more significant with respect to its de-activation. This can be justified as follows. Normally, the NBTI induced delay is

more significant than caused by PBTI for an inverter [90]. However, as all the NMOS transistors of the pull-down network are connected in series, their impact accumulate, resulting in higher impact on the word-line activation [88].

Since delay faults on the Address Decoders are considered to lead to read or write failures on the memory cells of the SRAM the aging impact constitutes a serious concern for the reduction of the memory's reliability.

3.4. Circuit Testing Techniques

A key requirement for obtaining reliable electronic systems is the ability to determine that the systems are error-free [91]. Although electronic systems contain usually both hardware and software, the interest of this thesis is on hardware testing.

Hardware testing is a process to detect failures primarily due to manufacturing defects as well as aging, environment effects and others. It can be performed only after the design is implemented on silicon by applying appropriate stimuli and checking the responses. Generation of such stimuli together with calculation of the expected response is called test pattern generation. Test patterns are in practice generated by an *automatic test pattern generation* tool (ATPG) and typically applied to the circuit using *automatic test equipment* (ATE).

With recent advances in semiconductor manufacturing technology, the production and usage of *very-large-scale integration* (VLSI) circuits has run into a variety of testing challenges due to the increased potential for defects, as well as the difficulty of detecting the faults produced by those defects. Traditional test techniques that use ATPG software to target single faults for digital circuit testing have become quite expensive and require an important amount of time to apply the test

patterns, while can no longer provide sufficiently high fault coverage for nanometer technology designs.

3.4.1. Built-In Self-Test (BIST)

One approach to alleviate these testing problems is to incorporate *built-in self-test* (BIST) features into a digital circuit at the design stage [8], [14], [92]-[94]. With logic BIST, circuits that generate test patterns and analyze the output responses of the functional circuitry are embedded in the chip or elsewhere on the same board where the chip resides. The BIST approach allows the tests to be performed under actual clock speeds, decreasing this way the test time.

There are two general categories of BIST techniques for testing random logic: (1) online and (2) off line BIST. General categories of logic BIST techniques is shown in Figure 3.12 [8].

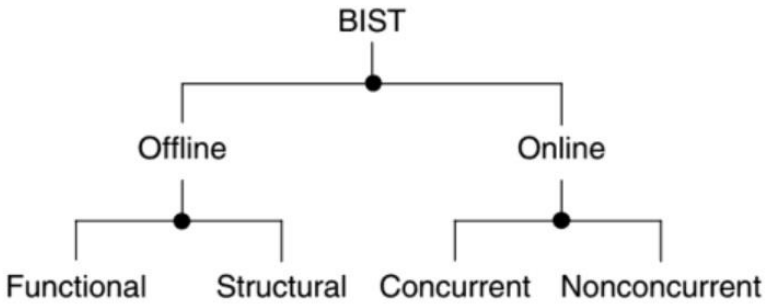

Figure 3.12 Logic BIST Techniques.

Online BIST

It is performed when the functional circuitry is in normal operational mode. It can be done either *concurrently* or *nonconcurrently*.

In **concurrent online BIST**, testing is conducted simultaneously during normal mode of operation. The functional circuitry is usually implemented using coding

techniques or exploiting duplication and comparison features [8]. When an *intermittent* or *transient* error is detected, the system will correct the error on the spot, e.g. rollback to its previously stored system states and repeat the operation, or generate an interrupt signal for repeated failures.

In **nonconcurrent online BIST**, testing is performed when the functional circuitry is in idle mode. This is often accomplished by executing diagnosis software routines (macrocode) or diagnosis firmware routines (microcode) [8]. The test process can be interrupted at any time so that normal operation can resume.

Offline BIST

It is performed when the functional circuitry is not in normal mode. This technique does not detect any *real-time errors* but is widely used in the industry for testing the functional circuitry at the system, board, or chip level to ensure the quality of the product.

Functional offline BIST performs a test based in the functional specification of the circuitry and often employs a functional or high-level fault model. Normally such a test is implemented as diagnostic software or firmware [95].

Structural offline BIST performs a test based on the structure of the circuit under test. There are two general classes of structural offline BIST techniques: (1) **external BIST**, in which test pattern generation and output response analysis are done by circuitry that is separate from the functional circuitry being tested and (2) **internal BIST**, in which the functional storage elements are converted into test pattern generators and output response analyzers. Such techniques are often used for board-level and system-level self-test. The most common BIST schemes assume that the functional storage elements of the circuit are converted into a scan chain or multiple scan chains for combinational circuit testing. Such schemes are much more common than those that involve sequential circuit testing.

Figure 3.13 shows a typical logic BIST system using the *structural offline* BIST technique. The test pattern generator (TPG) automatically generates test patterns for application to the inputs of the circuit under test (CUT). The output response analyzer (ORA) automatically compacts the output responses of the CUT into a *signature*. Specific BIST timing control signals including scan enable signals and clocks, are generated by the logic **BIST controller** for coordinating the BIST operation among the TPG, CUT and ORA. The logic BIST Controller provides a pass/fail indication once the BIST operation is complete. It includes comparison logic to compare the *final signature* with an embedded *golden signature* and often comprises diagnostic logic for fault diagnosis. As compaction is commonly used for output response analysis, it is required that for storage elements in the TPG, CUT and ORA are initialized to known states prior to self-test and no unknown values are allowed to propagate from the CUT to the ORA. In other words, the CUT must comply with additional *BIST-specific design rules*.

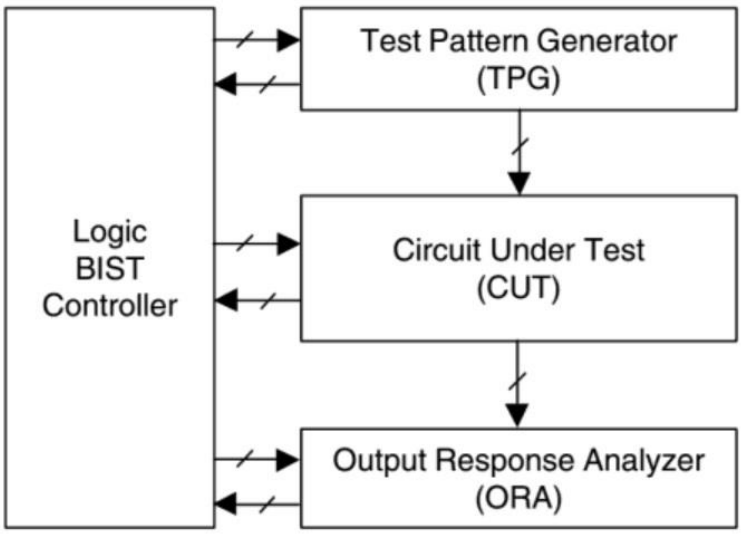

Figure 3.13 A Typical Logic BIST System.

There is a number of advantages when using the structural offline BIST technique rather than conventional scan:

- BIST can be made to effectively test and report the existence of faults on the board or system and provide diagnostic information as required; it is always available to run the test and does not require the presence of an external tester.

- Because BIST implements most of the tester functions on-chip, it is possible to conduct testing with fewer test pins per chip, which enables an increase in the number of chips that can be tested in parallel.

- Test costs are reduced due to reduced test time, tester memory requirements, or tester investment costs, as most of the tester functions reside on-chip itself.

However, there are also disadvantages associated with this approach. More stringent BIST-specific design rules are required to deal with unknown sources originating from analog blocks, memories, non-scan storage elements, asynchronous set/reset signals, tristate logic and multiple-cycle paths to name a few. Also, because pseudo-random patterns are commonly used for BIST Pattern generation, additional test points including control points and observation points have to be added to improve the circuit's fault coverage.

While *BIST-specific design rules* are required and the BIST fault coverage may be lower than that using scan, BIST does not eliminate the expensive process of software test pattern generation and the huge storage requirements to store the output responses for comparison. More importantly, a circuit embedded with BIST circuitry can be easily tested after being integrated into a system. Periodic in-system self-test, even using test patterns with less than perfect fault coverage, can diagnose problems down to the level where the BIST circuitry is embedded. This allows system repair to become trivial and economical [95].

3.4.2. I_{DDQ} Testing

While test generation methods such as BIST, focus on driving specific voltage values to circuit nodes and observing the voltage levels at the observable points such as the primary outputs, there are also techniques which are based on current measurement. These techniques are commonly referred as **I_{DDQ} test techniques** which target the current drawn by the CUT. The I_{DDQ} technique is based on measuring the quiescent current and can detect some of the faults which are not detectable with other testing techniques. Theoretically, the current monitoring technique is one of the most effective techniques for the discrimination of defective-free from defective circuits.

It relies on the observation that a defect-free CMOS integrated circuit has a quite small quiescent current (I_{DDQ}) in the steady state (due to transistor current leakages), with respect to a defective one. In case that the measured quiescent current is much higher than the expected level (taking into account possible process or temperature variations), then the circuit is considered outlier and is characterized as defective. The additional current is due to possible defects. Thus, the power consumption is the criterion to discriminate defective ICs. I_{DDQ} testing can be also used for reliability estimation [96].

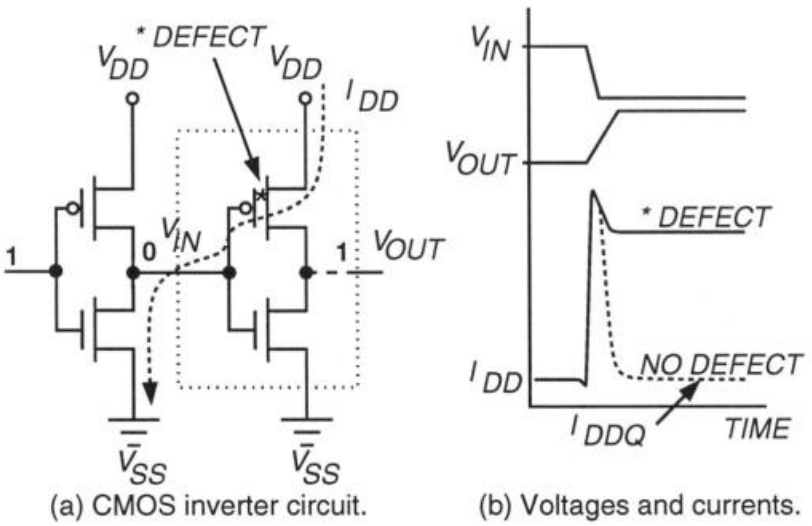

Figure 3.14 Basic principle of I_{DDQ} Testing.

Figure 3.14(a) shows a CMOS inverter, with a * indicating a defect in the pMOS transistor that causes its input impedance to drop from infinity to a finite value. The DC current flows in steady state along the path indicated by the arrow, and this elevates the steady state current, since current can still flow through the defective pMOS transistor [97]. Figure 3.14(b) shows the input and output voltages, and the current I_{DD} that flows through the transistors from the power supply. After switching is completed, this current corresponds to the *quiescent* current and is called I_{DDQ}. In the case of a defect-free circuit, the I_{DDQ} falls to a negligible value, whereas in the case of a defective circuit, it remains elevated long after the switching is over. Such faults are detected by measuring the I_{DDQ} at the time instant illustrated by the arrow.

The advantages of this technique is that it allows checking the chip for many possible faults with one measurement and that it may detect faults that are not found by conventional test vectors. However, the main disadvantage of this technique is the very slow testing process, which makes testing very expensive.

3.5. State of the Art in SRAMs Aging Monitoring

To sum up, as previously described, the performance of an SRAM memory cell is affected by BTI and HCI since these aging phenomena influence both the memory speed and the noise margins [58], [72]-[79], [83], [98] [99]. Moreover, recent research indicates that transistor aging in SRAM sense amplifiers also results in gradual speed performance degradation [72] and input offset voltage development [83], [98]. Additionally, the above mentioned aging mechanisms are shown to affect the Address Decoders [86]-[89] resulting in increased decoding delays (*WL* activation). Thus, as the performance of the SRAM's cells, sense amplifiers and decoders are continuously degraded, failures are expected to occur during memory operation. According to the above, it is mandatory to develop aging monitoring techniques for these blocks in order to obtain the ability to predict failures that

may occur during memory lifetime and repair the circuit accordingly to ensure aging tolerance and consequently its reliable operation.

Traditionally, in order to effectively deal with failure generation in the SRAMs under aging phenomena, designers use guardbands; extra margins are considered in the period of the clock signal used, as it is illustrated in Figure 3.15 in order to guarantee that the memory will operate correctly during its lifetime. Therefore, by design, the circuit operates at a lower frequency than it could operate in the worst case, which is a major disadvantage of this approach. It is obvious that such techniques negatively impact the performance of a circuit since they significantly affect speed.

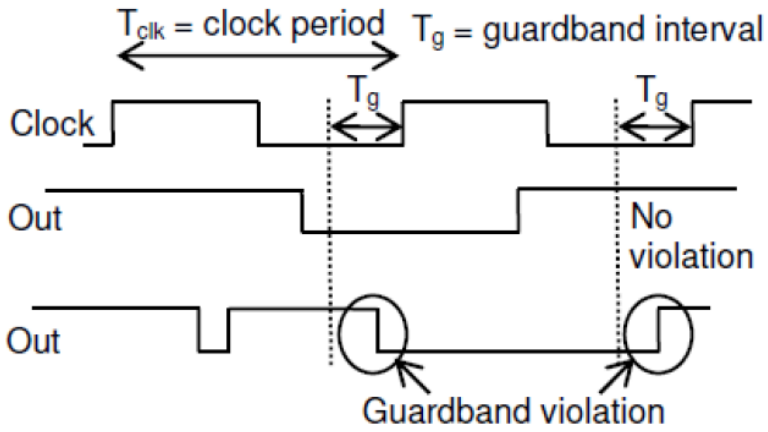

Figure 3.15 Guardband interval logic.

Recently, various techniques have been presented in the open literature for the monitoring of aging related performance degradation in SRAM memories. However, some of the cons related to these techniques stem from the fact that they cannot locate the defective parts thus, repairing operations are not feasible or they require a lot of additional circuitry leading to an excessive cost in the silicon area and design complexity.

3.5.1. Aging monitoring techniques in SRAM Memories

The I_{DDQ} current has been used in [76] for the characterization of NBTI induced SRAM performance degradation. Under NBTI influence the leakage current of the memory cells is reduced exponentially due to the V_t degradation. Thus, reductions to the expected leakage current of the SRAM indicate an aged memory. The presence of a power gating scheme is assumed and a current mirror based I_{DDQ} monitor is exploited as presented in Figure 3.16. This approach measures the accumulative leakage current of the whole memory array and thus the location of aged cells is not feasible.

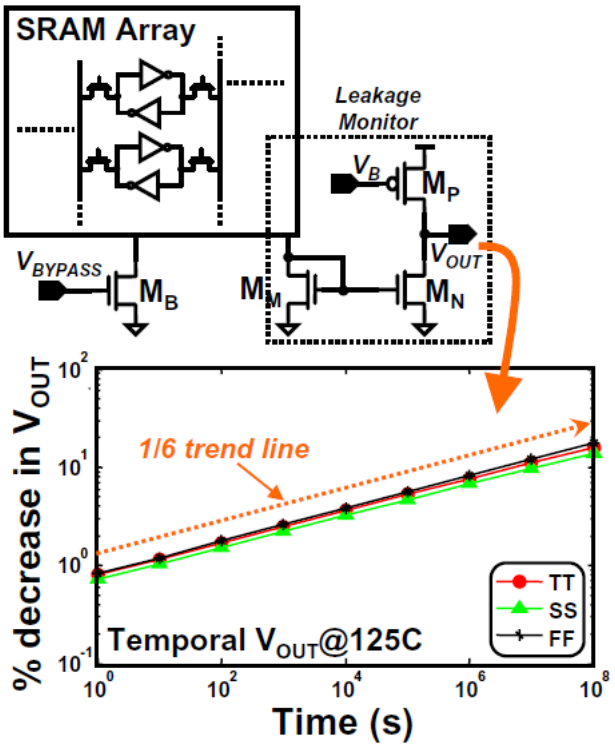

Figure 3.16 I_{DDQ} monitoring circuit for an SRAM array. Monitor output V_{OUT} can be used as a signature of NBTI degradation.

Special structures that are based on analog blocks (voltage comparators, phase comparators e.t.c.) are exploited in [77] for the stability characterization of an SRAM under the influence of NBTI (among other sources of vulnerability). The

authors mention that measuring the word-line pulse-widths calibrates out any timing uncertainty introduced by SRAM peripheral circuits, thus allowing characterization of the fundamental variability of the SRAM bitcells. The proposed method is used to identify sources of variability in dynamic stability by observing deviations from expected correlations between dynamic stability and static margins. However, the design complexity is too high.

In [100] an on-chip NBTI aging sensor is proposed for the off-line monitoring of the writing operations into the SRAM memory cells in order to detect degradation (Figure 3.17). A dedicated sensor (which is based on a sense amplifier topology) exists for every column in the SRAM array. The technique assumes the use of power gating schemes and the existence of a non-aged reference column of memory cells. Off-line write operations are performed to the memory cells and the corresponding reference cells. After each write operation the virtual V_{DD} node voltage levels of the memory cell under monitoring as well as the reference cell are compared in order to detect aged memory cells.

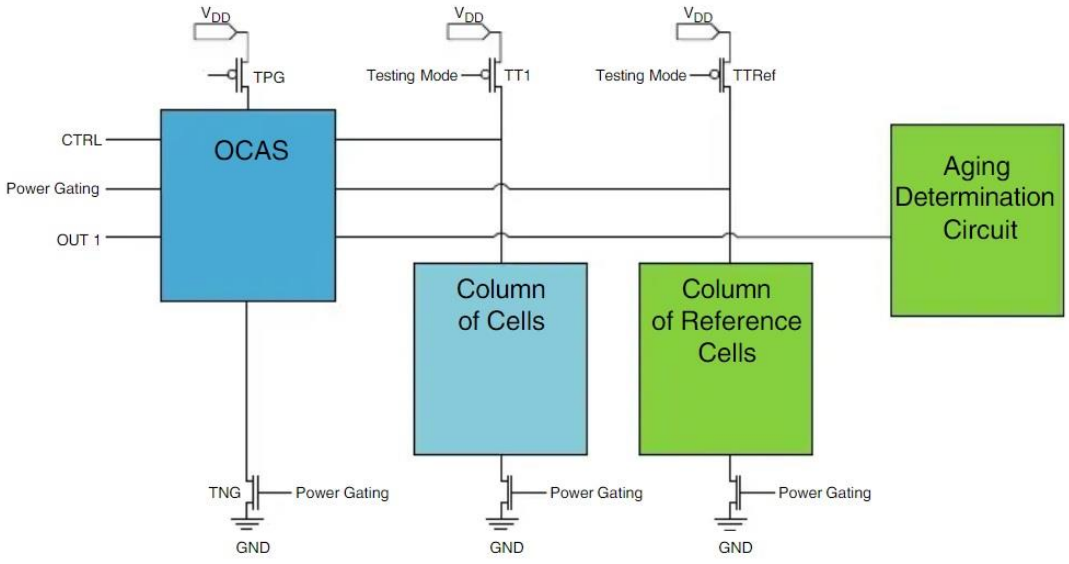

Figure 3.17 General block diagram of the hardware-based approach connected to one SRAM cell column.

Aging monitoring solutions based on embedded ring oscillators [101], [102] and voltage controlled oscillators [43] have been proposed. Also in that case the degradation of the whole SRAM is inspected and repair operations are not feasible. More specifically, in [101] a D-flip-flop uses a fresh reference ring oscillator (ROSC) to sample the output of an identically stressed ROSC. Each stressed oscillator is paired with its identical, fresh reference during measurements, and its frequency degradation is monitored with a beat frequency detection circuit. Small frequency shifts induced by circuit aging are magnified and aging is detected. In [102] an embedded SRAM ring oscillator (ESRO) based structure to characterize the BTI degradation is used. The unit-stage of ESRO, is composed of a blanked-cell (for control and stress) and an active-cell (for forming an inverter). The major function of the blanked-cell is to provide the stress voltage to the input of the active-cell inverter. The beat frequency is again extracted to detect aging.

A write margin degradation monitoring technique for SRAMs is presented in [103]. According to this technique, a sequence of read/write operations is performed on a cell while the word-line voltage is successively lowered. The maximum tolerated word-line voltage drop by the memory cell is related to the transistors' threshold voltage and it is exploited for degradation detection. This solution requires the generation of multiple voltage levels for the application of the aging monitoring procedure.

A bit-flipping technique has been also proposed in [104]. The authors proposed a *self-controlled* bit-flipping (SCF) technique which uses a flip-on-access (FOA) mechanism to avoid an interruption of normal cache operations. The SCF technique starts with the selection of a flag bit used to control the bit-flipping sequence. A flag bit is implicitly selected out of the data inputs of cache array (e.g. 1 bit out of 32 bits) through an offline selection algorithm. Ideally, the flag bit should be the bit position which has a 50% probability of storing a value of '1' or '0'. When the flag bit is '1', the value of all other bits is flipped before the write

operation and flipped back during the read operation. When the flag bit is '0', the value of all other bits remains the same during write and read operations. Once the flag bit is selected, it is connected to the Flip*in* and Flip*out* inputs of the FOA circuitry as illustrated in Figure 3.18. Since the flag bit is used to invert the content of the memory array during the read operation, its value should be preserved throughout the process. Although this approach is used for the alleviation of the aging effects on memory cells by reducing the aging rate, it does not eliminate the problem.

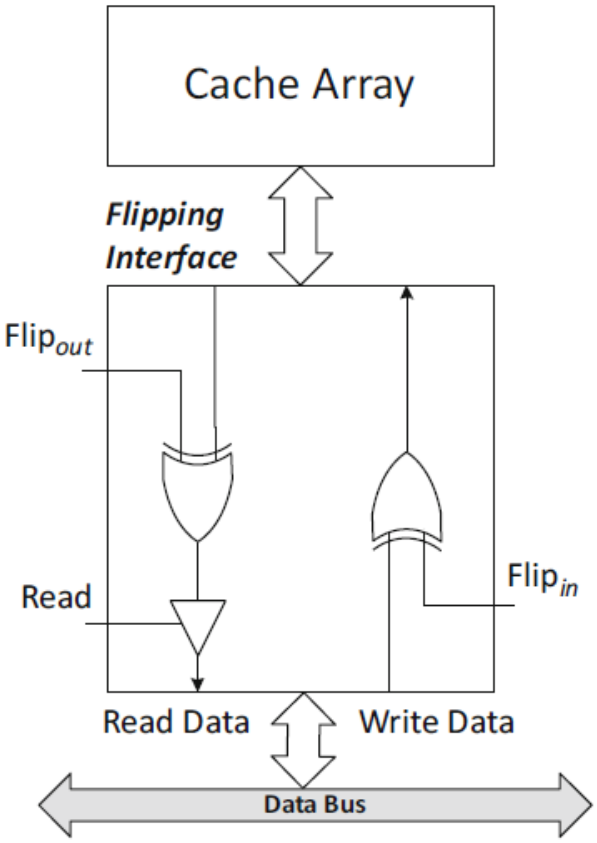

Figure 3.18 An FOA circuitry.

Error correcting codes (ECC) can be used to avoid aging related failures during the SRAM operation [55], [105]. However, in large memories and as the number of aged cells increases in time, it is necessary to exploit redundancy in order to

maintain reliability. Thus, ECC alone is not always a sufficient solution and monitoring schemes for the location of aged cells near failure are necessary.

An interesting memory cell BTI degradation estimation method is presented in [58]. Each bit-line slice in the memory array is modified as it is shown in Figure 3.19.

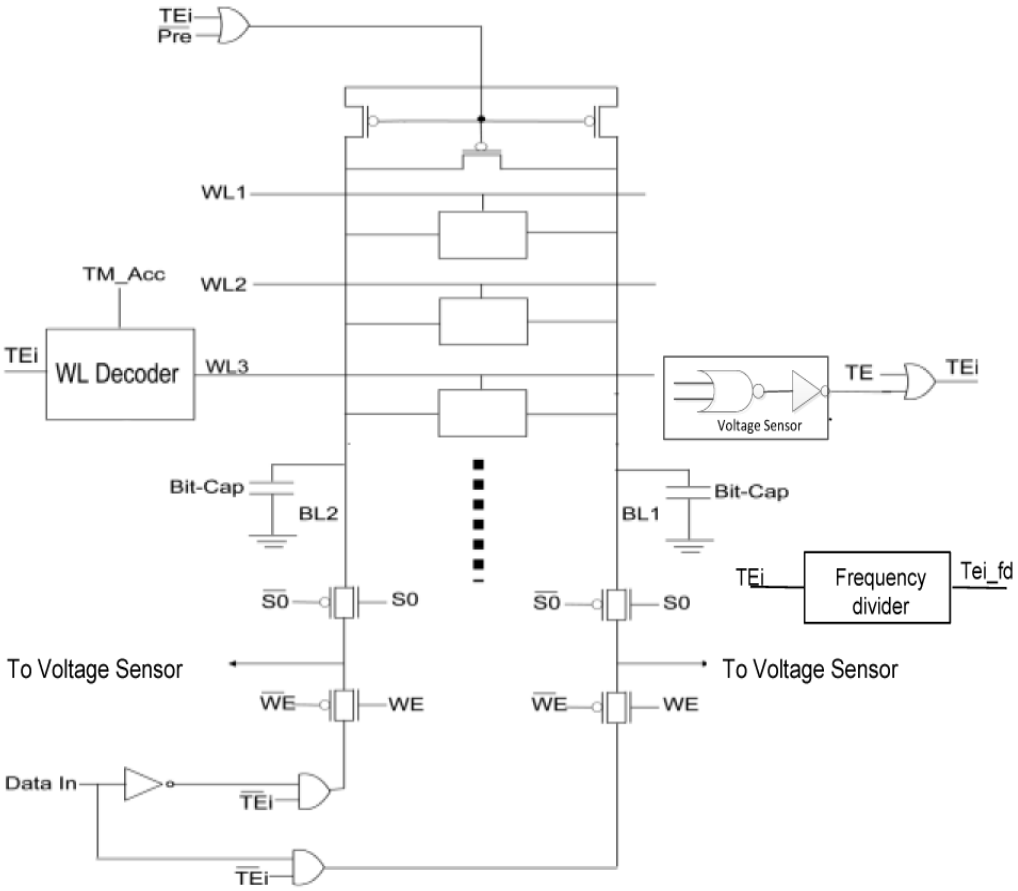

Figure 3.19 Bit-line slice modifications for the support of the BTI monitoring technique in [58].

According to this method, in NBTI testing the current strength of the pMOS transistors in the cell is used for the detection of aged devices. Sequences of successive pseudowrite operations are applied on a selected memory cell, where unlike the normal situation both bit-lines are forced to the ground through the modified write circuitry. Once the bit-lines are pulled-down the voltage sensing

circuit deactivates the write circuitry and the pMOS transistors start to charge the bit-lines. Obviously, the charging current is a function of the threshold voltage and it is influenced by NBTI so that NBTI aging affects the charging time. When the bit-line voltage exceeds a certain level the write circuit is re-activated forcing again the bit-lines to the ground. Consequently, an oscillation is generated which is related to the cell pMOS transistor strength. Transistor degradation results in oscillation frequency degradation. The generated oscillation frequency is digitized for evaluation.

Similarly, in PBTI testing the current strength of the nMOS transistors in the cell is used for the detection of aged devices. This time sequences of pseudoread operations are performed on a selected memory cell, where unlike the normal mode both bit-lines are forced to V_{DD} through the modified precharge circuitry. After the charging of the bit-lines the precharge circuitry is deactivated. Thus, the cell's nMOS tranistors start to discharge the bit-lines. When the bit-line voltage turns lower than a certain level the precharge circuit is re-activated forcing again the bit-lines to V_{DD}. As earlier, an oscillation is generated, where the frequency depends on the strength of the discharging nMOS transistors. Once again transistor degradation results to oscillation frequency degradation.

The above BTI monitoring method is periodically applied in the field of operation and detected over-aged memory cells (near failure) are replaced by spare cells using the embedded repair mechanisms in the memory. However, this approach is seriously affected by process and temperature variations. Moreover, note that the memory control logic must be redesigned in order to be able to support the proposed measurement process.

Finally, an approach for possible mitigation of BTI aging relies on the recovery property of this mechanism. For the case of BTI aging mitigation at a memory

cell, proper circuitry is included in the memory to periodically flip the stored data in it in order to reduce transistor degradation [106].

3.5.2. Aging monitoring techniques in SRAM Sense Amplifiers

From the above analysis, it is obvious that a lot of researchers have focused their work on the aging monitoring of SRAM memory cells but the aging monitoring of SRAM sense amplifiers has not been extensively investigated. A technique for the characterization of SRAM sense amplifier input offset voltage, for yield prediction, is presented in [107].

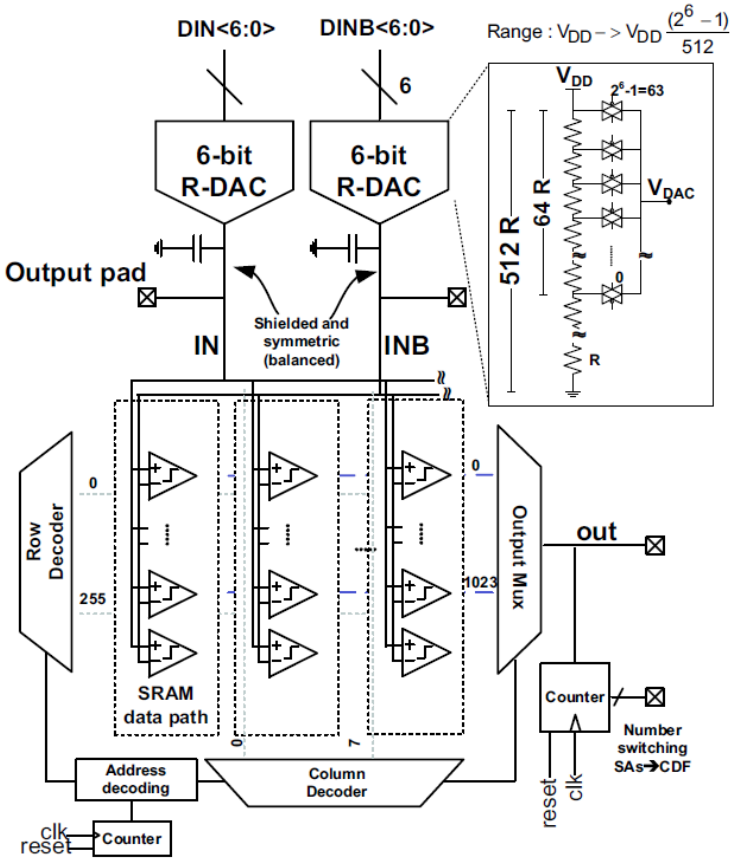

Figure 3.20 Implementation of the SA input offset monitor for the support of SRAM yield prediction in [107].

In this work, as shown in Figure 3.20, two resistor string 6-bit digital-to-analog converters (R-DACs) are used to drive the inputs of the sense amplifiers in a memory array. These R-DACs generate various bit-line voltage differences that are necessary for the sense amplifier input offset voltage estimation. Each sense amplifier is selected for test after the other with the help of a counter, while a second counter is exploited for the final yield evaluation. The silicon area cost and the design effort related to this characterization technique are quite high.

Another work in [108] proposes a performance sensor for memory cells and sense amplifiers aging detection. However, the influence of aging on the sense amplifier input offset voltage is not considered in this work. Finally, for the mitigation of BTI aging effect on a sense amplifier, relying on the recovery property of this mechanism, the research team in [83] proposes the input switching sense amplifier in order to attain a balanced workload and consequently succeed reduced transistor degradation. Due to the statistical nature of these solutions, BTI aging effects are alleviated but not eliminated. Consequently, aging prediction and repair schemes are mandatory for the long-term and reliable memory operation.

3.5.3. Aging monitoring techniques in SRAM Address Decoders

When the Address Decoders are affected by BTI or HCI they may also lead to significant performance degradation of the SRAM. Delay faults in the Decoder logic constitute an important factor why buyers return the purchased products [109], [110]. A few works analyze the impact of aging phenomena on the transistors of Address Decoders [86]-[89] but aging detection techniques have not been extensively investigated until today.

A couple of hardware techniques have been proposed for the detection of delay faults on SRAM Address Decoders during run-time, such as [111], [112], without providing though the ability to detect the faults before failure occurrence or

mitigate them. A most recent work in [113] proposes a hardware-based mitigation scheme for a special type of memory Address Decoder logic (presented in Figure 3.21) to reduce the impact of aging on it. The scheme is based on adapting the Decoder's workload during idle cycles by stressing the short paths and putting long paths into relaxation. As a result, the impact of aging on the address decoder's setup margin is reduced, aiming to prolong its lifetime.

In this work, a wordline decoder is implemented with the addition of a pre-decoding and post-decoding stage to activate and de-activate the wordlines. According to this research, in case a memory is idle, the input address of the address decoder is kept unchanged from the last valid applied input address. Hence, in case an application frequently keeps pre-decoder outputs with long paths activated during idle cycles the stress on them will be high and thus, they will significantly contribute to the decoder's degradation. To mitigate this aging effect, [113] proposes to utilize these idle cycles by applying addresses that activate pre-decoder outputs with short paths in order to ensure during idle cycles long paths are put into relaxation. The idea is based on the observation that the last output of the pre-decoder in Figure 3.21 has the lowest path delays and, therefore, is considered as a good candidate for mitigation.

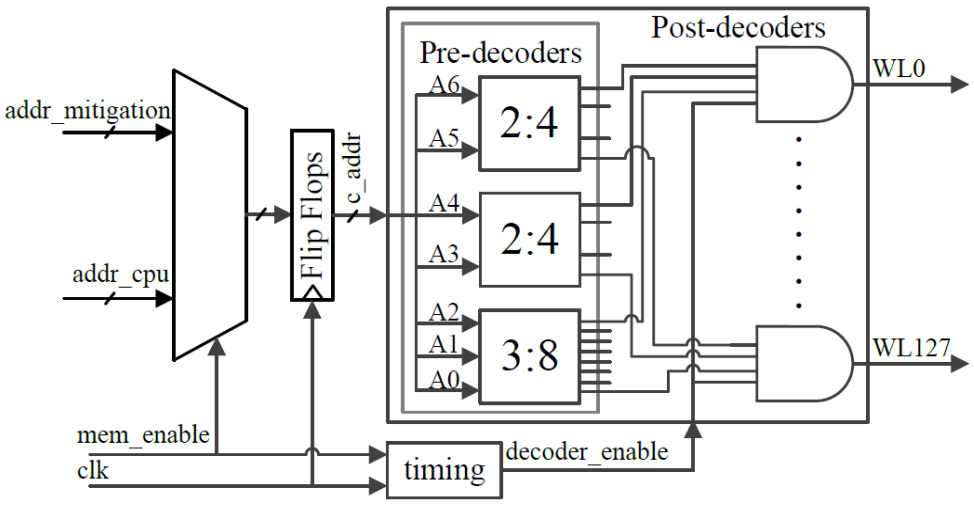

Figure 3.21 Address decoder mitigation scheme proposed in [113].

The proposed mitigation scheme is implemented with an additional multiplexer that is placed before the address flip-flops as presented in Figure 3.21. This multiplexer is used to select between the original address coming from the host (e.g., a CPU) and a mitigation address that stresses the short paths (and, hence, puts the long paths into relaxation). All bits of this mitigation address are high (hardwired), as this ensures that the last outputs of all pre-decoders are activated. The select signal of the multiplexer is connected to the memory's enable signal, so the mitigation address is applied to the address decoder during idle cycles.

This scheme suggests the implementation of only surrounding logic without any modification on the Address Decoder. However, it statically reduces the impact of aging but does not eliminate it while it is applicable to a specific type of Decoder.

CHAPTER 4

PROPOSED AGING MONITORING TECHNIQUES FOR MEMORY CELLS & SENSE AMPLIFIERS

In this Chapter we present self-healing techniques that focus on SRAM aging prediction for the early diagnosis of the memory status in order to react and maintain its reliable operation.

The proposed monitoring schemes address individually the aging monitoring and repairing of the SRAM Sense Amplifiers and Memory Cells with the use of a Differential Ring Oscillator. Next, a unified approach for the aging monitoring and repairing of both memory cells and sense amplifiers is presented. We also discuss the application of the proposed scheme during the manufacturing testing operations and finally, the simulation results for the validation of the proposed techniques are presented.

4.1. Aging Monitoring for SRAM Memory Cells

A continuous aging related (due to BTI or HCI phenomena) degradation of a memory cell's performance, as previously discussed, will significantly affect the SRAM's reliability, as it will eventually lead to failures in its operation. Thus, aging related degradation monitoring techniques for SRAM cells need to be developed, which will provide the ability of self-healing to the SRAM by predicting upcoming failures and reacting early in order to replace failing cells by exploiting existing repair mechanisms. An on-line aging monitoring technique for SRAM memory cells is presented next, that is based on the duty cycle of the signal provided by a low cost differential ring oscillator.

4.1.1. The monitoring circuitry

The proposed self-healing technique is a periodic on-line BTI and HCI monitoring scheme for detecting aged memory cells in an SRAM array. The idea is based on a small Differential Ring Oscillator (DRO), which is embedded in the memory array so that during the monitoring phase an SRAM memory cell plays the role of "active" load at its output. A dedicated DRO is placed at every bit-line in the memory array and drives an activated memory cell of this bit-line during the monitoring mode of operation. The duty cycle of the DRO signal is used for the discrimination of aged cells since according to our observations aging affects this characteristic.

In Figure 4.1 a bit-line slice (column) of an SRAM memory array along with the sense amplifier (SA) is presented. A bit-line access transistor pair M1-M2 (one per bit-line *BL* and bit-line bar *BLB* respectively) permits the access to the memory array, by exploiting the signal *CSEL* (column select).

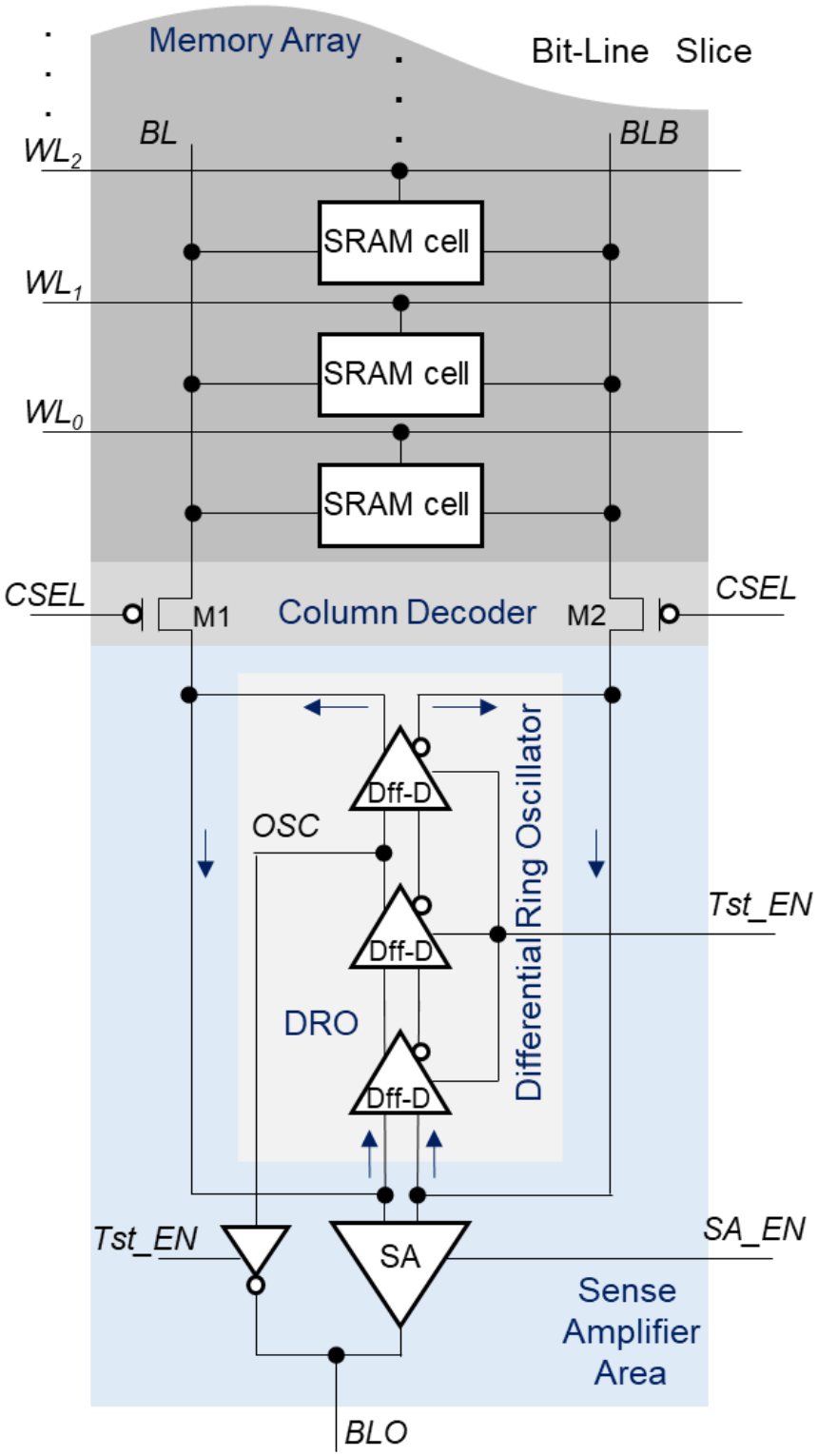

Figure 4.1 Monitoring circuitry for testing SRAM memory cells.

The DRO is placed in the area of the sense amplifier. It consists of three tri-stated differential delay elements (Dff-D cells), one of which feeds the pair of the bit-lines (*BL* and *BLB*). The *Tst_EN* (test enable) signal is used to activate (at logic high) the Dff-D cells in the monitoring mode of operation or deactivate them (at logic low) in the normal mode of operation. In the normal mode the outputs of the Dff-D cells are floating, while in the monitoring mode, an oscillation is generated in the Dff-D chain. As it will be further discussed, the duty cycle of the oscillation signal generated by the DRO, depends on the aging of the accessed SRAM cell in the corresponding bit-line.

The monitoring mode of operation in order to determine the cells' status is as follows. A desired word-line (*WL*) is activated (at logic high) in the array, while the bit-line precharge circuitry remains inactive (not shown in Figure 4.1 for simplicity). In addition, the desired bit-line pairs *BL* and *BLB* (columns) are accessed by setting the pertinent *CSEL* signals to low. Contrary to the normal operation, in monitoring mode neither the sense amplifier (SA) nor the write circuitry that exists in the sense amplifier area (this circuitry is also not shown in Figure 4.1 for simplicity) are activated. Thus, the *SA_EN* (SA enable) signal is set to low. Instead, the DRO is activated (*Tst_EN* is set to high) and the top Dff-D cell feeds the selected and activated SRAM cell in the array, which plays the role of an active load. During the oscillation phase, each time the differential signal at the output of this Dff-D cell changes state, the corresponding SRAM cell is written to the complementary state with respect to its current state.

A single-ended signal *OSC* is generated by the DRO, which is multiplexed with the readout signal of the sense amplifier (SA) through the use of a tri-state inverter. The *Tst_EN* signal controls the tri-state inverter. It should be noted that the SA is also a tri-state block controlled by the signal *SA_EN*. The signals *Tst_EN* and *SA_EN* are never activated simultaneously. The final signal (*BLO*) is propagated outside the memory array and its duty cycle is measured by an embedded circuit

for the final evaluation of the cell. As the duty cycle is used for the discrimination of the aged cells, there is no specific requirement for the period of the DRO's operation. According to the presented scheme, the *Tst_EN* signal, which is common for all bit-line slices, is the only extra signal used for the support of the monitoring operation, with respect to a typical SRAM.

4.1.2. The Differential Ring Oscillator (DRO)

The differential ring oscillator consists of three differential delay elements (Dff-D cells). A Dff-D cell is a tri-stated cross-coupled dual rail structure, as it is illustrated in Figure 4.2 that is based on the Differential Cascode Voltage-Switch (DCVS) logic design with the addition of two activation transistors, one nMOS (MDN3) and one pMOS (MDP3).

The activation transistors MDN3 and MDP3 are controlled by the *Tst_EN* signal and its complementary signal respectively. When the *Tst_EN* signal is low the Dff-D cells are inactive, disconnected from the power supplies, so that their nodes are floating. This is the normal mode of operation, where the SRAM operates most of the time since aging monitoring is a small duration periodic activity. It should be noted that as most of the time the SRAM is not operating in the monitoring mode, the Dff-D cells remain mostly inactive. Thus, a reduced performance degradation due to aging is expected on these cells and subsequently on the DRO.

In the monitoring mode all DROs of the memory array are activated in parallel. Thus, local power consumption is expected in the pertinent area and interference phenomena between adjacent DROs may be observed. In case that it is essential to alleviate these issues, a second test enable signal (or a third one) can be used so that by activating the DROs alternately the power consumption and the interference will be reduced. This is an easy to implement solution, as the routing

of these signals is not expected to pose serious constraints since the addition of two extra signals at the area of the SAs will be the only requirement.

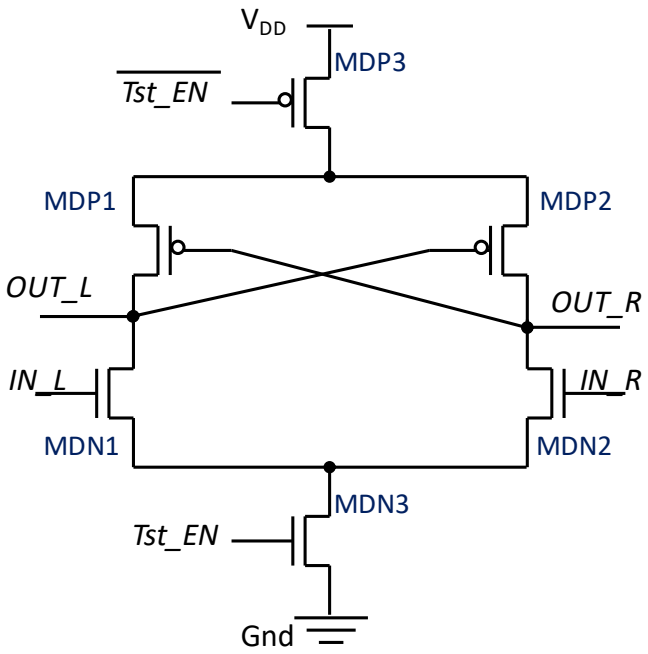

Figure 4.2 The Differential Delay Element (Dff-D cell).

4.1.3. Failure Prediction Methodology

The general memory array architecture of an SRAM is presented in Figure 4.3. During the monitoring mode of operation, a Counter generates all possible addresses for the activation of each word-line and the generated addresses feed the Row Decoder through a Multiplexer (MUX) activating a single word-line each time. The generated oscillation signals of all DROs are propagated in parallel through the corresponding *BLOi* signal lines and are transferred to a Digitizer block for the evaluation process.

As mentioned earlier, the duty cycle of the *BLO* signal is exploited for the monitoring of the aging influence on memory cells. In more detail, for every selected bit-line (by the column decoder), the corresponding activated (by the row decoder) SRAM cell plays the role of an "active" load at the output of the corresponding Dff-D cell of the ring oscillator. The Dff-D cell writes the SRAM cell

twice in every oscillation cycle. The aging of the SRAM cell alters the duty cycle of signal *OSC*/*BLO*. Thus, by measuring alterations on the duty cycle generated by the DRO, aged SRAM cells can be discriminated. The greater the alteration, the greater is the cell aging.

According to the discussion in Chapter 3.3, an SRAM cell affected by aging turns to be skewed, that is the two inverters have not equal driving strengths, and as a consequence the duty cycle of the *OSC* signal is also skewed.

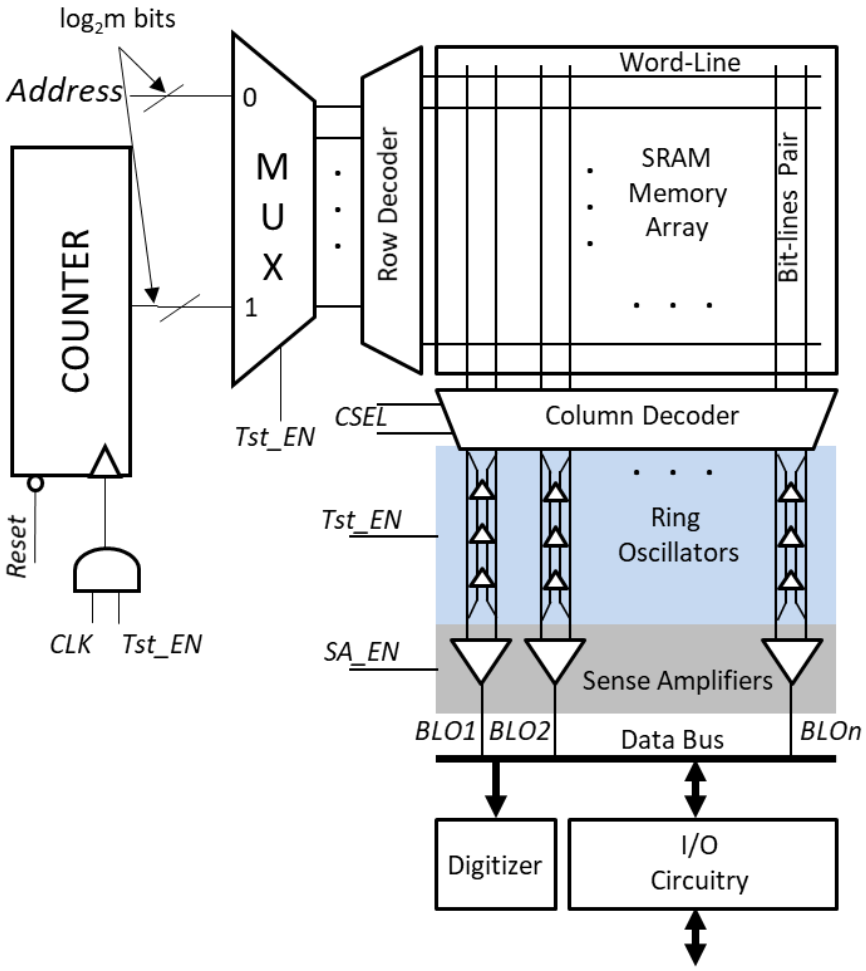

Figure 4.3 The general SRAM architecture.

In the Digitizer block, the duty cycle of the signal is measured and digitized in a serial manner (each time a single cell is evaluated). The result is compared with a

reference duty cycle value and in case that the duty cycle skew is above a reference value, an excess aging degradation is detected. The reference duty cycle is stored in a ROM, like those typically used for memory repair operations. After the evaluation of the current set of cells in the activated word-line, the address of the column decoder changes to select the next set and the procedure is repeated until the evaluation of all SRAM cells in the word-line. Upon completion of these steps, the next word-line is selected and so on until the evaluation of all cells in the memory array.

The reference duty cycle value corresponds to a memory cell that although it is aged enough it does not generate failures during the SRAM operation. Furthermore, a memory cell with a duty cycle just outside this reference limit also does not generate failures. However, further aging will soon result in the generation of failures and the degradation of the reliability levels. Thus, failure prediction can be achieved, through the application of the proposed method, in order to react early and self-heal the SRAM.

Periodic aging monitoring is a quite effective solution since aging degradation is a gradual phenomenon. Given that a memory cell has been seriously affected by aging, so that its operation is near failure generation, then after the detection of this status the SRAM can be properly repaired in order to ensure aging tolerance and retain a reliable memory (as it will be further discussed).

The monitoring procedure can be periodically applied, in the field of operation, at the system start-up or during idle times, aiming to increase the effective lifetime of the SRAM. However, in case that the reliability standards are too high, aging monitoring can be applied at the end of predetermined time intervals, like the refresh operations in DRAMs but at definitely much greater time periods. This kind of periodic activity will not in practice affect the performance of the system where the SRAM is embedded. In order to ensure that the data already contained

in the SRAM cells are not lost during the monitoring, when it is not operated at the system start-up, they first need to be temporarily stored. A copy of the SRAM can be stored at the main memory and upon completion of the monitoring written back to the cells, or a process like the refresh operations of the DRAM can be followed for every word-line. In the latter case, the data of each word-line under monitoring will be stored at the I/O register of the memory, the word-line will be then monitored for aging and upon completion of the monitoring, the data will be written back from the register to the cells of the pertinent word-line.

It should be noted that the Counter and the MUX are typical and reusable embedded circuitry in SRAMs for various test procedures and thus, in practice they do not constitute additional cost to the proposed technique.

4.1.4. The Digitizer

The Digitizer block is presented in Figure 4.4. It consists of an n-bit Shift Register (SR) and n Switches (these are n full pass gates – SWi, where n is the number of columns in the memory array under monitoring), an m-bit Counter, for which the Counter in Figure 4.3 can be reused (where 2^m = w that is the number of word-lines in the memory array), two identical counters (the High-Counter HC and the Low Counter LC) and optionally a Divider-Comparator block. The Data Register in Figure 4.4 is not part of the Digitizer; it is the standard I/O data register of the SRAM. A very simple state machine is utilized for the operation of the Digitizer and the corresponding operational diagram is shown in Figure 4.5.

As the bits (n) of the Shift Register SR can be high in number, the SR can be replaced by a $\log_2(n)$ Counter and a Decoder in order to activate again one by one all SWi switches.

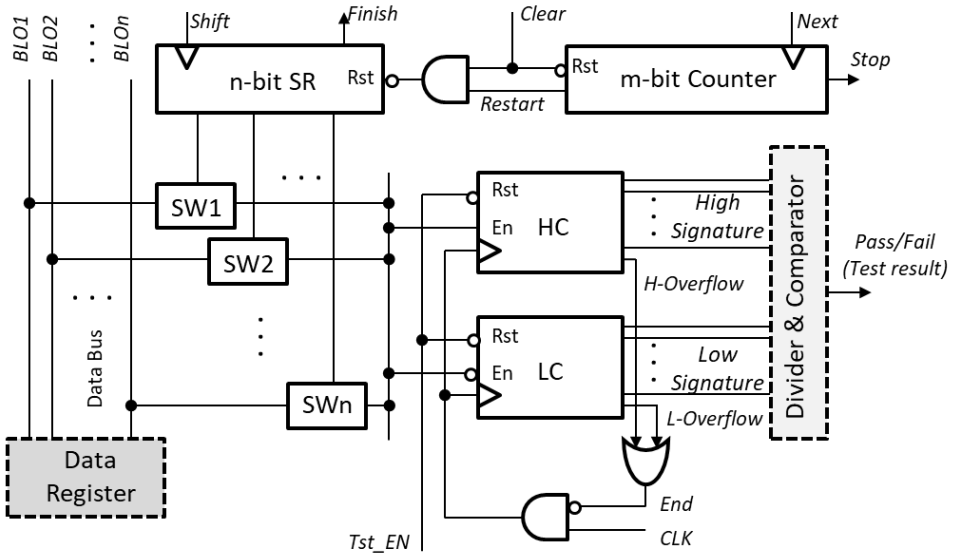

Figure 4.4 The Digitizer circuit.

In the normal mode of operation, the *Tst_EN* signal and the *Clear* signal are low so that the m-bit Counter is initialized to the zero state and the SR is initialized to the all zero state except the leftmost bit (which drives the first Switch SW1) that is set to high, while the two counters HC and LC are reset to the zero state. The Digitizer is in the Idle state.

In the monitoring mode of operation, the *Clear* signal is set to high. Next, the *Tst_EN* signal is also activated to high. The oscillation signal that is generated by the DRO of the first column is propagated through the corresponding output of the first sense amplifier (*BLO1*) and through the SW1 switch which acts only as activation signal for the enable input (*En*) of the HC and LC counters and a memory cell is tested. When the *BLO1* signal is high the HC counts up, while when the *BLO1* signal is low the LC counts up, under the supervision of the *CLK* signal. The *CLK* signal is the system clock and its frequency is higher, at least twice according to the Nyquist theorem, with respect to the frequency of the oscillation signal. Obviously, at any time instance, the value of each counter (high signature and low signature respectively) depends on the duty cycle of the *BLO1* signal. Whenever one of the counters reaches its maximum value, the pertinent

overflow signal is generated (*H-Overflow* or *L-Overflow*). Thus, the *End* signal is activated and both counters are frozen. The ratio of the High and Low signatures corresponds to the duty cycle ratio of the oscillation signal *BLO1*, and is used for decision making on the aging of the pertinent module (memory cell or sense amplifier) under monitoring. As an optional circuit, a local Divider can be exploited for the ratio generation, which is compared with the reference value. Alternatively, since in general an SRAM is embedded in a microprocessor system, the existing ALU can be used for the ratio generation and comparison. In case that the test result is "Fail" a repair operation follows, as it will be further discussed.

Next, the *Tst_EN* signal is deactivated to low and the two counters HC and LC are reset. In addition, a pulse is generated on the *Shift* signal for a single shift operation in the SR. Consequently, only the second Switch SW2 is now on. Then, the *Tst_EN* signal is activated and the oscillation signal through the *BLO2* output of the second column feeds the enable input (*En*) of the HC and LC counters. A second round of measurements follows for the evaluation of the second memory cell according to the above discussion, and so on.

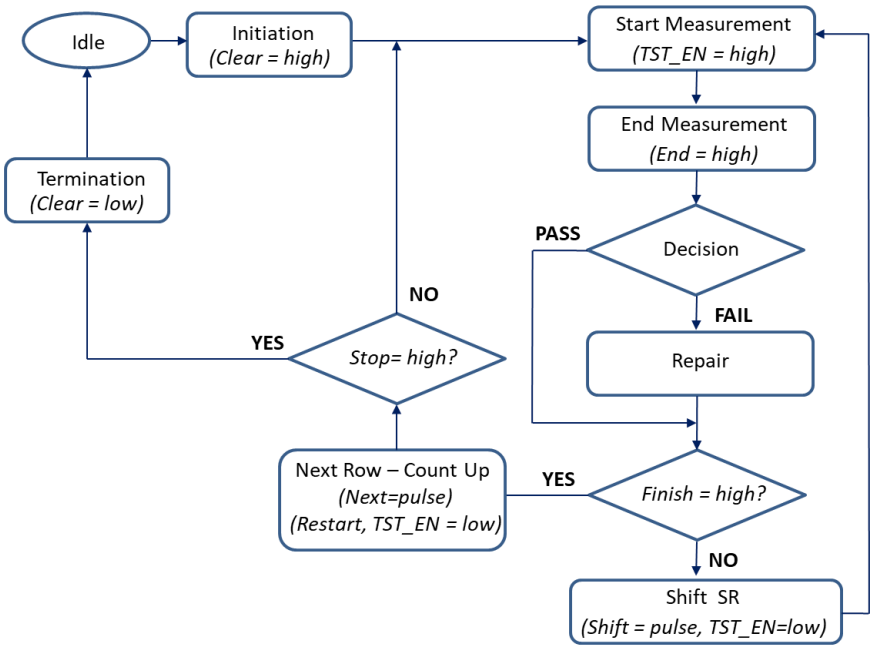

Figure 4.5 The operational diagram of the Digitizer.

After the evaluation of all memory cells in a row (e.g. after n shift operations in the SR), the *Finish* signal is generated by the SR. Then, a pulse is applied on the *Next* signal for a single count-up operation by the m-bit Counter in order to select and evaluate the next row in the memory array. Thus, the SR is re-initialized by the *Restart* signal that is generated by the m-bit Counter and the above measurement-evaluation-repair procedure is repeated for this row. When the cells of all rows in the memory array ($w=2^m$ in number) have been measured and evaluated (e.g. the m-bit Counter overflows), then the Stop signal is activated to high and the Digitizer block returns to the Idle state.

4.1.5. Repairing Methodology

Aiming to guarantee the reliable operation of the memory through its lifetime, after the detection and location of SRAM cells that are seriously affected by aging and are near failure generation, a proper self-repairing methodology must be considered. Towards this direction, the repair method proposed in [58] can be adopted. According to this, in case of a single over-aged cell in a memory array row, existing error correcting codes (ECCs), commonly used in SRAMs, are exploited to correct possible errors. Thus, single bit errors are tolerated. If a second cell in a row is also detected as over-aged, this can be replaced by a spare (redundant) one exploiting existing repair mechanisms that are embedded in the memory [114]-[116].

In the case of over-aged memory cell detection, an ageing mitigation technique such as the self-controlled bit-flipping [104] previously described in Chapter 3.5.1 or the periodic bit flipping can be also adopted in order to reduce the appearance rate of over-aged cells. Such a technique can be used to balance the signal probabilities of SRAM cells and reduce the impact of aging. In [117] it is proposed that the contents of the entire cache are periodically flipped at predefined time intervals. When the flipping process is activated, the first memory address is

accessed and the bit lines are read, inverted and written back to the same address. This process is continued until all addresses are accessed and the entire data inside the cache is flipped.

The above repairing methodology can be periodically applied in the field of operation, e.g. at the system start-up, aiming to increase the reliability and the effective lifetime.

4.1.6. Overall assessment

Towards the direction of the SRAM's self-healing, an aging monitoring technique for SRAM memory cells is presented. It is based on a simple, low silicon area cost, differential ring oscillator with the ability to detect overaged cells and exploit repair mechanisms to retain the SRAM's reliable operation. The proposed technique can be periodically applied and the duty cycle of the ring oscillator signal can be exploited to discriminate the degraded (skewed) cells. The ring oscillator is activated only during the monitoring mode and is disabled during the normal mode of operation for the power consumption elimination and the reduction of its transistor aging. Later in this chapter we will see how the effect of aging on the DRO, during the normal mode of operation, is eliminated. Furthermore, the influence of the monitoring scheme on the normal operation of the SRAM is negligible.

Finally, the proposed technique provides the ability to locate over aged SRAM cells in order to replace them (by exploiting existing repair mechanisms that are embedded in the memory) before the occurrence of a potential failure.

In the next section of this chapter, the previously presented technique is extended and analyzed for the aging monitoring of the SRAM's sense amplifiers.

4.2. Aging Monitoring for SRAM Sense Amplifiers

The impact of BTI and HCI aging on the transistors of the SRAM sense amplifiers is also crucial for the SRAMs' performance, as it has been previously explained in Chapter 3.3.2, since its presence results in gradual speed performance degradation [72] as well as in input offset voltage development [83], [98]. Consequently, failures are expected to occur during an SRAM operation under the continuous degradation of the sense amplifiers performance. Thus, it is of major importance to develop aging monitoring techniques for the sense amplifiers, that will provide the ability to predict upcoming failures in the read operations and early react to retain the reliable operation of the SRAM.

The on-line BTI and HCI aging monitoring technique presented in Chapter 4.1 for SRAM memory cells, was considered in order to achieve the effective aging monitoring of the SRAM sense amplifiers, too. The monitoring topology, which is based on the duty cycle of the signal provided by a low cost differential ring oscillator for the detection of over-aged sense amplifiers, will be next presented.

4.2.1. The monitoring circuitry

The proposed BTI and HCI monitoring architecture for monitoring over-aged sense amplifiers (SAs) and allow the self-healing of the SRAM, is illustrated in Figure 4.6. A Differential Ring Oscillator (DRO), similar to the one used for monitoring the SRAM memory cells, is embedded in the memory array and placed at every bit-line in a manner as to drive every SA during the monitoring mode of operation. According to our observations, a skewed SA due to aging alters the duty cycle of the oscillation signal that is generated by the DRO. Consequently, the duty cycle of this signal is exploited in order to identify over-aged SAs.

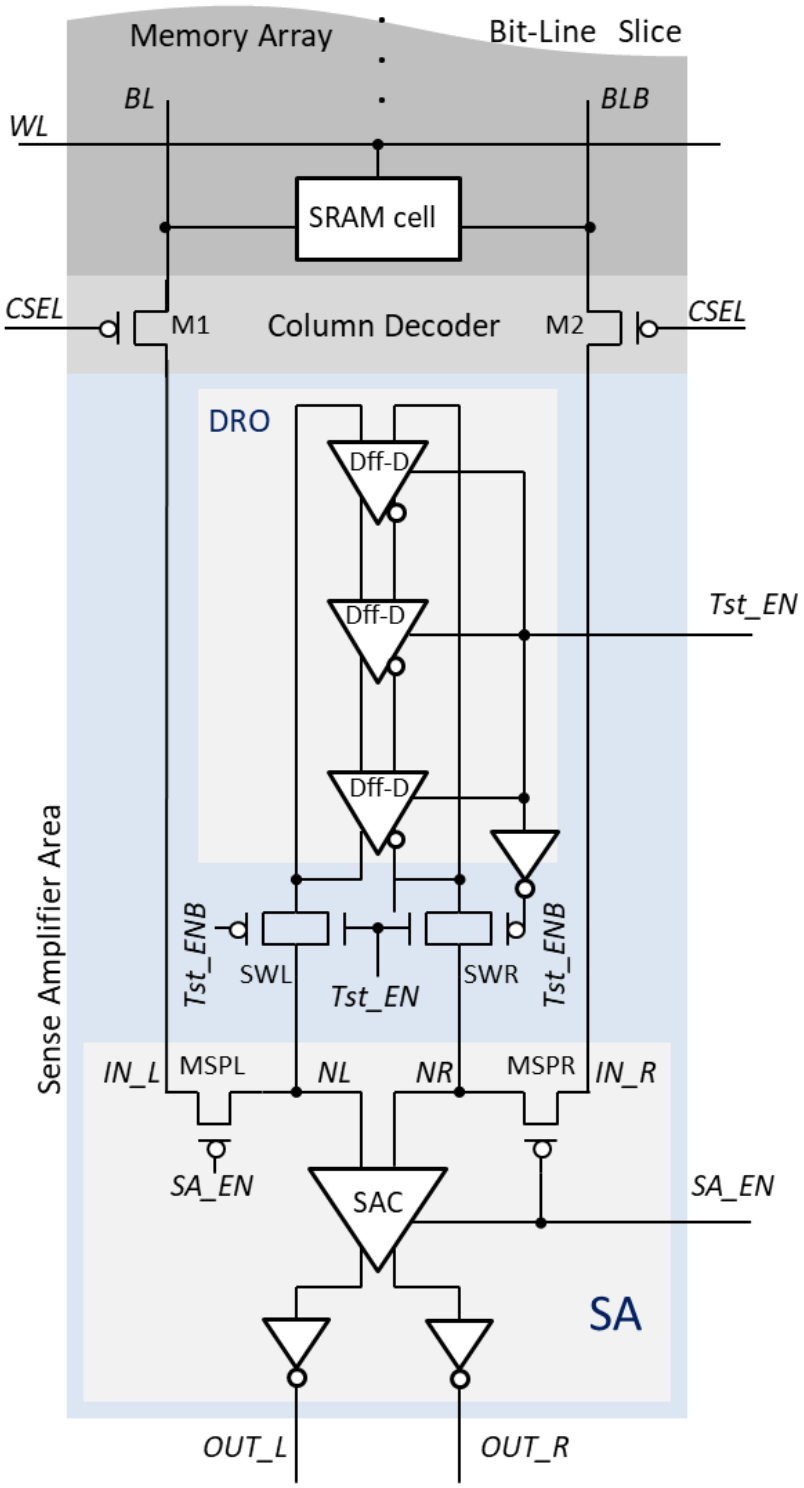

Figure 4.6 Monitoring circuitry for testing SRAM sense amplifiers.

Similarly to the case of the memory cell testing, in Figure 4.6 a bit-line slice (column) of an SRAM memory array along with the sense amplifier (SA) is presented. A bit-line access transistor pair M1-M2 (one per bit-line *BL* and bit-line bar *BLB* respectively) permits the access to the memory array, by exploiting the signal *CSEL* (column select). It should be noted that for simplicity, the write circuitry is not included in the figure.

Three tri-stated differential delay elements (Dff-D cells) are exploited again to form the DRO, which is located in the area of the SA. With respect to Figure 4.1 two additional switches (full pass gates) SWL and SWR (see Figure 4.6) are included in the DRO design in order to provide access to the internal nodes *NL* and *NR* of the SA core and isolate the DRO during the normal mode of operation. In the monitoring mode of operation the Dff-D cells and the two switches are activated (at logic high) by the *Tst_EN* (test enable) signal while they are deactivated (at logic low) in the normal mode of operation.

In the monitoring mode of operation, both the *SA_EN* and *Tst_EN* signals are set to high and the SA plays the role of an "active" load at the output nodes of a stage of the DRO. Since the DRO is activated, an oscillation is generated in the chain and is propagated to the outputs *OUT_L* and *OUT_R* of the SA where it can be retrieved for the needs of testing. During its monitoring, the SA operates similarly to a memory cell which is written to different values, rather than "sensing" the voltage difference of the bit-lines. The discrimination of aged SAs is succeeded with the exploitation of the duty cycle of the oscillation signal generated. This process is conducted outside the memory array where the oscillation signal is propagated in order to perform measurements of its duty cycle for the SA evaluation. Note that again the only extra signal for the support of the monitoring operation, with respect to a typical SRAM, is the *Tst_EN* signal, which is common for all bit-line slices.

The ability of the proposed monitoring circuitry to exploit the duty cycle of the generated signal for aging detection of both the SRAM memory cell and the SA, is based on the following observation. Without loss of generality, let us consider that in a given time instant during testing the *NL* node of the SA is at gnd while the NR node is at V_{DD}. The voltage level of node *NR* is defined as V_{NR}. Consequently, the gates of transistors MP2 and MN1 are at gnd and V_{DD} respectively and the gates of transistors MP1 and MN2 are at V_{DD} and gnd respectively (the latter transistors are off). Moreover, let us consider that due to aging the absolute threshold voltage of transistors MP2 and MN1 is increased by ΔV with respect to the threshold voltage of transistors MP1 and MN2.

Next, since the driving differential signal of the DRO oscillates, it will discharge node *NR* and charge node *NL*. For the discharging of node *NR* and the charging of node *NL*, the DRO output nodes that drive the SWR and SWL switches turn to gnd and V_{DD} respectively. A reliable flip of the SA state is ensured if we pull node *NR* low enough (below the threshold voltage of MN1) and pull node *NL* high enough (above V_{DD} minus the absolute threshold voltage of MP2). For the first case, at the threshold point, the following DC current equation stands. The pMOS transistor of the switch is in the cut-off region.

$$k_{nw}\left((V_{DD}-V_{tn})V_{NR}-\frac{V_{NR}^2}{2}\right)=k_{p2}\left((V_{DD}-|V_{tp}|-\Delta V)V_{Dsatp2}-\frac{V_{Dsatp2}^2}{2}\right) \quad \text{Eq. 4.1}$$

where k_{nw}, k_{p2} are the current gain factors and V_{tn}, V_{tp} are the threshold voltages of the nMOS transistor of SWR and the MP2 transistor respectively while V_{Dsatp2} is the saturation drain voltage of MP2. Without loss of generality, let us consider that $V_{tn} = |V_{tp}| = V_t$. Then solving for V_{NR} we take:

$$V_{NR} = V_{DD} - V_t - \sqrt{(V_{DD} - V_t)^2 - a\left((V_{DD} - V_t - \Delta V)V_{Dsatp2} - \frac{V_{Dsatp2}^2}{2}\right)} \quad \text{Eq. 4.2}$$

where α depends on k_{nw}, and k_{p2}. From equation Eq. 4.2 we observe that for an aged MP2 transistor the V_{NR} voltage level is lower than the pertinent value when MP2 is "fresh", due to the presence of ΔV. This indicates that it is much easier under aging conditions to flip the SA from the initial state to the opposite state since the requirement for a low V_{NR} voltage level is achieved easier. In a similar way the aging of MN1 further facilitate the SA flipping. However, since transistors MP1 and MN2 are not aged (or they are not as aged as MP2 and MN1), the SA flipping form the state where NR is at gnd and NL is at V_{DD} to the opposite state is not alleviated at all. Furthermore, the aging of MP2 and MN1 does not support this flipping. Consequently, in the presence of the aging conditions under consideration, the duty cycle of the generated oscillation signal, when the aged SA plays the role of an active load at the output of the DRO, is expected to alter with respect to this of the "fresh" SA. This observation stands even when the threshold voltage shift of MP2 transistor is not equal to this of MN1 transistor, e.g. ΔV is not common for these two transistors.

4.2.2. The Differential Ring Oscillator (DRO)

The DRO exploited for the testing of the sense amplifiers follows the same specifications and functionality with the description previously presented in Chapter 4.1.2. Accordingly, the structure of the Dff-D element of the DRO can be advised by the Figure 4.2. Let us however summarize its operation during normal mode, due to the presence of the extra access switches with respect to Figure 4.2.

When the SRAM operates in normal mode, the *Tst_EN* signal is low and the Dff-D cells are inactive and disconnected from the power supplies so that their nodes are floating. Since aging monitoring is a periodic activity of only a short duration, the SRAM operates most of the time in normal mode. In addition, in this mode, the switches SWL and SWR (in Figure 4.6) are also inactive and the DRO is isolated from the rest circuitry. Thus, due to the presence of leakage currents the internal nodes of the DRO are stabilized at intermediate voltage levels between V_{DD} and gnd. Consequently, a homogenous, low, voltage stress affects the transistors of the DRO (MDP1-2 and MDN1-2 in Figure 4.2) so that their aging is quite slow (transistors MDP3 and MDN3 are not under stress in the normal mode). Although this may affect the signal frequency of the standalone DRO, the duty cycle of this signal is not altered due to the homogenous voltage stress. Thus, the DRO can effectively support the SA aging monitoring throughout the SRAM lifetime.

4.2.3. Failure Prediction Methodology

For the aging monitoring of the SAs in an SRAM, the generated oscillation signal of all DROs are propagated through each SA to a Digitizer block for evaluation (the general architecture as previously analyzed, is presented in Figure 4.3). The *BLOi* signal is either the *OUT_L* or *OUT_R* signal of the i-th SA (see Figure 4.6). In the Digitizer block, the duty cycle of each signal is digitized (one after the other) and the result is compared with a reference duty cycle value, in order to detect possible excess aging degradation. After the evaluation of an SA the next SA is evaluated and so on until the evaluation of all SAs in the memory array.

The reference duty cycle value corresponds, as in the case of the memory cell, to an SA that although it is aged enough it does not generate failures during the SRAM operation. Moreover, if the duty cycle measured during the testing of an SA is just outside the limit of this reference value, it will also not generate failures. Nevertheless, if the aging progresses, failures are expected to be generated in short

time and the reliability levels will be degraded. According to the above and as it has previously been shown with the case of the memory cell monitoring, the proposed method can achieve failure prediction early in advance so as to react for the memory repair before any failures occur and thus achieving the SRAM's self-healing.

The aging monitoring method is again suggested for periodic application on the SRAM. For the case where aging has seriously affected an SA, so that it will generate failures in the near future, then after the detection of this status by the proposed technique, the SRAM's signals timing can be properly adjusted (as it will be further discussed) in order to retain the memory reliable operation.

The system's start-up or the idle times can be again used for the periodic application of the monitoring procedure. Alternatively, in order to meet requirements of high reliability, it can be applied at the end of predetermined time intervals, like the refresh operations in DRAMs. It should be noted that the memory's performance will not be affected by this kind of periodic activity.

4.2.4. The Digitizer

The Digitizer block used for the scheme previously presented for aging monitoring of sense amplifiers, is slightly modified to a more simple version from the one described in Section 4.1.4, that is without the inclusion of the m-bit Counter used for counting the word-lines. The alternative topology of the Digitizer is presented in Figure 4.7. It consists of an n-bit Shift Register (SR) and n Switches (these are n full pass gates – SWi, where n is the number of word-lines in the memory array, corresponding to the number of sense amplifiers under monitoring), two identical counters (the High-Counter HC and the Low Counter LC) and optionally a Divider-Comparator. The I/O Circuitry in Figure 4.7 is again not part of the

Digitizer. A very simple state machine is utilized for the operation of the Digitizer and the corresponding operational diagram is shown in Figure 4.8.

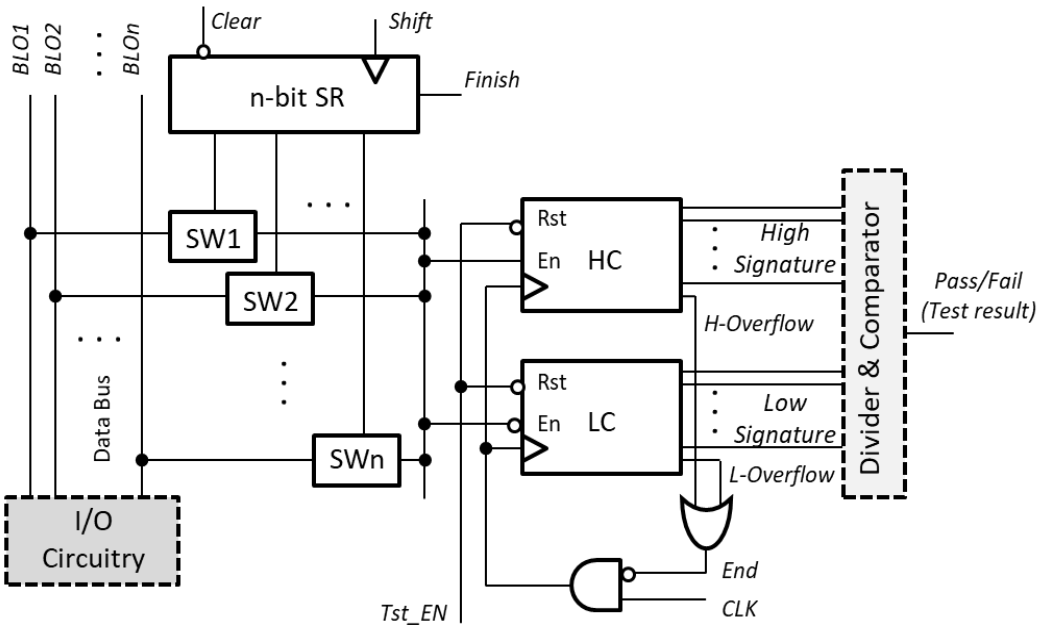

Figure 4.7 The Digitizer circuit.

In the normal mode of operation, the SR is initialized to the all zero state except the leftmost bit, while the two counters HC and LC are reset to the all zero state, setting the Digitizer in the Idle state.

In the monitoring mode of operation, with the *Clear* and *Tst_EN* signals set to high, the oscillation signal generated by the DRO of the first column through the *BLO1* output is propagated through the SW1 switch to the enable input (*En*) of both counters HC and LC and the corresponding SA is tested. Consequently, depending on the value of the *BLO1* signal, either the HC or the LC Counter counts up, under the supervision of the *CLK* signal. Whenever one of the counters reaches its maximum value, the pertinent overflow signal is generated (*H-Overflow* or *L-Overflow*), the *End* signal is activated and both counters are frozen. The ratio of the High and Low signatures corresponds to the duty cycle ratio of the

oscillation signal *BLO1*, which is used for decision making on the aging of the pertinent SA by comparing it against a reference value. As previously indicated, since generally an SRAM is embedded in a microprocessor system, the existing ALU can be used for the ratio generation and comparison. In case that the test result is "Fail" a repair operation follows, as it will be discussed next.

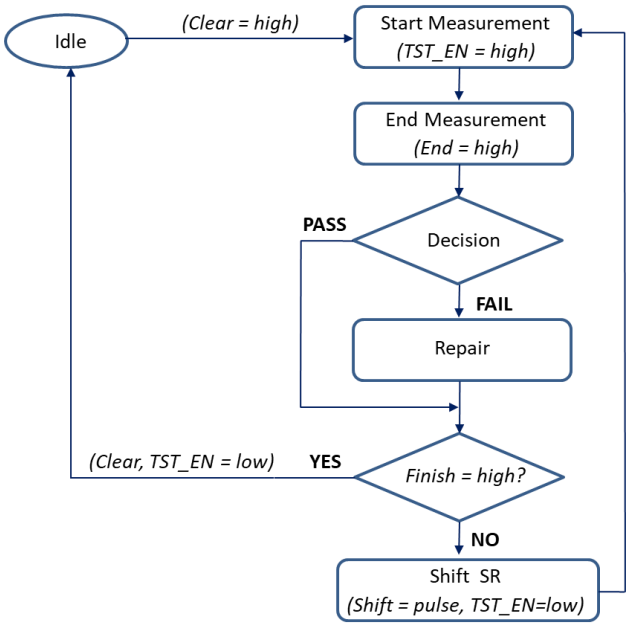

Figure 4.8 The operational diagram of the Digitizer.

Next, the *Tst_EN* signal is deactivated to low and the two counters HC and LC are reset. In addition, a pulse is generated on the *Shift* signal for a single shift operation in the SR. Consequently, the second Switch SW2 is on. Then, the *Tst_EN* signal is activated and the oscillation signal *BLO2* of the second DRO-SA system feeds the enable input (*En*) of both counters. A second round of measurements follows for the evaluation of the second SA according to the above discussion, and so on. After the evaluation of all SAs, the *Finish* signal is generated by the Shift Register and the Digitizer returns to the Idle state.

4.2.5. Repairing Methodology

In the case of an over-aged SA, the memory enters the repair phase. The established offset voltage can be tolerated by increasing the voltage difference on the bit-lines in order to force the SA to sense correctly the stored value. One possible solution under the occurrence of a single over-aged SA in a memory array, is to exploit existing single error correcting codes (ECCs), commonly used in SRAMs, for error masking. Thus, single bit errors can be tolerated. When no ECC is used and an SA is detected as over-aged or ECC is exploited and a second SA is also detected as over-aged, during the repairing phase, the SA activation delay time is increased by a constant value in order to set the new aging tolerance margin. Thus, the time interval between the activation of the world-line (*WL* set to logic '1') and the activation of the SA (*SA_EN* set to logic '1') is increased. In parallel, the reference duty cycle ratio value in the register is also properly adapted by a predetermined at the manufacturing stage constant step. This adjustment of the activation delay will cause an impact on the system timing; nevertheless the repairing step is important for the needs of self-healing and in order to maintain the reliable operation of the SRAM.

4.2.6. Discussion on the Monitoring Procedures

Aiming to ensure the SAs' reliable operation their monitoring procedure is scheduled according to Figure 4.9. After system power up, the reference duty cycle is transferred from the ROM to a dedicated register. Then, the SA monitoring phase starts in order to detect degraded SAs. This step is required due to possible SAs' aging during previous system activations. In case that all SAs are healthy, the system enters the normal mode of operation.

If an SA is characterized as over-aged, the repairing procedures are applied and the new reference value is adapted. Next, a second SA monitoring phase is applied since the previous repair procedure may not cover all over-aged SAs. If this is

true, a second repair phase follows in order to further adjust the SA activation delay and the reference value and so on until all SAs are healthy according to the new duty cycle ratio reference value. Then, the system enters the normal mode of operation.

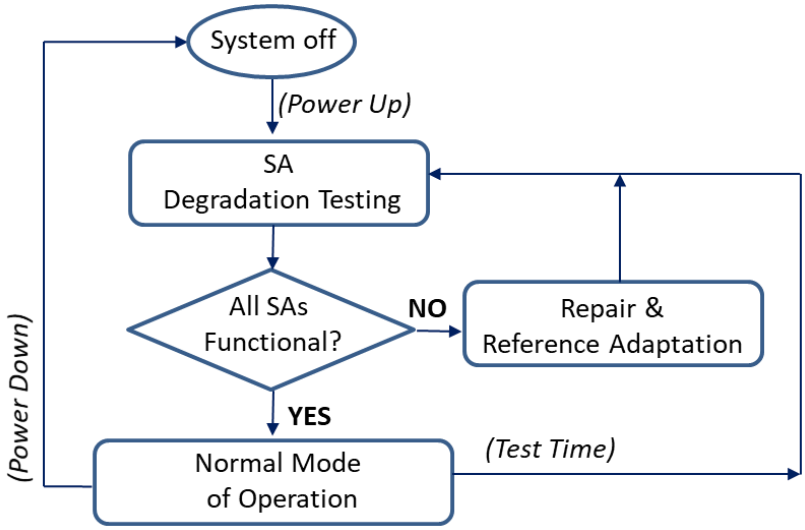

Figure 4.9 Monitoring scheduling.

4.2.7. Overall assessment

The detection of over-aged (near failure) SRAM sense amplifiers can be succeeded with the use of the proposed periodic BTI or HCI related aging monitoring technique. The monitoring circuit is constructed by exploiting a simple, low cost, differential ring oscillator. During monitoring mode, the duty cycle of its oscillation signal is used to discriminate degraded sense amplifiers. During the normal mode, the ring oscillator is deactivated and isolated from the rest circuit in order to alleviate the influence of aging on the monitoring topology.

After the detection of an over aged sense amplifier (which is near failure generation), the proposed technique provides the ability to adjust the timing of the memory signals and succeed in maintaining the reliable operation of the SRAM through a self-healing approach.

4.3. Alternative Aging Monitoring for SRAM Sense Amplifiers

An alternative approach of the previously presented technique was also developed for the needs of the BTI and HCI aging monitoring on SRAM sense amplifiers. The idea was once more based on the use of a low cost differential ring oscillator considering however some alterations on the structure of the circuitry. For the detection of over-aged sense amplifiers, the frequency ratio of two signals is used and the ring oscillator turns to be reconfigurable with the addition of a pair of switches, as it will be next described.

4.3.1. The monitoring circuitry

The alternative proposed scheme for BTI and HCI aging monitoring on SRAM SAs is presented in Figure 4.10. A reconfigurable differential ring oscillator (rDRO) is constructed for every sense amplifier, where the SA can be part or not of this oscillator. The oscillation frequency of the rDRO measured in two modes is used for the detection of aged SAs.

In Figure 4.10 the bit-line slice of a memory array is presented. Similarly to the previous approach in Chapter 4.2, the bit-line access transistor pairs M1-M2 (one per bit-line *BL* and bit-line bar *BLB*) permit the access to the memory array, by exploiting the signal *CSEL*. For the read operations the SA is activated by the *SA_EN* signal (active high). Once again, for simplicity reasons the write circuitry is not shown in the figure.

The rDRO, which is placed in the SA area, consists of three tri-stated differential delay elements (Dff-D cells) and the *Tst_EN* (test enable) signal is used for the activation (at logic high) of the Dff-D cells in the monitoring mode of operation or their deactivation (at logic low) in the normal mode of operation. In the monitoring mode of operation, an oscillation is generated in the chain and is propagated to the *OSC* output of the rDRO for the needed measurements.

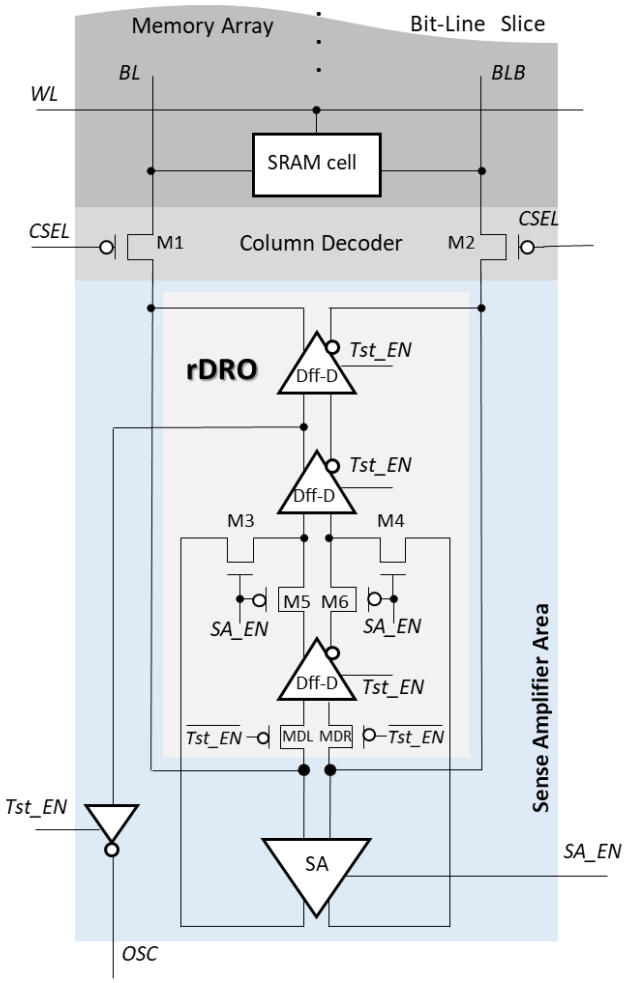

Figure 4.10 The alternative circuitry for monitoring the aging on SAs.

The rDRO is reconfigurable by means that the bottom Dff-D cell can be replaced by the SA (which is also a tri-stated differential module as presented in Figure 3.9). Thus, two oscillating signals can be generated, one by the three Dff-D cells of the ring oscillator without the SA and one by two Dff-D cells of the rDRO and the SA in the loop. The ratio of the frequencies of these two signals is exploited for the monitoring of the SA performance degradation. The frequency ratio is adopted in this approach aiming to compensate process variations.

The reconfiguration is achieved by the insertion of four transistors (two nMOS M3-M4 and two pMOS M5-M6 – see Figure 4.10). The *SA_EN* signal feeds the gates of all four transistors so that when *SA_EN* is low the pair of pMOS

transistors is on, while when *SA_EN* is high the pair of nMOS transistors is on. In the first case, the rDRO is constructed by the three Dff-D cells (Config-1) while in the second case the rDRO is constructed by the two upper Dff-D cells and the SA (Config-2). Two pMOS transistors MDL and MDR are also exploited in order to isolate the rDRO from the rest of the circuit during the normal mode of operation.

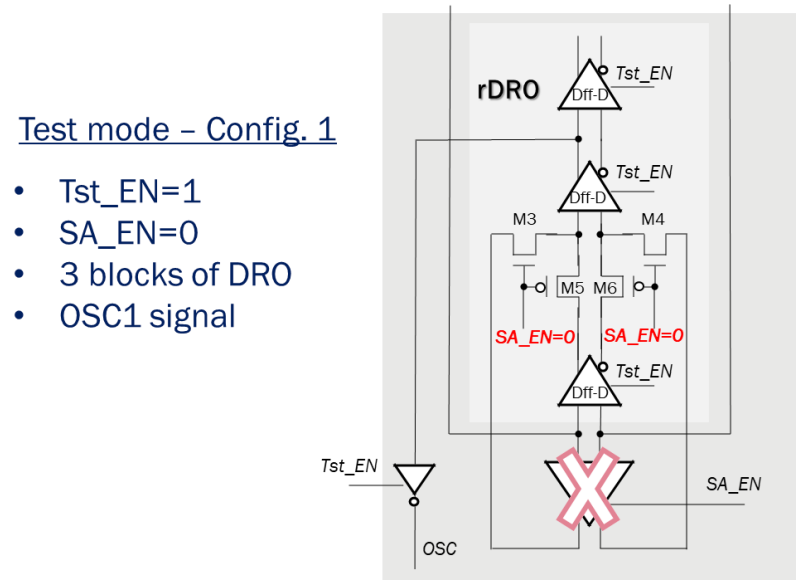

Figure 4.11 Config-1 of the monitoring circuit.

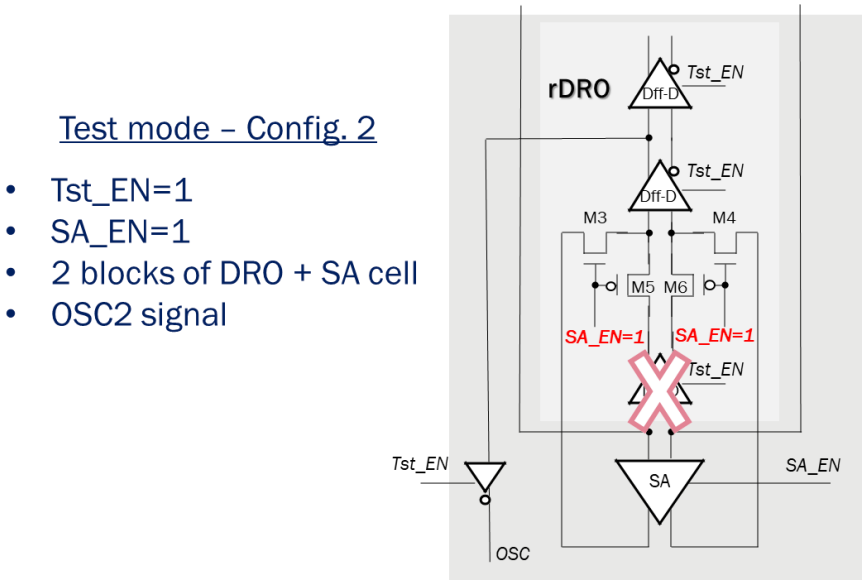

Figure 4.12 Config-2 of the monitoring circuit.

In more detail, during the monitoring mode, the operation is as follows. The *Tst_EN* signal is turned to high allowing the rDRO to connect to the circuit (transistors MDL and MDR of Figure 4.10 are activated). Initially, the first configuration is activated (Config-1 as presented in Figure 4.11) by setting *SA_EN* at low. The oscillating signal OSC_1 is generated by the ring oscillator which is propagated outside the memory array for frequency digitization and storing. Next, the *SA_EN* is turned to high so that the second configuration is activated (Config-2 as presented in Figure 4.12), where the SA is active. A new oscillating signal OSC_2 is generated by the ring oscillator which is also propagated outside the memory array for frequency digitization and storing. Right after the two configurations are completed, the monitoring mode is deactivated by setting the *Tst_EN* signal to low. Next, the ratio of the two frequencies is compared with an embedded reference value in order to detect over aged SAs.

4.3.2. The reconfigurable Differential Ring Oscillator (rDRO)

The Dff-D cells of the rDRO exploited in the alternative monitoring approach for the case of over-aged sense amplifiers follow the same specifications and functionality with the description previously presented in Chapters 4.1.2 and 4.2.2 accordingly. With the use of only one extra signal (*Tst_EN*) and its complementary, the cells of the rDRO are either activated or deactivated and thus the circuit enters the monitoring mode or the normal mode of operation.

Since the proposed circuitry is used for periodic monitoring of BTI and HCI aging in SRAM SAs, most of the time the memory operates in the normal mode during which the rDRO is inactive (*Tst_EN*=low and the Dff-D cells are disconnected from the power supplies so that their nodes are floating). During only very small time intervals, where the SA degradation is evaluated, the rDRO is activated. In addition, during the normal mode of operation, the transistors MDL and MDR (see Figure 4.10) are also inactive and the DRO is isolated from the rest of the

circuitry. Thus, due to the presence of leakage currents the internal nodes of the DRO are stabilized at intermediate voltage levels between V_{DD} and gnd. Consequently, the transistors of the Dff-D cells are not frequently stressed and the oscillation frequency (period) of the rDRO in the first configuration (Config-1) is almost not altered over time.

4.3.3. Failure Prediction Methodology

As previously described, SAs are frequently activated for the needs of read operations in the SRAM. Thus, the corresponding transistors are continuously stressed and their threshold voltage is degraded leading to increased propagation delays. As a result, the transistor aging in the SA alters (increases) the propagation delay in the rDRO at the second configuration (Config-2) and the oscillation frequency is decreased. Thus, measuring the oscillation frequency in the two configurations and calculating their ratio, aging monitoring can be achieved and over aged SAs can be discriminated.

According to the proposed methodology, the *OSC* signal generated by each configuration is transferred (one after the other) to a Digitizer block for evaluation. In the Digitizer block, the frequency of each signal is digitized and then the frequencies' ratio is calculated. The result is compared to a reference ratio, which corresponds to the case of a "fresh" SA, in order to detect possible excess aging degradation. After the evaluation of an SA the next SA is evaluated and so on until the evaluation of all SAs in the memory array.

The reference frequency ratio is derived during the design phase or during prototyping and it is stored in a simple and very small ROM in the chip (like these that support the repair operations in an SRAM). The reference can be the same for all dies under a worst case scenario that is based on statistical analysis of circuits' measurements. Alternatively, a dedicated per die reference can be used

that is derived after measurements during manufacturing testing. Given that the frequency ratio is defined as the ratio of the frequency at the first configuration to this of the second configuration, we select the reference ratio so that an SA with a frequency ratio just above this value although it is considered as over-aged, it does not generate failures during the SRAM operation. Thus, this over-aged SA can be detected by the monitoring procedure before failures appearance in the SRAM.

The monitoring procedure can be periodically applied, in the field of operation, at the system start-up or during idle times, aiming to increase the effective lifetime of the SRAM. Periodic monitoring is an effective solution for this alternative approach as well, since aging degradation is a gradual phenomenon.

4.3.4. The Digitizer

The frequency-to-digital conversion in the Digitizer can be achieved by exploiting various techniques like the use of counters, beat frequency detectors [101] or vernier delay lines [118].

4.3.5. Repairing Methodology

After the start-up of the system, the reference duty cycle is transferred from the ROM to a dedicated register and the memory enters the monitoring phase. This step is required due to possible SAs' aging during previous system activations. In case that all SAs are healthy, the system enters the normal mode of operation. In case of an SA that has been seriously affected by aging, so that its operation is near failures generation, after the detection of this status by the proposed technique the repairing methodology also suggested for the latch type SA in Chapter 4.2.5 is adopted for the SRAM's self-healing and in order to maintain the reliable operation of the memory. During the repairing phase, the SA activation delay time is increased by a constant value predetermined at the manufacturing stage in order to set a new aging tolerance margin. This way, the time interval

between the activation of the world-line (*WL* set to logic '1') and the activation of the SA (*SA_EN* set to logic '1') is increased. In parallel, the reference frequency ratio value in the register need to be also properly adapted by the ratio calculated from the second configuration (Config-2).

All SAs are examined one after the other and the repair procedure is applied each time the frequency ratio calculated and compared to the reference value is found outside the allowed limit, until all SAs provide an acceptable ratio. Then, the system enters the normal mode of operation.

4.3.6. Overall assessment

The alternative BTI and HCI monitoring technique for SRAM sense amplifiers presented above, which is based on a simple, low cost, reconfigurable differential ring oscillator (rDRO), also provides the ability of self-healing to an SRAM. The monitoring procedure is periodically applied. The oscillation frequency of the rDRO is used to discriminate degraded sense amplifiers. The influence of the monitoring scheme on the normal operation of the SRAM is very small. In order to alleviate the aging of the monitoring circuitry, the ring oscillator is disabled and isolated from the rest of the circuit during the normal mode.

The proposed approach provides the ability, after the detection of an over aged sense amplifier (which is near failure generation), to adjust the SA activation time in order to maintain the reliable operation of the memory through self-healing operations. However, the use of the presented technique would increase significantly the cost and complexity in the case where it was adopted for aging monitoring of the memory cells, as additional configurations and storage elements would be needed to examine separately the cells and the SAs.

4.4. Unified Aging Monitoring Approach for SRAM Memory Cells and Sense Amplifiers

The above analysis was focused on individual techniques for either the aging monitoring on SRAM memory cells or SRAM sense amplifiers. As has been discussed in detail, both modules (memory cells and sense amplifiers) are prone to aging stress, which results to their gradual degradation and thus the appearance of reliability issues in the SRAM memory operation. Aiming to deal with the presence of aging in SRAM memory cells and SAs and achieve an extended and more complete self-healing option, this section combines the above aging monitoring techniques and extends the functionality of the proposed scheme in order to provide a consolidated aging monitoring technique for both modules. This technique focuses on SRAM aging prediction and the early diagnosis of the memory status by examining the duty cycle of a ring oscillator signal. A detailed repairing method for the aging mitigation of both the memory cells and the sense amplifiers in an SRAM is also analyzed for the case where transistors have excessively degraded.

4.4.1. The monitoring circuitry

In Figure 4.13 the aging (due to BTI and HCI) monitoring architecture for the two SRAM modules (sense amplifiers and memory cells) is presented for a bit-line slice. The scheme uses the differential ring oscillator (DRO) presented in Section 4.1.1 so that during the monitoring phase an SRAM memory cell or a sense amplifier under monitoring acts as the "active" load at its output. A dedicated DRO is placed at every bit-line slice in the memory array, to drive either the corresponding sense amplifier or an activated memory cell of this bit-line while operating in monitoring mode. For the detection of over-aged memory cells or sense amplifiers, the duty cycle of the generated oscillation signal by the DRO is exploited. The simulation results presented later in this Chapter, show that the

duty cycle of this signal is altered with the aging of the modules (since they become skewed).

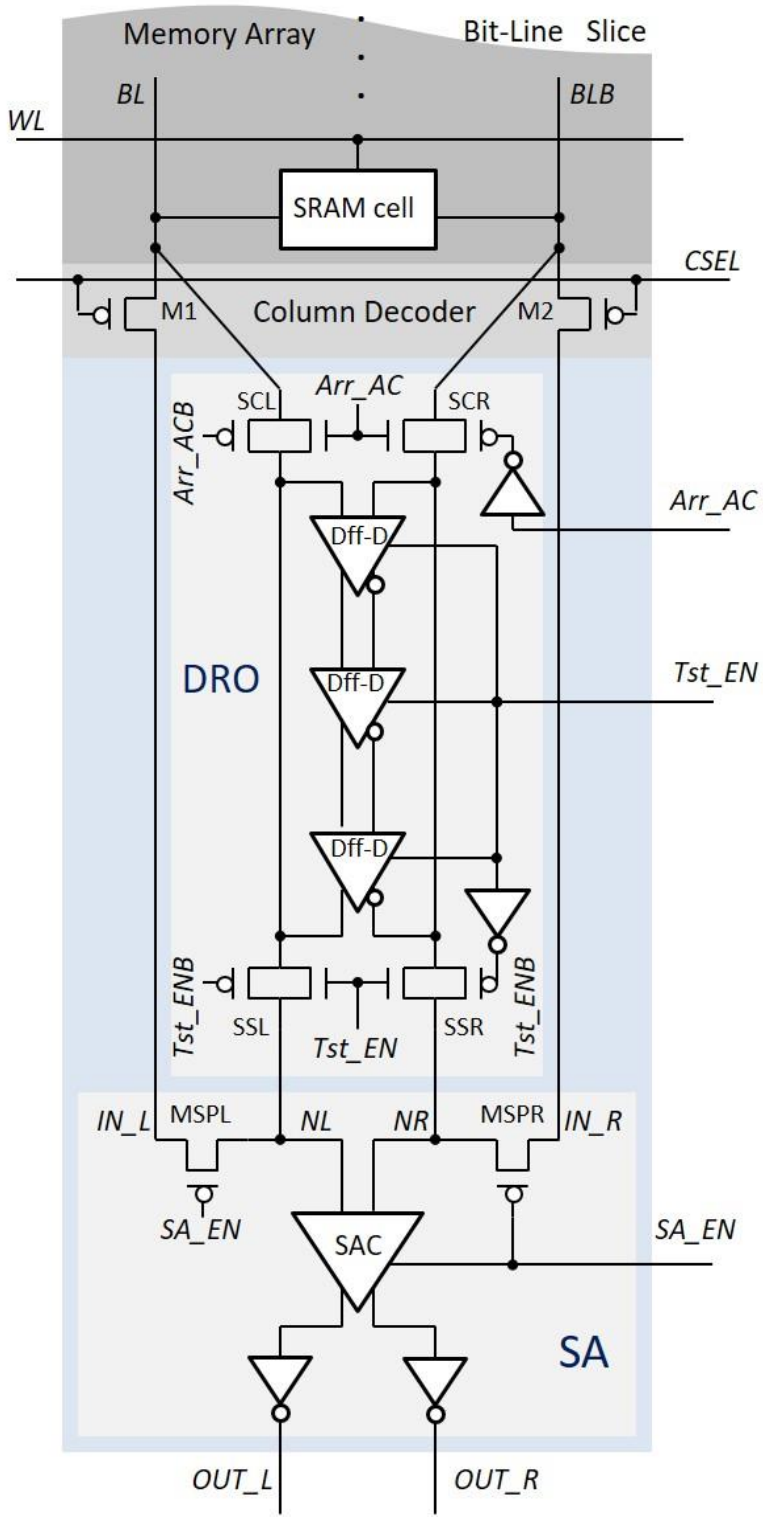

Figure 4.13 The proposed circuitry of the unified approach for aging monitoring.

In the normal mode and during the read operation, the column decoder signal *CSEL* is set to low so that transistors M1, M2 (for bit-line *BL* and *BLB* accordingly) are activated and allow the access of the sense amplifier to the memory cells of the array. The *CSEL* signal is never activated during the monitoring mode. It should be noted that for simplicity the circuitry for writing operations does not appear in Figure 4.13.

The three tri-stated Dff-D cells that form the DRO are activated during the monitoring mode by setting the test enable signal *Tst_EN* to high, while in normal mode the signal *Tst_EN* is set to low. When the DRO is active it drives either the internal nodes *NL* and *NR* of the sense amplifier core (SAC) through a pair of switches (full pass gates) SSL and SSR or the bit-lines *BL* and *BLB* of the corresponding bit-line slice through a second pair of switches SCL and SCR (see Figure 4.13). The first pair of switches (SSL and SSR) is activated by the signal *Tst_EN* while the second pair (SCL and SCR) is activated by the memory array access (*Arr_AC*) signal (both signals are active at logic high). Two monitoring modes of operation exist, one for the monitoring of the memory cells and one for the sense amplifiers.

Initially, the monitoring mode for testing a memory cell is discussed. The sense amplifier remains inactive in this mode (*SA_EN* is low). The word-line (*WL*) of the cell under monitoring is activated (at logic high), while the bit-line precharge circuitry (not shown in Figure 4.13 also for simplicity) remains inactive too. In addition, the bit-line pair *BL* and *BLB* is accessed by setting the *Arr_AC* signal to high. The DRO is activated (*Tst_EN* is set to high) and the generated oscillation signal drives the selected memory cell in the array, which acts as the active load. Each time the differential signal of the ring oscillator changes state the corresponding memory cell is written to the complementary state with respect to its present state. Given that the *Tst_EN* signal is active the switches SSL and SSR

are on so that the oscillation signal is further propagated to the *OUT_L* and *OUT_R* outputs of the sense amplifier.

During the monitoring mode of operation for the sense amplifier monitoring, the *SA_EN* signal is activated. In addition, the *Tst_EN* signal is also activated while the *Arr_AC* signal remains inactive. This time, the oscillation signal generated by the DRO feeds only the sense amplifier, which now plays the role of the active load. Once again this signal appears at the outputs *OUT_L* and *OUT_R* of the sense amplifier.

In both cases, the duty cycle of the oscillation signal is used for the discrimination of aged modules (memory cells or sense amplifiers). The oscillation signal is propagated outside the memory array in order to perform measurements of its duty cycle for the aging evaluation. Then, the duty cycle is compared with an embedded reference value aiming to detect the over-aged modules. Note that two reference values are stored upon exercising the monitoring procedure, one for the memory cells and one for the sense amplifiers. For the support of the monitoring operation two extra signals are introduced, the *Tst_EN* and the *Arr_AC* signals, with respect to a typical SRAM, which are common for all bit-line slices. The monitoring procedure is applied to all word-lines and the sense amplifiers' row, one after the other, in order to examine all cells and sense amplifiers in the array.

The proposed scheme is capable of detecting aging by using the duty cycle of the oscillation signal according to the analysis previously described in Section 4.2.1, which is presented for the SA case but also stands for the SRAM cell.

4.4.2. The Differential Ring Oscillator (DRO)

The DRO exploited for the proposed unified approach for the aging monitoring of both the memory cells and the sense amplifiers follows the structure previously

presented in Chapter 4.1.2 and the Dff-D element of the DRO can be advised by the Figure 4.2.

In the normal mode of operation when the *Tst_EN* signal is low, the cells of the DRO are detached from the power supplies and thus deactivated with their internal nodes floating. This mode occupies most of the SRAM's operation time considering that the aging monitoring is applied only periodically and for a short period of time. During monitoring mode the *Tst_EN* signal is set to high, activating the chain of the Dff-D cells which generate an oscillation for testing either the memory cell or the sense amplifier depending on the monitoring mode under consideration.

4.4.3. Failure Prediction Methodology

The methodology that is analyzed next is used for periodic monitoring, in the field of operation, of the memory cells and sense amplifiers aging in SRAMs. In the monitoring mode, either for a memory cell or a sense amplifier in a bit-line slice, the generated oscillation signal of the DRO reaches the outputs of the corresponding sense amplifier and is transferred to a Digitizer block for evaluation (the overall structure is illustrated in Figure 4.3). The output of the i-th sense amplifier (either the *OUT_R* or the *OUT_L* in Figure 4.13) corresponds to the *BLOi* signal in Figure 4.2. The Digitizer measures and digitizes the signal's duty cycle in a serial manner (each time a single cell or sense amplifier is evaluated). Then, any excess aging degradation is detected upon comparison of the digitized value and a reference value of the duty cycle. The reference duty cycle is stored in a ROM like these typically used for memory repair operations. A dedicated word of the ROM is used for the reference duty cycle value of the cells and the corresponding value of the sense amplifiers, since in general these are expected to be different. After the evaluation of a memory cell or sense amplifier the next cell

or sense amplifier is evaluated and so on until the evaluation of all cells or sense amplifiers in the memory array.

The reference duty cycle value is defined based on the same idea described in the previous schemes presented in Sections 4.1 and 4.2. More specifically, it corresponds to a memory cell or a sense amplifier accordingly that even though it is affected by aging, failures are not generated while the SRAM operates. In addition, failures are not generated when the duty cycle of a memory cell or a sense amplifier is slightly over this reference value, but will be soon generated if the aging progresses. Consequently, by applying the proposed methodology failures can be predicted early in advance and the degradation of the SRAM's reliability levels will be alleviated.

The monitoring procedure can be again periodically applied in the field of operation, at the start-up of the system or during idle times, in order to extend the reliable SRAM lifetime. Nevertheless, if there are very high reliability requirements, the monitoring procedure can be applied at the end of predetermined by the manufacturer time intervals, as already indicated. As previously described in Chapter 4.1.3, in order to ensure that the data of the memory cells will not be lost, either the whole memory array needs to be stored to the main memory before the initiation of the monitoring mode, or a read-store-monitor-rewrite procedure should be followed for the cells of every word-line similar to the refresh operations of the DRAMs.

4.4.4. The Digitizer

The Digitizer circuit exploited for the present unified scheme, is the one presented in Section 4.1.4. It consists of an n-bit Shift Register (SR), n Switches (full pass gates – SWi), an m-bit Counter (where $2^m = w+1$ and w is the number of the word-lines in the memory array), two identical counters (the High-Counter HC

and the Low Counter LC) and optionally a Divider-Comparator block. Its operation follows the state machine presented in Figure 4.5.

In monitoring mode, the signal generated by the ring oscillator of the first memory column is propagated via the pertinent *BLO1* sense amplifier output and via the SW1 switch to activate the enable (*En*) input of the HC and LC counters. Depending on the monitoring mode, a memory cell or the sense amplifier of this column is tested. The two counters HC and LC, count up when the *BLO1* signal is high or low respectively. The *CLK* signal is the system clock acting as supervisor of the counters with a frequency twice or more higher compared to that of the oscillation signal. The procedure stops and both counters freeze when either HC or LC reaches its maximum value and the *End* signal is turned on. The decision of whether the module under monitoring (memory cell or sense amplifier) is over-aged, is made according to the ratio of the High and Low signatures, which corresponds to the duty cycle of the oscillation signal on *BLO1*. If "Fail" is the outcome of the testing, a self-repairing mechanism is activated as it will be further described and the SRAM is self-healed.

Next, the two counters HC and LC are reset, a single shift operation in the SR is achieved making the only active switch to be SW2. Then, the *Tst_EN* signal is set to logic '1' and the enable input (*En*) of the LC and HC counters is fed by the oscillation signal of the second column's output (*BLO2*). The second module is then evaluated following a second round of measurements, and so on. Upon having evaluated all modules in a row, a pulse is applied for a single count-up operation by the m-bit Counter in order to select and evaluate the next row in the memory array. Thus, the SR is re-initialized and the above measurement-evaluation-repair procedure is repeated for this row. When the modules of all rows in the memory array (w in number) plus the row of the sense amplifiers have been measured and evaluated, then the *Stop* signal is activated to high and the Digitizer block returns to the idle state.

4.4.5. Repairing Methodology

The proposed repairing methodologies presented previously for the case of the memory cell and the sense amplifier are applied in the current technique depending on the module identified as over-aged. In the case of a single over-aged cell in a memory array row, the existing error correcting codes (ECCs) are exploited to correct possible errors. Thus, single bit errors are tolerated. If a second cell in a row is also detected as over-aged, this can be replaced by a spare (redundant) one exploiting existing repair mechanisms that are embedded in the memory [114]-[116].

In the case of an over-aged SA, the established offset voltage can be tolerated by increasing the voltage difference on the bit-lines in order to force the SA to sense correctly the stored value. This can be achieved by increasing the time interval between the activation of the world-line (*WL* set to logic '1' in Figure 4.13) and the activation of the SA (*SA_EN* set to logic '1').

Obviously, this will result to the increment of the overall response delay of the SA and consequently the system performance degradation. However, aiming to negate aging effects and ensure the reliable operation of the SRAM, the proper adjustment of the time interval between the two activation signals is a viable solution.

4.4.6. Discussion on the Monitoring Procedures

In order to ensure the efficient operation of the SRAM self-healing scheme, the pertinent duty cycle ratio reference value of the SA must be updated after every repair phase. This is crucial in order to avoid characterizing, just after repairing and under the new operating conditions, as over-aged the same SA that triggered the repair operation.

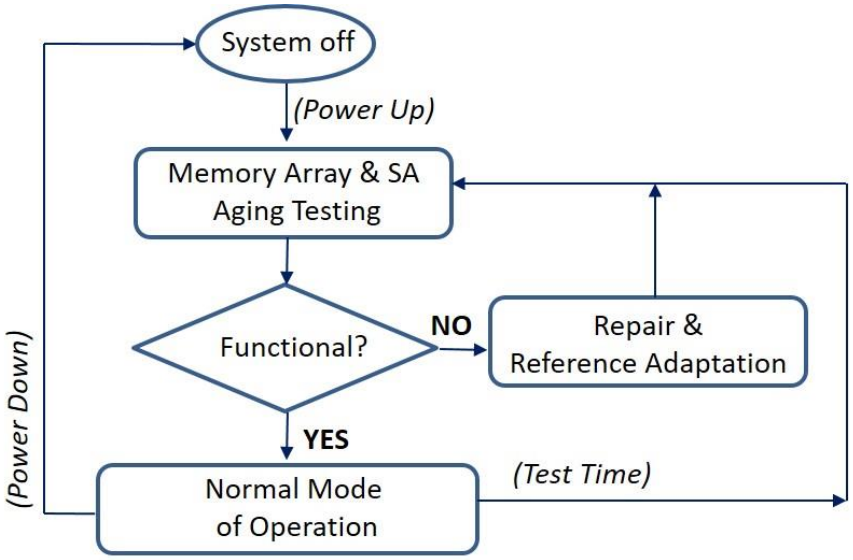

Figure 4.14 Monitoring scheduling.

Moreover, after each system power-up, the memory operating status, due to earlier repair actions, must be retrieved. According to the adopted monitoring procedure in Figure 4.14, at the startup of the system a monitoring phase starts for testing the memory cells and sense amplifiers for the case of aging occurrences at the memory cells or sense amplifiers during previous activations of the system. Initially, the reference duty cycle values for the memory cells and the sense amplifiers are transferred from the ROM to two dedicated registers. Next, the memory testing procedures are applied. If according to the testing results all memory cells and sense amplifiers are found to be healthy, the system enters the normal operating mode; otherwise it enters the memory repairing phase. During this phase, if a cell is detected as over-aged, memory redundancy can be exploited for repair. If a sense amplifier is detected as over-aged, the memory timing is properly adjusted and a new aging tolerance margin is set (SA activation delay is increased by a predetermined step). At the same time, proper adjustment is made to the reference value in the corresponding register using a constant predetermined step.

Then, for the case that the previous repairing procedures do not confront with the actual aging status of the memory, another testing phase must follow. If needed, a second repair procedure is also applied and the pertinent duty cycle ratio reference values are re-adjusted. The procedure is repeated until all SRAM sense amplifiers and memory cells are reported as healthy (functional) by the testing procedures in accordance to the actual reference values of the duty cycle ratio. As the SRAM has completed the self-healing process, the system enters the normal mode of operation.

4.4.7. Manufacturing Testing Operations

The proposed unified aging monitoring technique can be also applied as part of the manufacturing testing procedures in the fab for the testing of the ring oscillator, the memory cells and the sense amplifier modules. The ring oscillator can be exercised by applying a slightly modified version of the proposed method previously described. For that purpose, the signals *SA_EN* and *Arr_AC* are left deactivated while the *Tst_EN* signal is set active. Thus, the memory cells or the sense amplifier do not affect the oscillation signal generated by the ring oscillator. This allows the identification of defective or over-skewed (due to transistor mismatches) ring oscillators. Then, in order to maintain the aging monitoring capability and the SRAM reliable operation the column of the defective ring oscillator is replaced by a spare one.

After the testing of the ring oscillators, the testing of the memory cells and the sense amplifiers follows. The proposed aging monitoring scheme is exploited as previously presented for the detection of over-skewed memory cells or sense amplifiers with unacceptable offset voltage levels, which are related to local process variation induced transistor mismatches. Over-skewed memory cells or sense amplifiers are identified and the memory can be repaired according to the discussion in Section 4.4.5. As a result, having completed the above testing

procedures, any potential misfunctionalities related to local process variations either in the memory cells, the ring oscillators or the sense amplifiers are avoided.

4.4.8. Overall assessment

The proposed unified approach for the periodic aging monitoring aims to allow the self-healing of the SRAM by detecting over-aged (near failure) SRAM modules (cells or sense amplifiers) with the use of a simple, low cost, differential ring oscillator embedded in the SRAM circuit. A monitoring mode is introduced and the duty cycle of the oscillator is used for the detection of excess aging degradation in the sense amplifiers or memory cells. When operating in normal mode, the ring oscillator is kept inactive and detached from the remaining circuitry so that the impact of aging on the monitoring scheme is alleviated. This topology can be reused during testing procedures at the fab in order to characterize and test the SRAM.

Upon detecting an over-aged (near generating failures) sense amplifier or memory cell, the presented scheme provides the ability to react and repair the SRAM operation in a proper manner in order to achieve its self-healing and maintain the reliable operation of the memory.

4.5. Simulation Results

In order to validate the proposed aging monitoring techniques, a plethora of simulations were performed. The first subsection that follows, focuses on the simulation results of the unified aging monitoring approach for the memory cells and the SAs, as well as its performance and power consumption effects. Next, the simulation results of the alternative technique for aging monitoring of the SAs and the pertinent performance and power consumption effects are presented.

4.5.1. Evaluation of the Unified Aging Monitoring Approach for Memory Cells and SAs

The proposed unified monitoring scheme has been validated by simulations on the memory bit-line slice topology in Figure 4.13 that has been designed in the 90nm CMOS technology of UMC (V_{DD}=1V), using the CADENCE Virtuoso platform and the SPECTRE simulator respectively.

The designed bit-line slice consists of 256 memory cells according to typical SRAM standards [119]. The transistors' widths for the Dff-D cell, the switches (SCL, SCR, SSL, SSR) and the sense amplifier are presented in Table 4.1.

Table 4.1 Transistor widths of the unified approach

Cell	Transistor	Width
SRAM cell (Figure 3.5)	MCP1, MCP2	120nm
	MCN1, MCN2	140nm
	MCN3, MCN4	120nm
Dff-D cell (Figure 4.2)	all pMOS	120nm
	MDN1, MDN2	200nm
	MDN3	335nm
Latch-type SA (Figure 3.9)	MSP1, MSP2	400nm
	MSP3	800nm
	MSN1, MSN2	1425nm
	MSN3	960nm
Switches SCL, SCR, SSL, SSR (Figure 4.13)	all pMOS	800nm
	all nMOS	800nm

The read response delay time of the "fresh" memory, without the monitoring circuitry and in the typical conditions, with respect to the word-line (*WL*)

activation is 92.80ps. This time corresponds to the time interval from the activation of a word-line until the SA can provide its response to one of its outputs.

As indicated in [83], in the cross couple topology of the memory cell or the sense amplifier core and considering typical memory workloads, a pair of diagonal transistors, are maintained in relaxation, whereas all the other transistors face stress (asymmetric aging). Without loss of generality, in the presented simulations, transistors MCP1 and MCN2 of the memory cell in Figure 3.5 are considered to age. On the other hand, MCN1 and MCP2 transistors are considered to be in relaxation. Accordingly, transistor aging is considered for MSP1, MSN2, MSP3 and MSN3 of the sense amplifier core (SAC) in Figure 3.9, while transistors MSP2 and MSN1 remain in relaxation and V_t degradation does not influence them. It is obvious that the opposite condition is equivalent. In the simulations, a voltage source of proper polarity is used at the gates of the aged transistors for the modeling of aging [120] as illustrated in Figure 4.15 where its voltage level is equal to the induced threshold voltage shift $|\Delta V_t|$ (aging level).

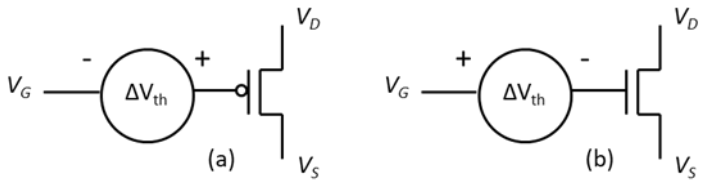

Figure 4.15 Sub-circuit simulation models for (a) NBTI in the pMOS transistor and (b) PBTI in the nMOS transistor.

Efficiency of the Monitoring Circuit for the Memory Cells

As we mentioned in Section 3.3.1, transistor aging affects the memory noise margins. Considering that the Static Noise Margin (SNM) is the maximum amount of noise that the cell can tolerate before its data is corrupted and its state is changed, simulations were applied to obtain the static characteristic curves of the

cross coupled inverters of the cell (butterfly curve). The SNM was measured as the size (side) of the minimum square between the two largest squares that were fit in the two eyes of the curves. With the use of the butterfly curve the cell stability was explored [121].

Below, the SNM for both the hold and the read operation is calculated. A controlled voltage source V_n is used as a noise source at one of the cell's inputs which is swept from 0V (gnd) to 1V (V_{DD}) as illustrated in Figure 4.16.

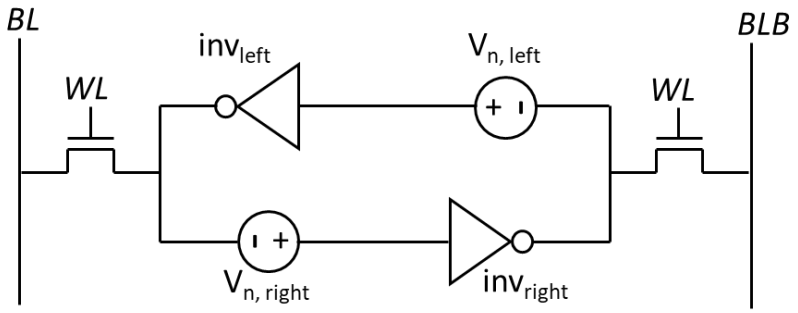

Figure 4.16 The cross-coupled inverter pair with static noise sources $V_{n,left}$ and $V_{n,right}$.

The read and hold SNMs are determined as follows:

- Read SNM: is determined by enabling the word-line to connect the memory cell internal nodes to the pre-charged at V_{DD} bit-lines, and the V_n is swept from the ground to V_{DD}. The V_n that flips the cell gives the Read SNM (RSNM).

- Hold SNM: is determined by disabling the word-line to isolate the memory cell internal nodes from the bit-lines, and the V_n is swept from the ground to V_{DD}. The V_n that flips the cell gives the Hold SNM (HSNM).

Aging induced V_t degradation disturbs the inverters' static characteristic curves and reduces the RSNM and HSNM. In Figure 4.17 the RSNM for a "fresh" and an aged cell with $|\Delta V_t|$ equal to 60mV is presented.

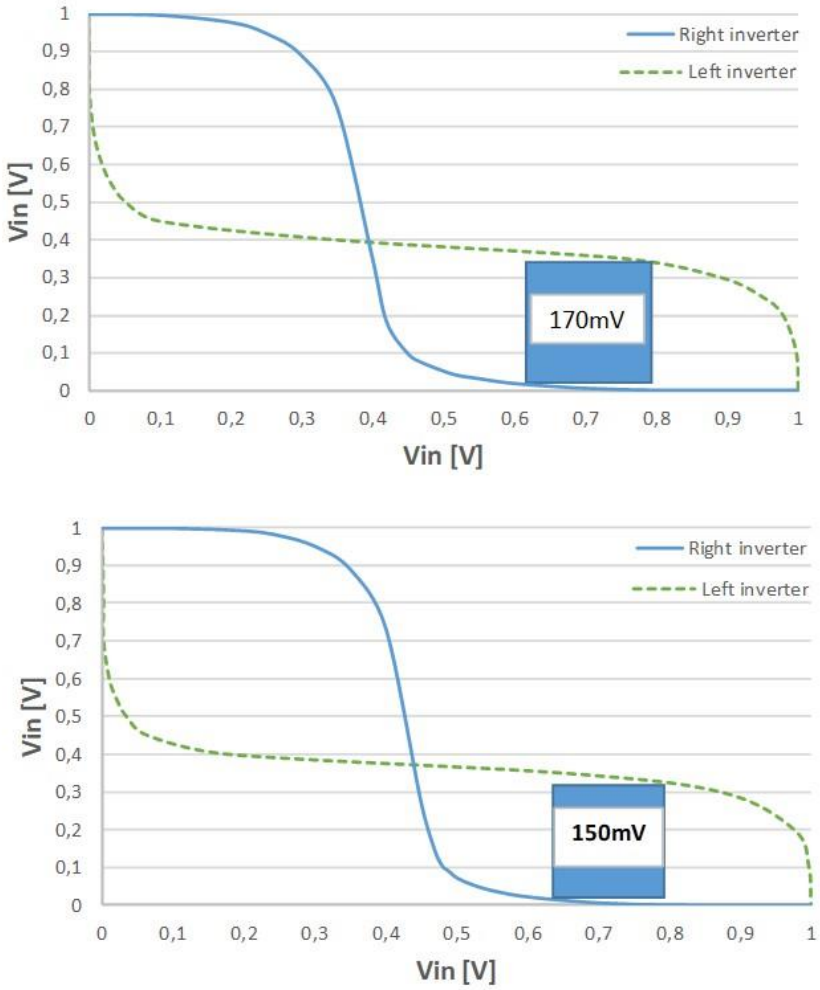

Figure 4.17 HSNM measurement for **(a)** a fresh cell and **(b)** an aged cell (for $|\Delta Vt|$ = 60mV).

Next, in Figure 4.18 the HSNM for a "fresh" and an aged cell with $|\Delta V_t|$ equal to 60mV is presented accordingly.

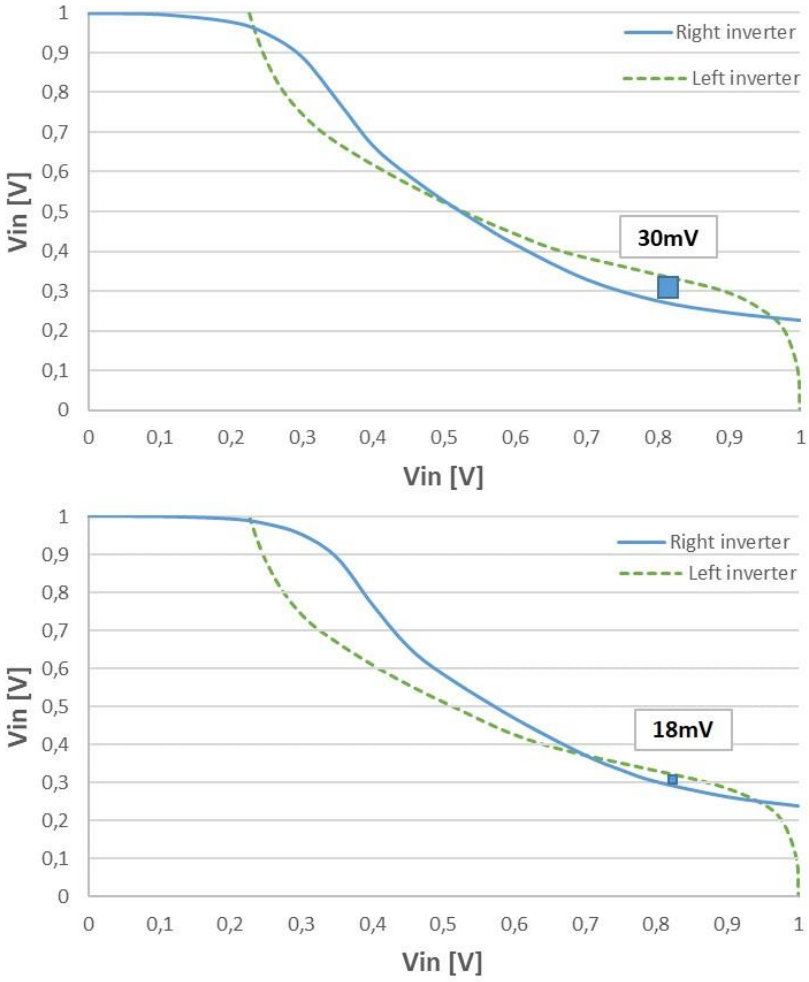

Figure 4.18 RSNM measurement for **(a)** a fresh cell and **(b)** an aged cell (for |ΔVt| = 60mV).

For the case of a memory cell, the ratio of the DRO oscillation signal (high to low ratio of this signal during the monitoring mode) with respect to the transistor threshold voltage shift ($|\Delta V_t|$) for the transistors MCP1 and MCN2 in Figure 3.5, at the typical process conditions, is presented in Figure 4.19. We can easily observe that the ratio of a skewed cell ($|\Delta V_t| > 0$) has a clear deviation with respect to a "fresh" cell ($\Delta V_t = 0$), which permits the detection of aged memory cells.

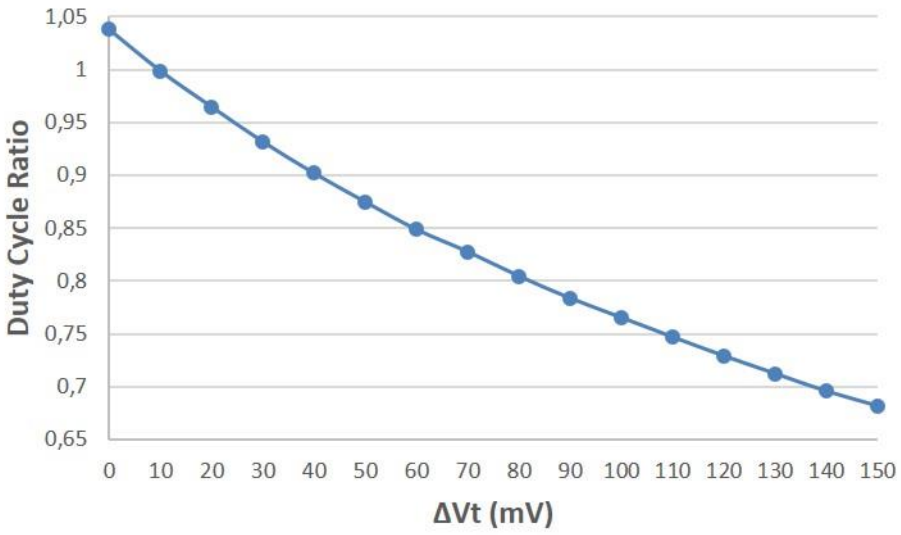

Figure 4.19 Duty cycle ratio vs V_t degradation (ΔV_t shift) of the memory cell.

The following Figure 4.20 presents the HSNM and RSNM degradation of the memory cell as the ΔV_t increases (from ΔV_t=0 for a cell with fresh transistors to ΔV_t=100mV for a cell with over-aged transistors). These curves indicate a significant reduction of the corresponding noise margins under the presence of progressive aging. More specifically, for the case of ΔV_t=100mV the HSNM presents a reduction of 17,6% while the RSNM presents a 100% reduction with respect to a fresh memory cell.

Taking into consideration the curve presented in Figure 4.19, the duty cycle of the DRO is reduced by 26,26% with respect to a fresh cell, when ΔV_t=100mV for a cell with over-aged transistors. Combining all these measured values, Figure 4.21 presents the relation of the HSNM and RSNM evolution with the duty cycle evolution under various aging conditions (from ΔV_t=0 to ΔV_t=100mV).

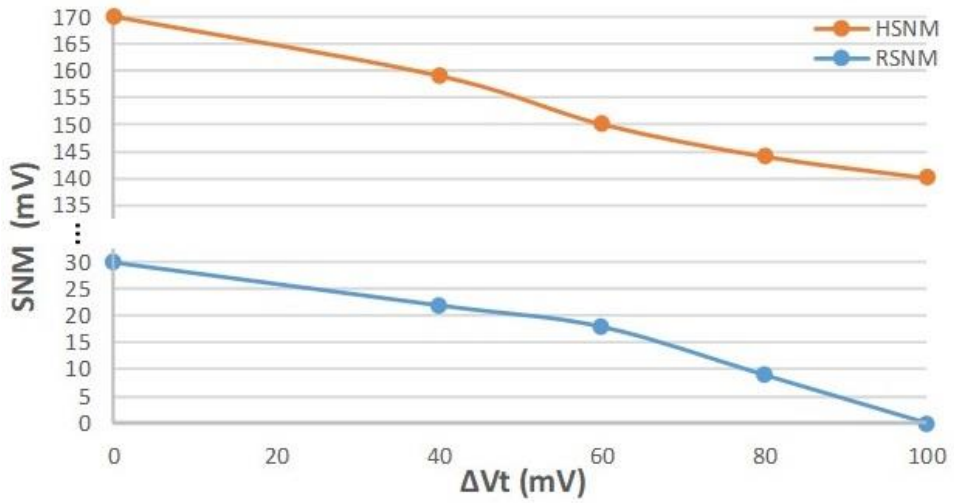

Figure 4.20 HSNM and RSNM vs V_t degradation (ΔV_t shift).

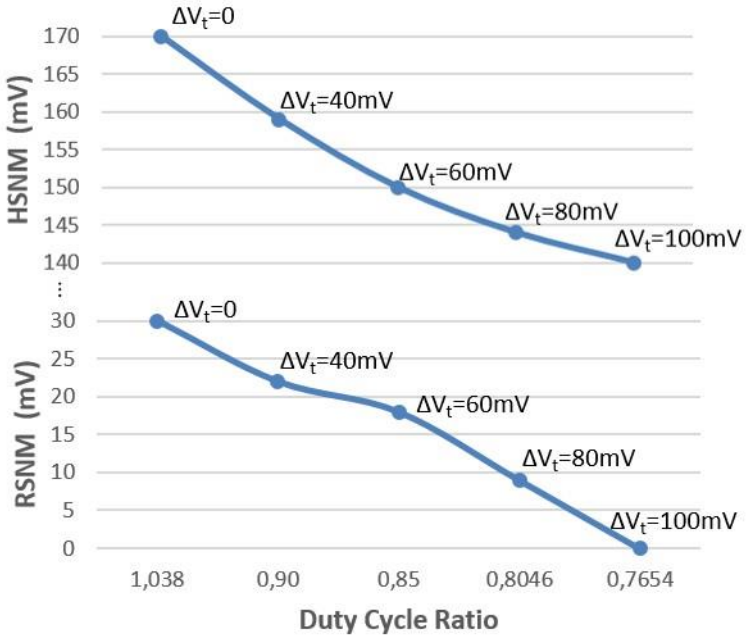

Figure 4.21 HSNM and RSNM vs Duty Cycle ratio under aging.

The waveform of the oscillation signal that is generated by the monitoring circuit for a "fresh" memory cell (without any transistor degradation - no transistor threshold voltage shift exists - $\Delta V_t=0$) is presented in Figure 4.22(a). Furthermore,

in Figure 4.22(b) the signal waveform is shown for the case of an aged memory cell (transistor degradation exists due to an absolute threshold voltage shift equal to $|\Delta V_t|$=60mV for the two indicated transistors). It is obvious from these figures that the duty cycle of the generated signal is decreased in the presence of a V_t shift.

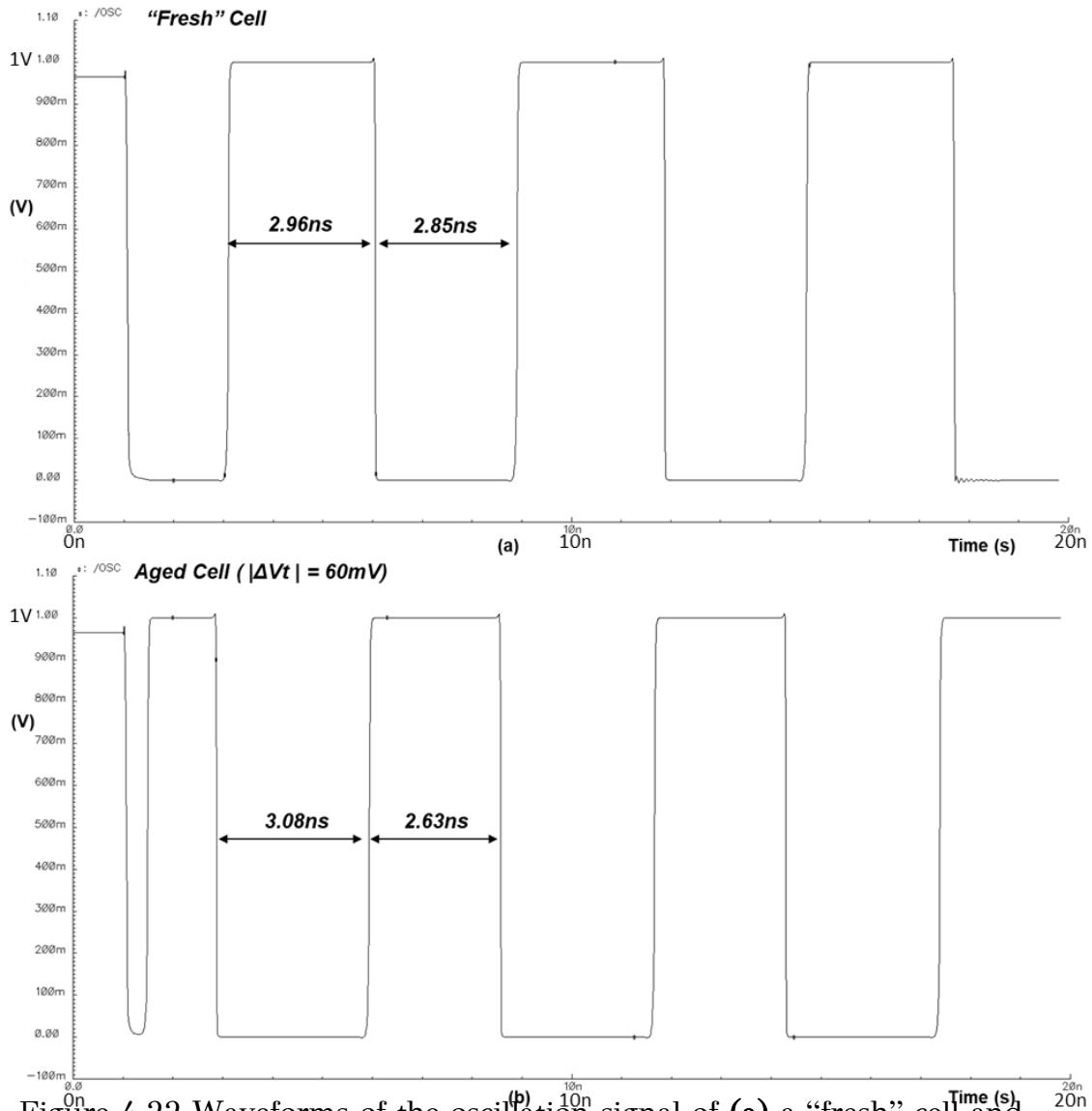

Figure 4.22 Waveforms of the oscillation signal of **(a)** a "fresh" cell and **(b)** an aged cell ($|\Delta Vt|$ = 60mV).

The aging of the memory cells should be also examined under the presence of process, temperature and supply voltage variations. The statistical models of the

used technology are exploited in order to conduct statistical (Monte Carlo) analysis aiming to explore the influence of process variations. 1000 Monte Carlo runs were applied for the testing procedure of a memory cell, for the case of "fresh" cells and the case of aged cells with $|\Delta V_t|$=150mV. The oscillation signal duty cycle ratio distribution appears in Figure 4.23. As it is expected, with transistor aging the distribution shifts towards duty cycle ratios of lower values. Furthermore, we observe that a number of the "fresh" cells present a serious skew because of the presence of transistor mismatches (time-zero variability). These fresh cells should be also repaired, since they will also lead to failure generation.

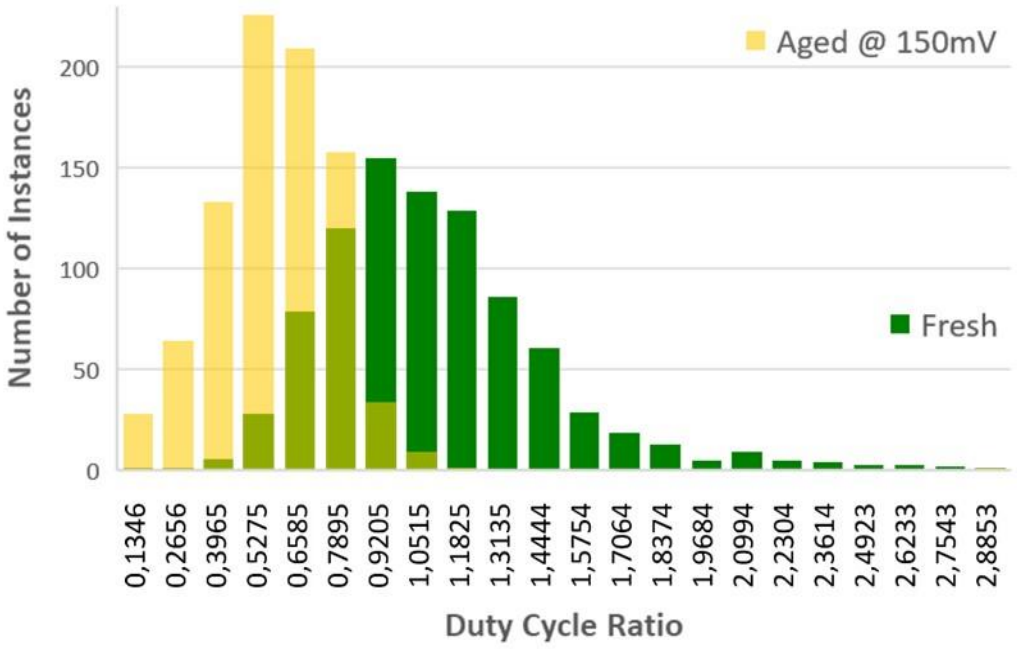

Figure 4.23 Process variation related circuits' distribution for a "fresh" and an aged (|ΔVt| = 150mV) SRAM cell (the center value of the bins is presented).

Considering Figure 4.23, a limit on the duty cycle ratio (reference value of the duty cycle) is defined at the manufacturing stage according to the SRAM specifications, the statistical analysis of measurements on the fabricated circuits and possibly simulation data as well as the skew which is considered as tolerable.

This reference value is digitized and then stored in a small ROM in the memory. In case that during a monitoring session the duty cycle ratio from a memory cell is below this reference value, the cell is characterized as over-skewed (either if the cell is "fresh" or aged) and a repair procedure must be applied to ensure the reliable operation of the memory.

Figure 4.24 and Figure 4.25, present the influence of supply voltage variations (a ±10% variation over the nominal value is considered) and temperature variations respectively, on a "fresh" and an aged memory cell with a threshold voltage shift equal to 60mV and 150mV. In both cases there is enough headroom for aging detection under variations. It should be noted here that since the monitoring procedures take place during time intervals where the system is inactive, the levels of voltage variability are not expected to be high.

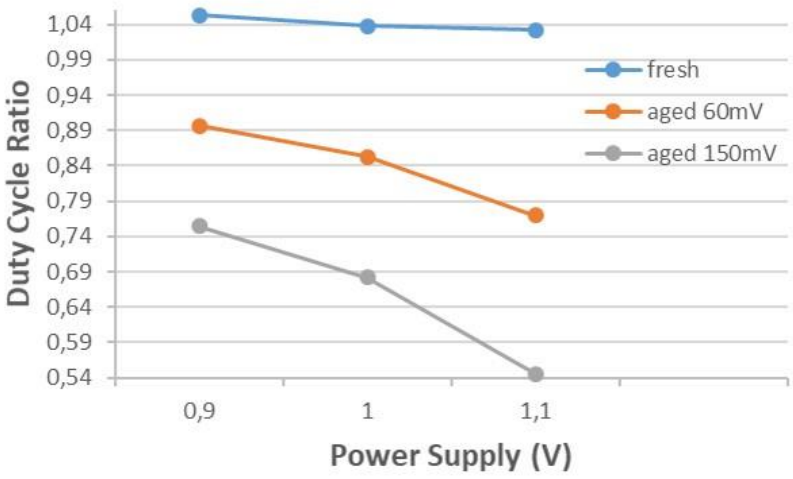

Figure 4.24 Voltage variation effects on the duty cycle ratio.

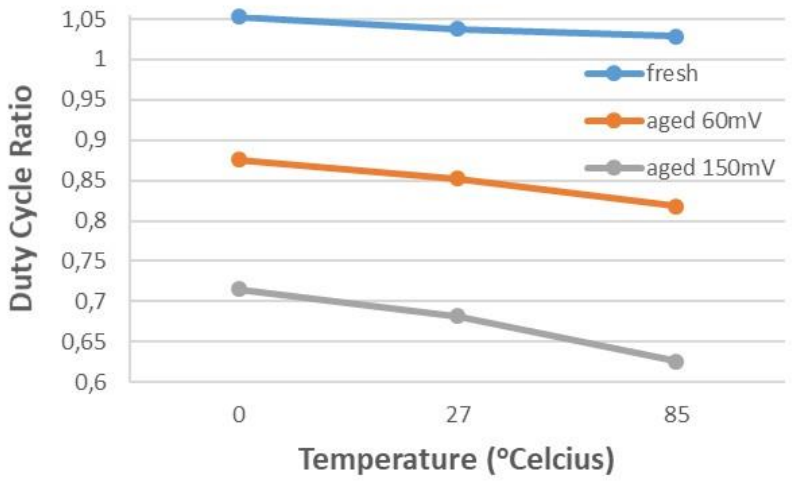

Figure 4.25 Temperature variation effects on the duty cycle ratio.

Efficiency of the Monitoring Circuit for the Sense Amplifiers

In order to explore how the operation of the sense amplifier is affected by aging, its input offset voltage was measured under various conditions of threshold voltage shift for the MSP1, MSN2, MSP3 and MSN3 transistors of the sense amplifier core.

Towards this direction, for every threshold voltage degradation level, the minimum voltage difference on the bit-lines was measured so that the correct response would be provided by the sense amplifier (independently of the required time). The measured voltage difference corresponds to the input offset voltage. Figure 4.26 presents the increment of the input offset voltage according to the transistors' threshold voltage shift $|\Delta V_t|$.

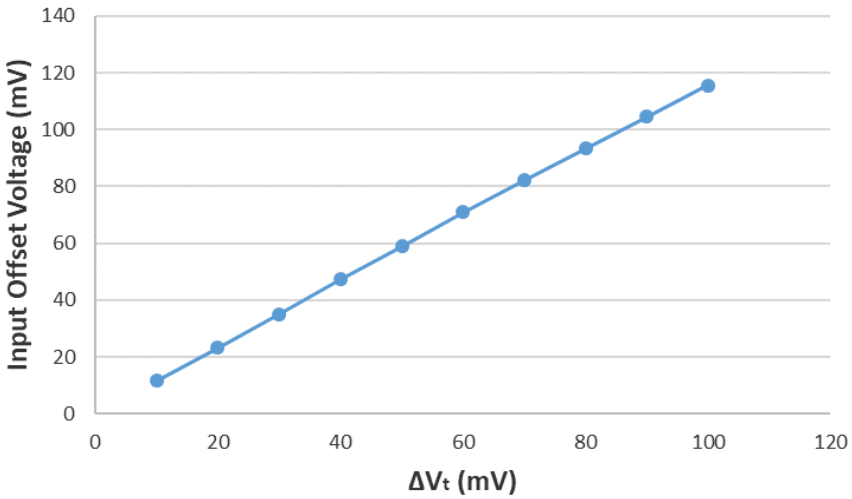

Figure 4.26 Sense amplifier input offset voltage vs V_t degradation (ΔV_t).

Furthermore, in a second simulation, the minimum bit-line voltage difference that is required on the bit-lines in order for the SA to provide its response within the same time duration, after the activation of the *SA_EN* signal, as a "fresh" SA was measured. The pertinent graph is presented in Figure 4.27. As the threshold voltage shift $|\Delta V_t|$ increases the bit-line voltage difference should also increase in order for the sense amplifier to respond within the same time duration. This means that the time interval between the world-line activation (*WL* turns to high in Figure 4.13) and the SA activation (*SA_EN* turns to high) must be increased in order to be able to establish the required voltage difference on the bit-lines.

Obviously this will result in the increment of the overall response delay of the sense amplifier. However, aiming to negate aging effects and ensure the reliable operation of the SRAM, the proper adjustment of the time interval between the word-line activation and the sense amplifier activation is a viable solution. The delay on the activation of the SA vs the transistors' threshold voltage shift $|\Delta V_t|$, which is required for the cancellation of the input offset voltage effects, is presented in Figure 4.28.

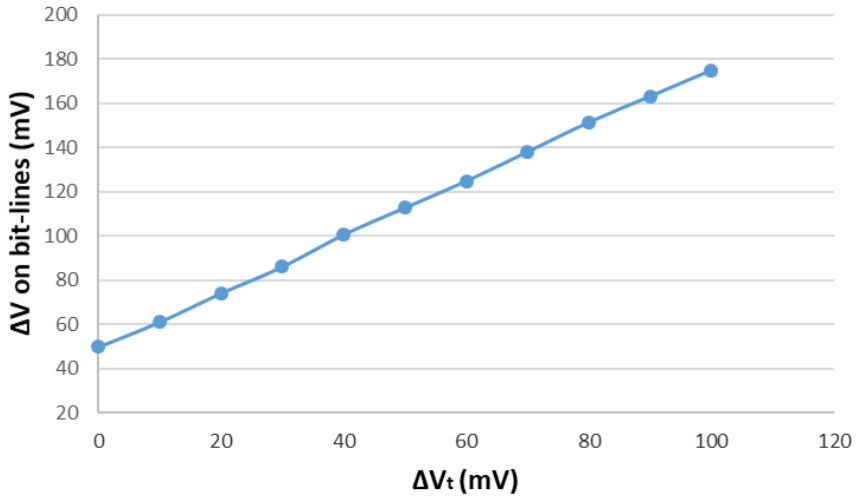

Figure 4.27 Minimum Bit-Lines voltage difference for successful read operation vs V_t degradation.

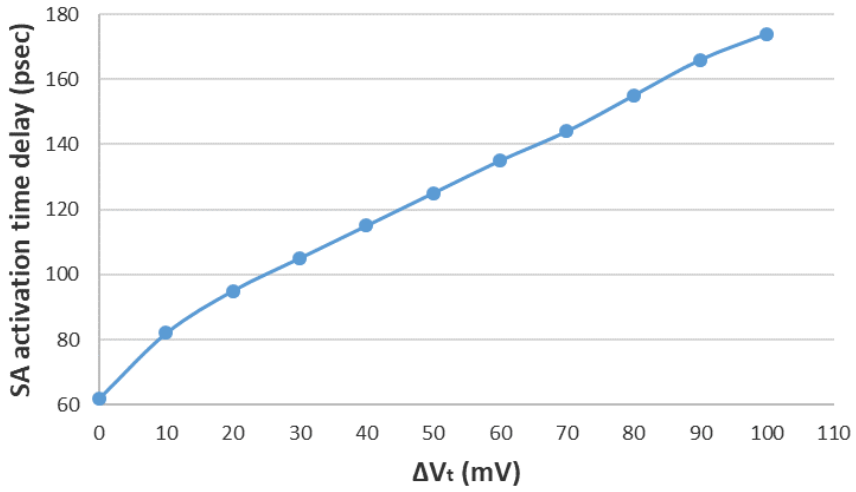

Figure 4.28 Sense amplifier activation delay for input offset voltage effects cancellation vs V_t degradation.

The plot in Figure 4.29 presents the duty cycle ratio provided by the monitoring circuit against the threshold voltage shift $|\Delta V_t|$ in the four transistors (MSP1, MSN2, MSP3 and MSN3) of the SAC for the typical process conditions. In a similar manner to the memory cell, in the sense amplifier case as the threshold voltage shift $|\Delta V_t|$ increases the duty cycle ratio is almost linearly decreased, too.

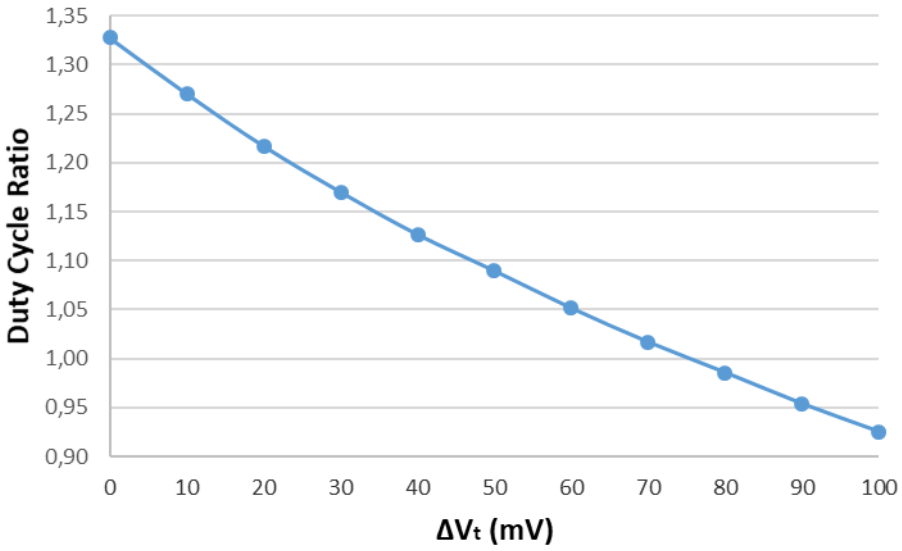

Figure 4.29 Duty cycle ratio vs sense amplifier's transistors V_t degradation (ΔV_t shift).

Combined information is exploited from the graphs in Figures 4.28 and 4.29, in order to determine the increment of the time interval between the word-line activation and the sense amplifier activation for the repairing purposes. Next, the corresponding reference value is properly adjusted (decreased) by a constant step that is predetermined at the manufacturing stage exploiting the information in Figure 4.29.

In Figure 4.30 (a) the simulated waveform of the output signal of a "fresh" SA (without any transistor degradation - no transistor threshold voltage shift exists $\Delta V_t=0$) during the monitoring mode is presented. Next, in Figure 4.30(b) the same signal waveform is illustrated for an aged SA (transistor degradation exists due to an absolute threshold voltage shift equal to $|\Delta V_t|=60mV$ for the above four transistors). From the simulations, we observe that the V_t degradation under consideration the duty cycle of the generated signal is altered.

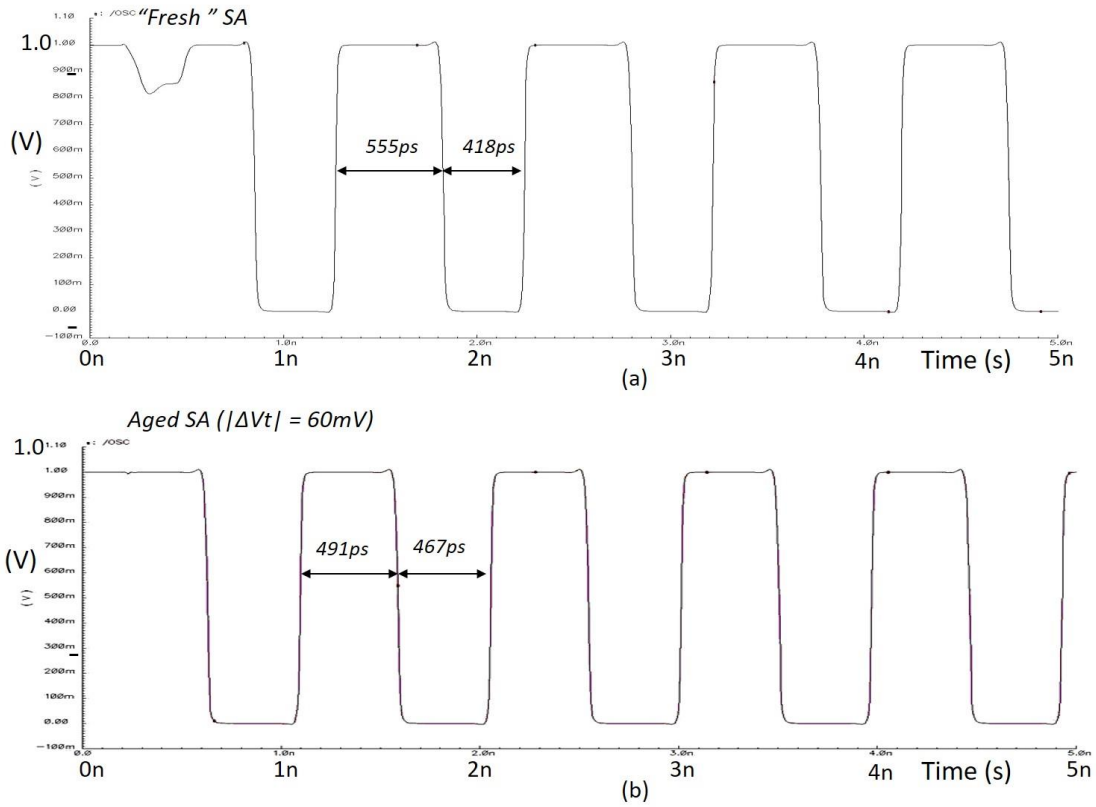

Figure 4.30 Waveforms of the oscillation signal for **(a)** a "fresh" sense amplifier and **(b)** an aged sense amplifier (|ΔVt|=60mV).

The proposed technique for the aging monitoring of the SRAM SAs has been also examined for its immunity to process, temperature and voltage variations in order to validate that it can accurately diagnose the presence of aging under all circumstances. Initially, using the statistical models of UMC, statistical (Monte Carlo) simulations have been carried out for the "fresh" as well as for various aging conditions of the SA and the monitoring circuitry. In Figure 4.31, the circuits' distribution with respect to their duty cycle ratio is presented after 1000 Monte Carlo runs, both for the "fresh" case as well as for the 150mV aging condition. As expected, with the aging of the sense amplifier, the distribution is shifted towards duty cycle ratios of lower value. The results are similar to these for the case of the memory cells.

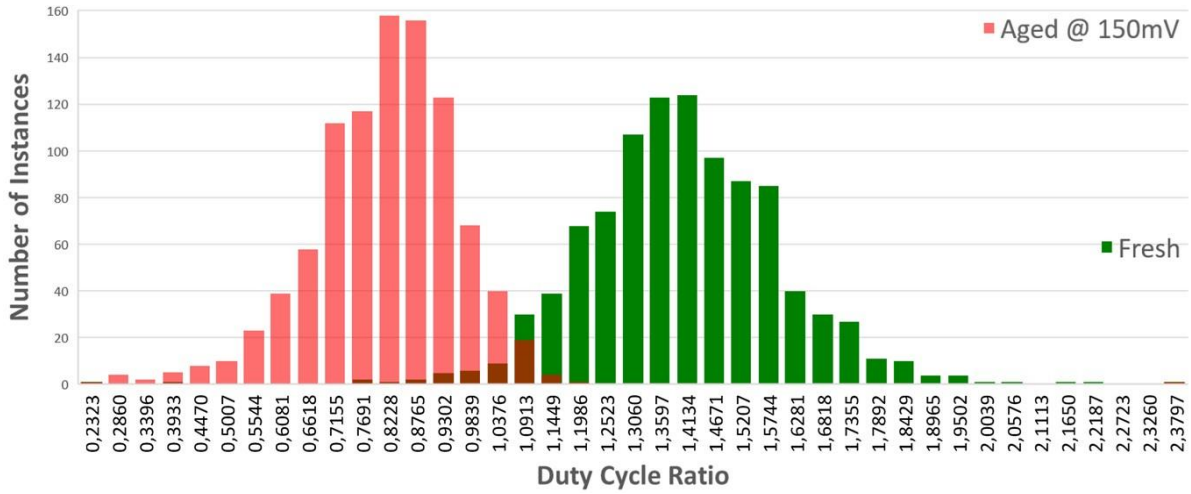

Figure 4.31 Process variation related circuits' distribution for a "fresh" and an aged ($|\Delta V_t|$ = 150mV) sense amplifier (the center value of the bins is presented).

Due to transistor mismatches (time-zero variability), a portion of the "fresh" SAs will be seriously skewed. Depending on the memory specifications, the statistical analysis of measurements on fabricated circuits and possibly simulation data as well as the acceptable aging tolerance defined from the manufacturing stage, a duty cycle ratio limit (initial reference duty cycle value) is specified. This reference value, after digitization, is stored in a simple and very small ROM, like these that support the repair operations in an SRAM. An SA ("fresh" or not) with a duty cycle ratio lower than this reference value is considered over-aged and a repair operation must be applied according to the procedures discussed in Chapter 4.2.5. The reference duty cycle corresponds to an SA that is aged enough although it does not generate failures during the SRAM operation. Moreover, it stands that an over-aged SA with a duty cycle ratio which is just outside this reference limit does not also generate failures but can be detected by the periodic monitoring procedure. Further aging of this SA will soon result to the appearance of failures and consequently the reliability reduction.

For further validation of the proposed monitoring circuitry, the influence of voltage variations (±10% of the nominal value) and temperature variations has

been also explored and the results are presented in Figure 4.32 and Figure 4.33 respectively, for the "fresh" SA and aged SAs with a threshold voltage shift of 60mV and 150mV. We should mention here that a monitoring procedure is applied at idle time intervals of system inactivity so that we do not expect a high voltage variability during the monitoring phase.

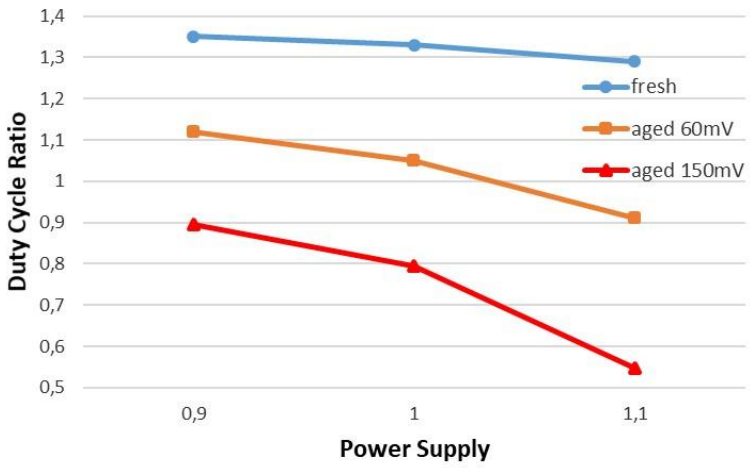

Figure 4.32 Voltage variation influence on the sense amplifier monitoring procedure.

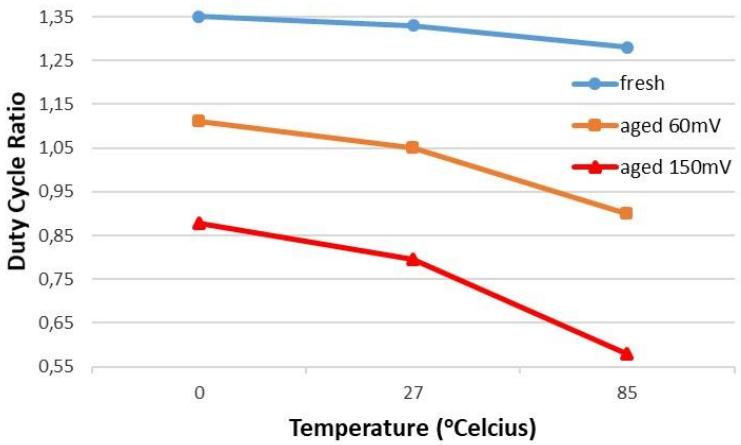

Figure 4.33 Temperature variation influence on the sense amplifier monitoring procedure.

Efficiency of the DRO

In the normal mode of operation, the switches SSL, SSR, SCL and SCR (in Figure 4.13) are off and the DRO is isolated from the rest circuitry and power-gated from both power supplies. Thus, due to leakage currents its internal nodes are stabilized at intermediate voltage levels between the two power supplies; V_{DD} and gnd. As a consequence a low, homogenous, voltage stress affects the ring oscillator's transistors (MDP1, MDP2 and MDN1, MDN2 in Figure 4.2) in a way that they age very slowly. It should be noted also that in the normal mode of operation MDP3 and MDN3 transistors are not stressed. Even though there might be a slight effect on the frequency of the ring oscillator's signal, the homogenous voltage stress does not alter the signal's duty cycle. Consequently, the ring oscillator is capable of efficiently supporting the aging monitoring task for all the lifetime duration of the SRAM.

For the validation of this statement, the duty cycle ratio of the DRO for various (homogenous) ΔV_t shifts is presented in Figure 4.34. Even for unexpected ΔV_t values the variation is shown to be too small (e.g. for ΔV_t= 100mV it is only 3.7%).

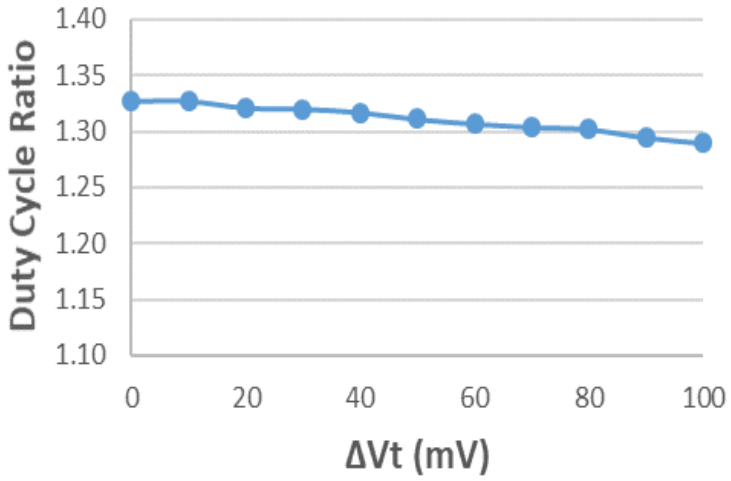

Figure 4.34 Duty cycle ratio of the oscillation signal vs V_t shift (ΔV_t).

Efficiency of the Digitizer

The efficiency of the used Digitizer block (see Figure 4.4), mainly depends on the frequency of the system clock signal (*CLK*) that is used for the sampling of the generated oscillation signal as well as the measurement time during the monitoring sessions. A quantization error is introduced by this procedure. Simulation results are presented in Figure 4.35, where the measured by the Digitizer duty cycle ratio is compared with the expected ratio for various threshold voltage shift values.

The Digitizer design is in the same 90nm technology of UMC as the bit-line slice. Both the HC and LC are 10-bit counters. The circuit is fed with the SA output signal (*BLO*). The used clock frequency is 3GHz while the oscillation signals under monitoring are around 1GHz. As we observe, the maximum quantization error introduced by the Digitizer is 3.82%. This error can be further reduced by increasing the number of bits of the LC and HC counters and/or the clock frequency.

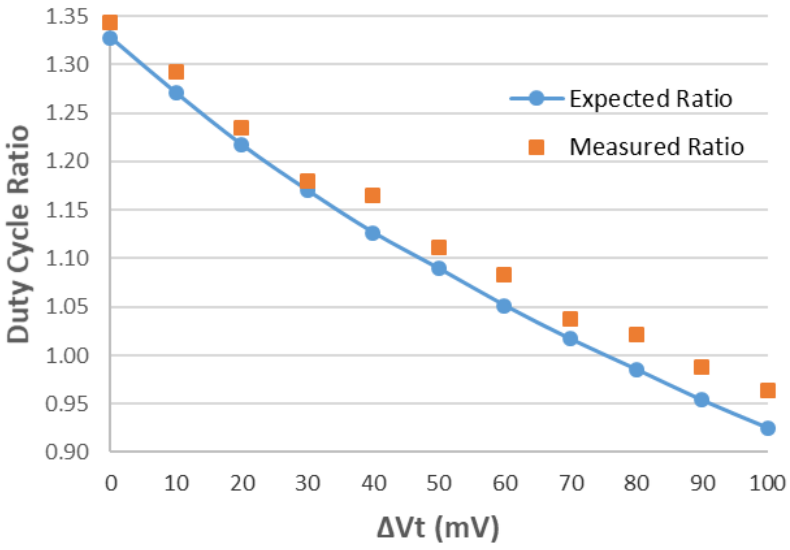

Figure 4.35 Voltage variation effects on the duty cycle.

Performance and Power Consumption Effects

According to the simulations, during normal mode, the impact of the monitoring circuitry on the response delay time of the SA is 2.80%. As the ring oscillator is detached from the bit-lines and the sense amplifier when operating under normal mode, this low degradation on the speed is a result of the additional parasitic capacitance of the switches SCL, SCR, SSL and SSR (Figure 4.13) that are attached on the bit-lines and the inputs of the sense amplifier core.

In terms of power consumption, the Dff-D cells and the Digitizer are inactive during the normal mode of operation so that their impact is negligible. The only influence is related to the small parasitic capacitance that is inserted on the bit-lines due to the presence of the four switches of the DRO. The power consumption increment during the normal mode is measured to be only 1.24%.

The silicon area cost of the DRO circuit is approximately 5.5% per bit-line slice, which incorporates the precharge circuit, 256 memory cells, the access transistors and the sense amplifier (the write circuitry and the decoders are not included in this estimation). Given that two memory blocks share a common sense amplifier, the actual area overhead is smaller than 2.75%. For memory arrays of larger size, the overhead is less than the one mentioned. It should be also noted that the whole memory uses a single Digitizer of quite small area overhead.

Based on the simulation results presented, the following pros of the proposed technique are considered:

- the normal SRAM operation is not altered,
- only slight silicon area overhead is added (less than 2.75% in the memory array),
- the impact on the SA response delay time during the read operation is only 2.8% while during the write operation the influence is negligible,
- the power consumption is increased only by 1.24%.

4.5.2. Evaluation of the Alternative Aging Monitoring Technique for SAs

In order to validate the proposed aging monitoring scheme, the memory bit-line slice in Figure 4.10 was designed and simulated in the 90nm CMOS technology of UMC (V_{DD}=1V), using the CADENCE Virtuoso platform and the SPECTRE simulator. It consists of 256 word-lines. The SA under consideration in these simulations is the pMOS cross-coupled amplifier as presented in Figure 3.10. The transistors' widths for the Dff-D cell, the access transistors used and the SA are presented in Table 4.2. The response delay of a "fresh" SA in the typical conditions is 42ps. Aging induced to the transistors of the SA is considered as an increase of the threshold voltage shift |ΔVt|. In the simulations, a voltage source of proper polarity is used at the gates of the aged transistors for the modeling of aging [120] as previously presented in Figure 4.15, where its voltage level is equal to the induced threshold voltage shift $|\Delta V_t|$ (aging level).

Table 4.2 Transistor widths of the alternative monitoring technique

Cell	Transistor	Width
Dff-D cell (Figure 4.2)	all pMOS	300nm
	all nMOS	200nm
PCCEQ SA (Figure 3.10)	MSP1, MSP2	360nm
	MSP3	120nm
	MSN1, MSN2	120nm
	MSN3	180nm

In Figure 4.36(a) the simulated waveform of the *OSC* signal for the rDRO in the first configuration (Config-1) is presented. Next, in Figure 4.36(b) the same signal waveform is illustrated for the rDRO in the second configuration (Config-2) with a "fresh" SA (without any transistor degradation, that is, no transistor threshold voltage shift exists ΔV_t=0). Finally, in Figure 4.36(c) the *OSC* signal waveform is shown for the rDRO in the second configuration (Config-2) with an aged SA

(transistor degradation exists due to an absolute threshold voltage shift equal to $|\Delta V_t|$=50mV for both pMOS and nMOS transistors). From the simulations we observe that under V_t degradation the oscillation frequency of the ring oscillator is decreased.

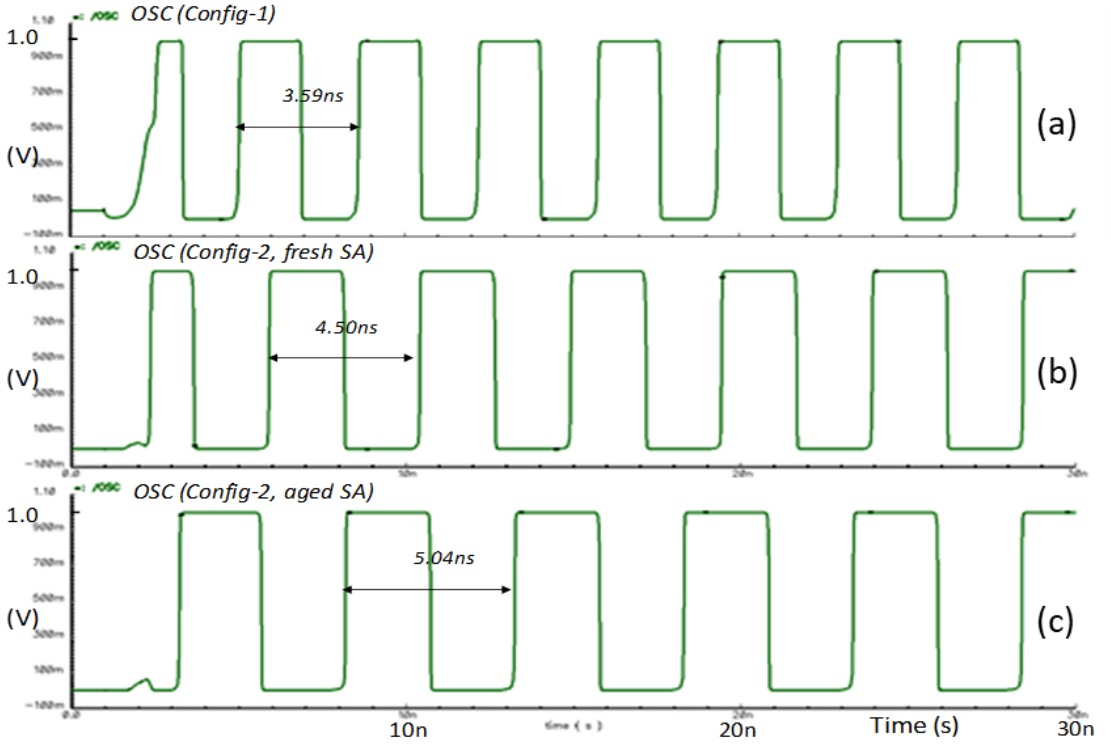

Figure 4.36 Simulated Waveforms.

The simulations were repeated for various transistor threshold voltage degradation levels in the SA. The results are illustrated in the plot of Figure 4.37. In this figure, the frequency ratio is plotted against the threshold voltage shift $|\Delta V_t|$ in the transistors of the SA. As the threshold voltage shift $|\Delta V_t|$ increases, the ratio also linearly increases. Depending on the acceptable aging tolerance defined from the manufacturing stage, by selecting a reference ratio that corresponds to a specific $|\Delta V_t|$, allows the detection of over aged SAs in order to further adjust the SA activation delay time and retain the SRAM reliable operation.

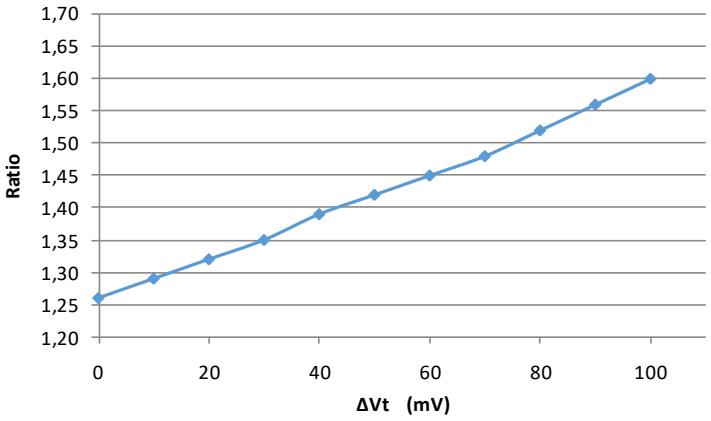

Figure 4.37 Frequency ratio vs V_t degradation ($|\Delta V_{tl}|$ shift).

4.5.2.1. Performance and Power Consumption Effects

According to the simulations, the impact of the monitoring circuitry on the memory access time is 4.3%. In addition, the power consumption of the Dff-D cells in the normal mode of operation is negligible, since the transistors MDL and MDR are inactive (*Tst_EN* to low) and the rDRO is inactive and isolated from the rest of the circuit. The silicon area cost of the ring oscillator is equal to 7.81 SRAM cells per pair of bit-line slices since it is a common practice an SA to be shared between two memory arrays. For a bit-line slice of 256 cells (this is a total of 512 cells per SA) the silicon area overhead of the monitoring circuitry is only 1.9%.

Finally, it should be noted that a single Digitizer is exploited for the whole memory array so that the silicon area cost is small especially for large memory arrays.

CHAPTER 5

PROPOSED AGING MONITORING TECHNIQUE FOR SRAM DECODERS

In this Chapter we present an embedded circuit for the aging monitoring in SRAM Decoders along with an approach to react for memory repair after the detection of an over-aged Decoder in order to achieve its self-healing.

5.1. Aging Monitoring for SRAM Decoders

Similarly to the memory cells and the sense amplifiers in an SRAM array, it has been proven that Address Decoders are also affected by BTI and HCI phenomena, the presence of which plays an important role to the overall performance degradation of the memory. As Address Decoders age, delay faults appear during their operation usually leading to read or write failures. Thus, it is imperative to turn to self-healing solutions considering the Address Decoders, as well. That is, to develop aging-tolerant design techniques that will provide the ability to sense aging levels on SRAM Address Decoders, predict upcoming failures and react to

retain the reliable operation of the SRAM memory. A BTI and HCI aging monitoring scheme is proposed next for SRAM row Decoders along with a repairing approach aiming to retain the reliable operation of the SRAM.

5.1.1. The monitoring circuitry

The circuit proposed for monitoring the performance degradation of the Decoders' transistors aims to early diagnose over-aged Decoders and properly react in order to achieve self-healing and prolong the lifetime of the SRAM. Although the analysis that follows focuses mainly on row Decoders, the proposed technique can be also applied to column Decoders.

Without loss of generality, in the analysis that follows we consider the m-to-2m NAND type dynamic logic row Decoder topology (m input address bits to 2m rows) previously presented in Figure 3.3 of Section 3.2.4. It is composed of 2^m slices and each slice consists of m+1 nMOS transistors, a precharge transistor plus an inverter which provides the word-line activation signal (*WL*).

The proposed scheme for aging monitoring suggests the addition of a simple, low cost embedded circuit consisting of two extra word-line slices along with a Counter, a MUX, a simple Test Controller and a Comparator. In Figure 5.1 the overall architecture of the proposed scheme is presented.

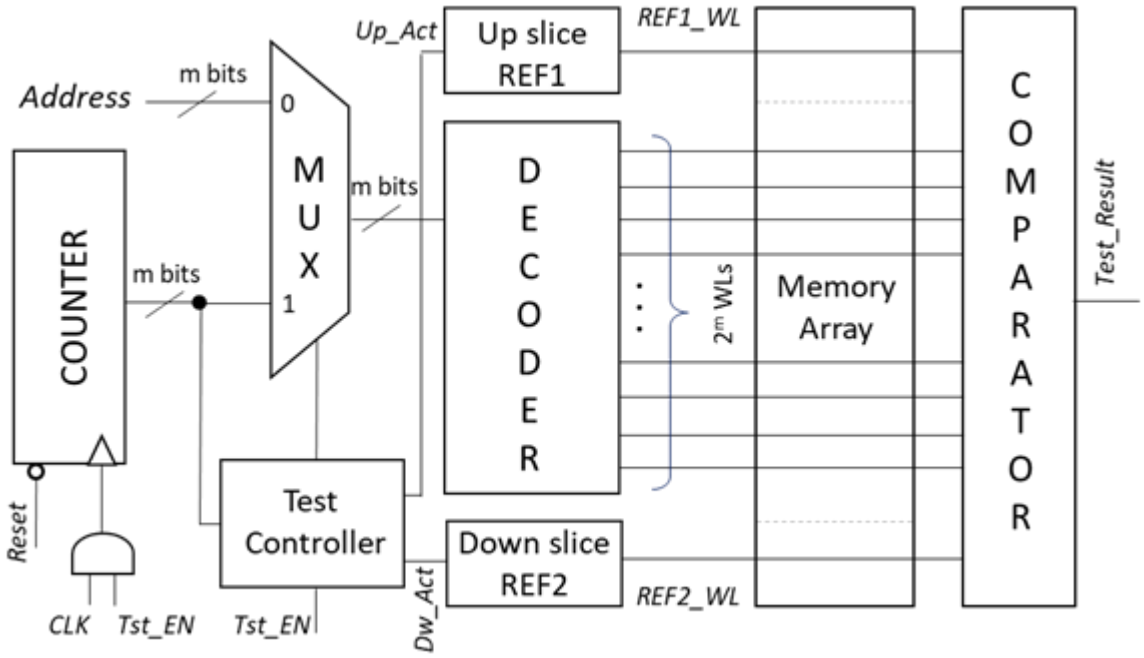

Figure 5.1 The overall architecture of the proposed decoding scheme.

According to our observations, as a Decoder undergoes continuous stress and its transistors begin to age, the time needed for the activation of a word-line tends to increase. Based on this status, in order to achieve aging monitoring of the Address Decoders, two additional word-lines are inserted in the memory structure as references of the word-line activation time that corresponds to a "fresh" Decoder. As a result, the comparison between the activation time of a word-line and that of a reference word-line will allow the identification of an over-aged Decoder.

According to Figure 5.1, the two additional word-line slices are attached one at the top (Up slice REF1) and one at the bottom (Down slice REF2) sides of the Decoder and the Memory Array. The circuit enters the monitoring mode of operation with the exploitation of the test enable signal *Tst_EN* activated at high. During this mode, an m-bit Counter generates all possible addresses for the activation of each word-line and the generated addresses feed the Decoder through a Multiplexer (MUX). Both the MUX and the extra word-line slices are controlled by a Test Controller, which is a simple state machine. When the *Tst_EN* is high

the address generated by the Counter feeds the Decoder and selects one word-line for testing; in addition, one of the two reference word-lines REF1, REF2 is activated, either by the *Up_Act* or the *Dw_Act* signals that are generated by the Controller. In the monitoring mode, the signals of the two activated word-lines (the reference and the one under test) are propagated to the Comparator. The *Test_Result* signal, which is the response of the Comparator, is used for the discrimination of over-aged Decoder slices. In our monitoring scheme, only one additional signal, the *Tst_EN,* is required. It should be noted that during the monitoring mode, the memory is in the read mode of operation. Thus, when a word-line is activated, the pertinent memory cells are read. However, no value is written to the cells and thus their contents are not altered.

5.1.2. The Comparator

The architecture of the Comparator is presented in Figure 5.2. As seen in the Figure, 2^m diode connected access transistors MWA_i (where i= 0 to 2^m-1, plus two transistors for the additional word-line slices), a XOR gate, two clear transistors MCl_1 and MCl_2, four reset transistors MR_1 - MR_4 and a flip-flop compose the Comparator. A diode connected nMOS transistor (MWA_i) is attached to each word-line in order to permit the access to a XOR gate only when the word-line is activated to high, by the pertinent input address (Figure 5.2). A XOR gate is exploited to compare the response of the reference word-line (*REFi_WL*) with the response of the word-line under test (WL_i) in order to evaluate the time delay between them and identify over-aged Decoder slices. When a slice of the Decoder has not aged, the two responses reach the inputs of the XOR gate at the same time and their comparison generates a logic '0' at the XOR output, indicating that the Decoder's slice remains in a "fresh" state. It should be noted that each address generated by the Counter is propagated simultaneously to the whole Decoder and thus no delays are expected between the activation of the slice under monitoring and the reference slice.

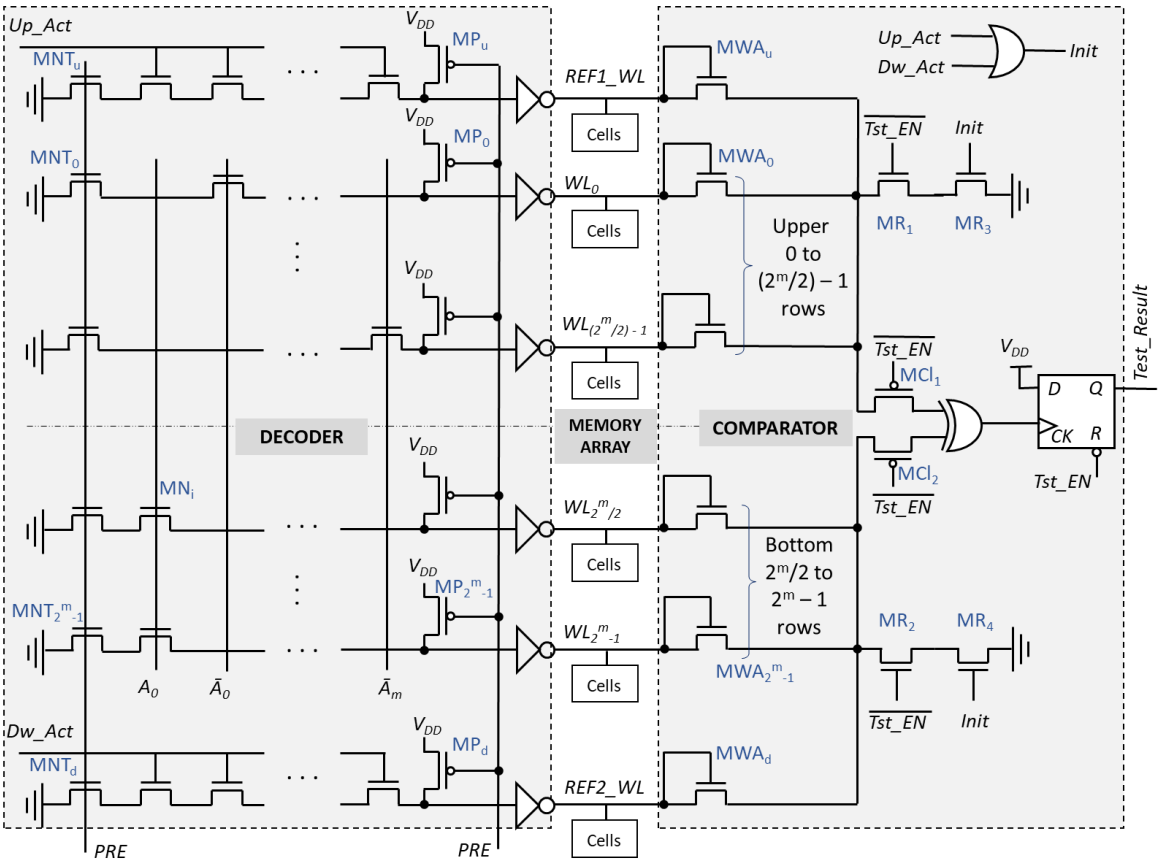

Figure 5.2 Decoder, Memory Array and Comparator.

The output of the XOR gate feeds the clock input *CK* of a flip-flop. The *D* input of the flip-flop is permanently connected to V_{DD}. When the aging has caused a significant delay between the two responses, the XOR output turns to high and triggers the flip-flop so that the *Test_Result* signal is set to high indicating that the Decoder slice is over-aged. It should be noted that the crucial transistors of both reference word-line slices remain in a "fresh" state as it will be further discussed. Both inputs of the XOR gate and the flip-flop output are initialized to zero (reset) before a Decoder's slice testing. For the initialization, the *Tst_EN* signal should be low while both *Up_Act* and *Dw_Act* signals should be high.

The output of the upper reference word-line *REF1_WL* is compared to the generated output of a selected word-line (WL_i) from the lower half Decoder slices

$(2^m)/2$, while the output of the lower reference word-line *REF2_WL* is compared to the generated output of a selected word-line (WL_i) from the upper half Decoder slices.

The SRAM mostly operates in the normal mode of operation, as the duration of the aging monitoring activity is too small. During this mode signals *Tst_EN*, *Up_Act* and *Dw_Act* are set to low, deactivating this way the reference slices REF1, REF2. Thus, the nMOS network of the reference Decoder slices stay inactive and the pertinent transistors are not under stress and remain in a "fresh" state. The reset paths through transistors MR_1 - MR_4 are also off in the normal mode. The clear transistors MCl_1 and MCl_2 also block any potential current leakage toward the XOR gate when a word-line is activate during normal mode of operation. As presented in Figure 5.3, the NOT gates included within the XOR gate are tri-stated gates, keeping the internal nodes of the XOR gate floating when in normal mode, thus the transistors of this gate do not undergo any stress leading to aging.

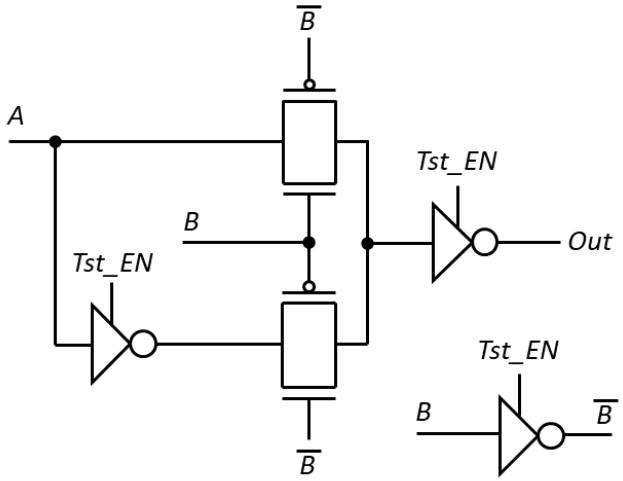

Figure 5.3 The XOR gate.

As the diode connected transistors MWA_i of the Comparator face AC stress when a word-line is activated in normal mode (WL_i='1'), the topology can be altered according to Figure 5.4.

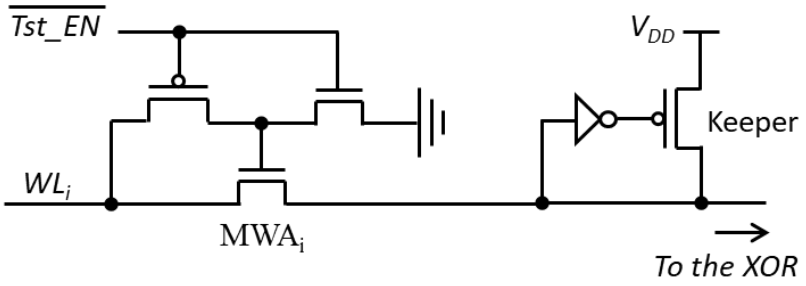

Figure 5.4 Alternative scheme of the diode connected nMOS transistors MWA_i of the Comparator.

Two minimum size transistors (one nMOS and one pMOS) per line are added for the elimination of MWA_i transistor aging during normal mode. Both transistor are fed by the complement of the *Tst_EN* signal. In the monitoring mode of operation the pMOS transistor is on enabling the formation of the diode connected structure, while in the normal mode the nMOS is on to set the MWA_i transistor in the non-conducting state. The silicon area cost introduced by this scheme is low. Finally, a keeper structure can be exploited in order to ensure that each input of the XOR gate during the monitoring mode will receive a robust logic '1' when the pertinent word-line will be activated. It should be noted, that a single keeper is added for each of the two inputs of the XOR gate.

5.1.3. The Memory Monitoring and Repair Approach

In the monitoring mode, the Counter activates one after the other the Decoder slices and the signal of the corresponding word-line is propagated to the Comparator. In the case where the activated slice is from the upper half of the Decoder, the REF2 slice is also activated by the Controller (while when the lower half of the Decoder is activated then the REF1 slice, is also activated) and its output is propagated to the Comparator, as well. The two signals are compared in order to detect possible excess aging degradation on the Decoder slice under test. Given that the transistors of the Decoder slice in question have been seriously affected by aging, the output of the examined slice will reach the Comparator

significantly delayed. In that case, the XOR gate will generate a robust pulse at its output which triggers the *CK* signal of the flip-flop (Figure 5.2) and sets the *Test_Result* signal to high indicating the presence of excess aging and the need for proper memory repair. When the slice is not overaged the pulse that may be generated by the XOR gate is not capable to trigger the flip-flop and the output of the Comparator will be low. After the evaluation of a Decoder's slice, the test logic is initialized and the next slice is evaluated and so on until the evaluation of all slices. The signals timing diagram for the testing of a slice during the monitoring mode is presented in Figure 5.5.

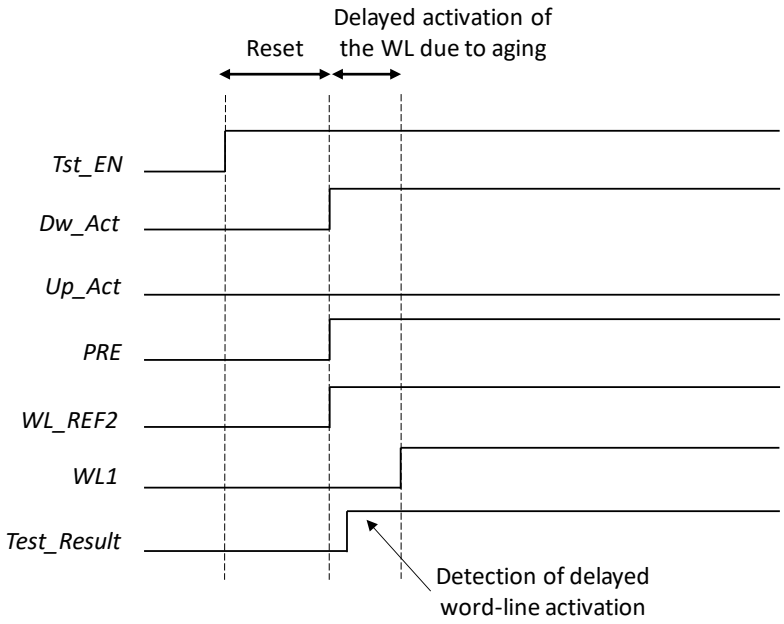

Figure 5.5 Timing Diagram.

For the repair of the SRAM and for achieving its self-healing in case of an overaged Decoder slice, existing spare memory rows can be exploited to replace the affected ones, as it is common in memory self-repair operations [58], [114]-[116]. Possibly, in modern SRAMs extra spare memory rows can be considered by design for aging alleviation. The monitoring process can be periodically applied at the start-up of the system or during idle times, in order to early detect the

presence of aging and properly react in order to ensure the reliable operation of the SRAM. Alternatively, the monitoring can be applied at the end of predetermined time intervals, such as the refresh operations in DRAMs, but as previously discussed, at greater periods of time. Such a periodic operation will not affect the SRAM's performance or this of the system where it belongs.

5.2. Simulation Results

The above aging monitoring scheme has been validated by simulations on the Decoder topology in Figure 5.2 that has been designed in the 90nm CMOS technology of UMC (V_{DD}=1V), using the CADENCE platform. We considered a Decoder with an address space of 10-bits (1024 word-lines) and 64 memory cells per word-line. In the simulations, a voltage source of proper polarity is used at the gates of the aged transistors for the modeling of aging [120] where its voltage level is equal to the induced threshold voltage shift $|\Delta V_t|$ (aging level).

Transistor aging (e.g. threshold voltage shift $|\Delta V_t|$) is considered for the transistors in the nMOS network (MN_i) of a Decoder slice as well as the pMOS transistor of the corresponding inverter in Figure 5.2. These transistors are crucial from the speed performance point of view. Aiming to explore the influence of aging on the Decoder operation, the time needed for the activation of the word-line WL_i was measured under various threshold voltage shift conditions for the above transistors. For every threshold voltage shift, the minimum activation delay time of the *WLi* was measured and the results are presented in Figure 5.6.

From the simulations performed, we observed that when V_t degradation passes above 50mV, the XOR gate generates a robust pulse which triggers the flip-flop indicating the time violation due to the aging. Below this limit, the delay of the Decoder is considered to be within the typical time margins of the memory. When the Decoder is under 50mV V_t shift in typical conditions, the delay added to the word-line for its activation compared to that of a fresh state is 13,16psec.

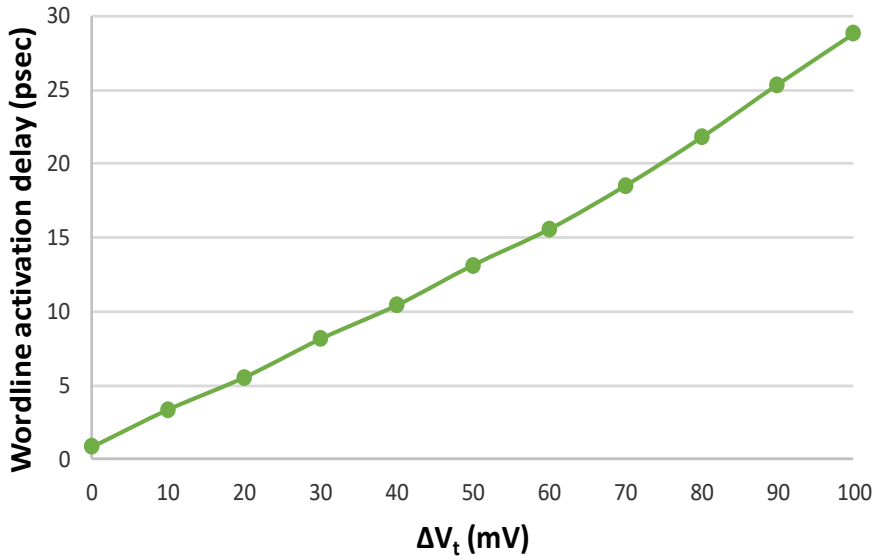

Figure 5.6 Word-line activation delay vs transistors' V_t degradation.

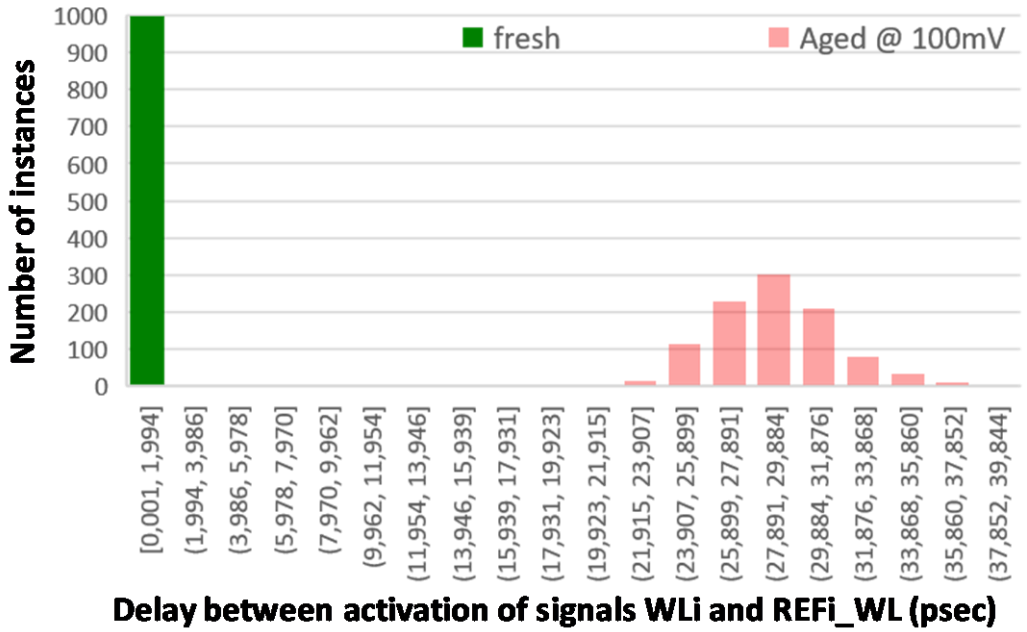

Figure 5.7 Process variation related circuits' distribution for a "fresh" and an aged ($|\Delta V_t|$ = 100mV) Decoder.

The aging of the Decoder was next examined under the presence of process variations. The statistical models of the used technology were exploited in order to conduct statistical (Monte Carlo) analysis. 1000 Monte Carlo runs were applied working in the monitoring mode, for the case of "fresh" Decoder slices and the case of an aged Decoder slice with $|\Delta V_t|$=100mV.

The distribution of the delay between the activation time of the word-line WL_i signal of the slice under consideration and the activation time of the *REFi_WL* signal of the reference slice is presented in Figure 5.7. As it is expected, with transistor aging the distribution shifts towards higher delay values. The extra delay for the word-line activation caused by aging may lead to false readings or writings on the SRAM.

5.2.1. Performance and Silicon Area Cost

The impact of the monitoring circuitry on the speed performance of the Decoder is too small and stems from the presence of the MUX, which is a typical block in commonly used memory built-in self test schemes and thus can be re-used by other test procedures. From the simulations performed it was measured that the delay of the Decoder in the presence of the proposed Comparator is 213.13psec, while its delay without the addition of the Comparator is 203.02psec.

According to these measurements, the additional delay for the activation of the word-line in the presence of the proposed Comparator (due to the additional parasitic capacitances) is only 10.11psec (the delay is increased by 4.99%), thus the induced speed performance degradation is too small. The power consumption of the Comparator in the normal mode of operation is negligible. Furthermore, the dynamic power consumption of the REF1 and REF2 slices in the normal mode of operation is negligible, since they are inactive.

The silicon area cost of the two reference word-line slices is also negligible as only two rows are added per Decoder of 2^m rows. The same applies for the cost of the Comparator which incorporates only one transistor per word-line, a single XOR gate with four reset transistors, an OR gate and one flip-flop. Should the altered topology of the diode connected transistor be used, two more transistors (one pMOS and one nMOS) per word-line plus two keepers are added to the Comparator's cost. The total silicon area cost of the monitoring circuitry is measured to be 1.18%, which incorporates the two additional reference word-lines and the Comparator (including the altered nMOS diode-connected network and the Keepers).

It should be noted, that the two reference Decoder slices are considered to drive two of the existing dummy word-lines at the borders of the memory array and thus, no further cost is added to the array. For memory arrays of greater size than 1024 x 64 cells, the overhead is less than the one mentioned. The area cost for the rest additional circuitry (Controller, Counter and MUX) can be shared among various testing schemes as it is a typical and reusable circuitry for SRAM embedded testing.

The transistors' widths of the proposed additional elements for the aging monitoring of a 10-bit Decoder are illustrated in the following Table 5.1.

Table 5.1 Transistor widths for the monitoring of a 10-bit Decoder

	Transistor	Width
2 Reference Lines (Figure 5.2)	Precharge pMOS	240nm
	nMOS network	2640nm
	NOT Gates	2880nm
Comparator (Figure 5.2)	Diode connected nMOS	246.240nm or 240nm/word-line (incl. the 2 reference lines)
	Reset transistors (MR_1-MR_4)	480nm
	Clear Transistors (MCl_1-MCl_2)	480nm
	XOR Gate	1680nm
	Flip Flop	4080nm
Additional cost due to the altered diode-connected transistor MWA_i	Additional pMOS	246.240nm or 240nm/word-line (incl. the 2 reference lines)
	Additional nMOS	123.120nm or 120nm/word-line (incl. the 2 reference lines)
	2 Keepers	1200nm

5.3. Overall Assessment

Aging phenomena such as Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI) have an important impact on the reliable operation of nanometer

technology SRAMs, as they affect the performance characteristics of the memory blocks. Aiming to provide SRAM self-healing options, aging prediction techniques are mandatory in order to timely react, before errors generation, and maintain the memory reliable operation. The presented scheme suggests an embedded circuit for the aging monitoring in SRAM Decoders along with an approach to react for memory repair after aging detection. The proposed low cost and easy to implement embedded architecture consists of two additional word-line slices, a Counter, a Test Controller, a MUX and a Comparator.

The contribution of the proposed monitoring scheme is summarized as follows:

- it offers early diagnosis of an over-aged Decoder
- it provides the ability of early reaction upon the diagnosis of an over-aged Decoder
- there is no alteration of the memory normal operation and the influence on its performance characteristics is reduced
- it introduces only 1.18% additional silicon area cost to the total array structure and
- it provides the ability to avoid the aging of the monitoring circuitry when the SRAM is operating in normal mode.

CHAPTER 6

CONCLUSIONS

Reliability is one of the major design challenges in designing robust CMOS circuits and systems. The reliability issue is exacerbated as the CMOS devices scale down towards nanometer technologies. In the nanometer era, process variations and aging phenomena lead to a degradation in the performance and reliability of an electronic system, hence limiting its expected lifetime. Variation-aware design techniques, such as conservative safety margins, can be re-used to reduce the impact of aging on system reliability; however, the applications of such methods make it harder to develop competitive products and may lead to the elimination of performance gains of technology scaling. Therefore, there is a need for innovative approaches to improve the resilience of an integrated circuit to aging-induced failure without affecting its performance.

The reliable operation of nanometer technology static random-access memories (SRAMs) crucially impacted by process variations and aging mechanisms like Bias Temperature Instability (BTI) and Hot-Carrier Injection (HCI), since the performance characteristics of the memory cells, the used sense amplifiers and the decoders are affected. As a result, aging prediction techniques are mandatory in order to timely react, before errors generation, and maintain the memory reliable operation.

The contributions of this dissertation in the area of SRAM memory testing, on-line testing, monitoring and self-healing are summarized as follows:

Initially, an aging-tolerance oriented design technique was developed for individual periodic aging monitoring on SRAM memory cells and sense amplifiers along with repairing mechanisms for the alleviation of the aging effects. Then, the technique was properly modified in order to compose a unified scheme achieving the aging monitoring of both memory cells and sense amplifiers in a memory array. The proposed technique, based on the use of a simple, low cost, embedded Differential Ring Oscillator (DRO) is able to sense aging levels and predict upcoming failures in the memory and provides the ability to the system to early react in order to retain the reliable operation of the memory.

The features of the proposed unified aging monitoring scheme are:

- the ability to early react for memory repair after the detection of an over-aged module (memory cell or sense amplifier) that is near failure,
- the reduced influence on the performance (speed and power) of the memory; without altering the normal operation of the memory,
- the reduced silicon area overhead,
- the ability to avoid the aging of the pertinent monitoring circuitry during the normal mode of SRAM operation and
- the ability to reuse it for manufacturing memory characterization and testing.

Furthermore, an embedded circuit for the aging monitoring in SRAM Decoders was developed and an approach to react for memory repair after aging detection was presented. The proposed embedded architecture suggested the implementation of low cost and complexity additional circuitry, consisting of two additional wordline slices, a Counter, a Test Controller, a MUX and a Comparator.

The features of the proposed aging monitoring scheme once again are:

- the early diagnosis of an over-aged Decoder
- the early reaction upon the diagnosis of an over-aged Decoder
- the reduced influence on the performance (speed and power) of the memory; without altering the normal operation of the memory,
- part of the monitoring circuitry is in compliance with existing SRAM testing techniques,
- the ability to avoid the aging of the monitoring circuitry when the SRAM is operating in normal mode.

Our future plans are mainly focused towards the direction of developing alternative repairing mechanisms for the case of excess aging degradation on SRAM memory cells and Decoders, aiming to minimize the needs of memory redundancy to overcome the aging effects. We will also study the effects of other aging mechanisms such as the Time-Dependent Dielectric Breakdown (TDDB) which has been also indicated in literature that significantly affects the transistors of the SRAMs. In this direction aging monitoring techniques could be developed to support the self-healing of the SRAM under the presence of such mechanisms.

BIBLIOGRAPHY

[1] L-T Wang, C. E. Stroud, N. A. Touba, "System-On-Chip Test Architectures: Nanometer Design For Testability," Morgan Kaufmann Publishers, 2008.

[2] G. E. Moore, "Cramming more components onto integrated circuits," Electronics, vol. 38, no. 8, April 1965.

[3] C. H. Díaz, S-M. Kang, Ch. Duvvury, "Modeling of electrical overstress in integrated circuits," Springer, 1995.

[4] D. Ghosh, R. Sharman, H.R. Rao, S. Upadhyaya, "Self-healing systems - survey and synthesis," Decision Support Systems, vol. 42, no. 4, pp. 2164–2185, 2007.

[5] K. Khalil, O. Eldash and M. Bayoumi, "A cost-effective self-healing approach for reliable hardware systems," Proceedings IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5, 2018.

[6] P. Koopman, "Elements of the Self-Healing System Problem Space," 25th International Conference on Software Engineering (ICSE) - Workshop on Software Architectures for Dependable Systems (WADS), Oregon, USA, 2003.

[7] Z. Zhang, Q. Yao, Y. Xiaoliang, Y. Rui, C. Yan and W. Youren, "A self-healing strategy with fault-cell reutilization of bio-inspired hardware," Chinese Journal of Aeronautics, vol. 32, no. 7, pp. 1673-1683, 2019.

[8] M. Abramovici, M.A Breuer, and A.D Friedman, Digital Systems Testing and Testable Design, IEEE Press, Revised Printing, 1994.

[9] V. D. Agrawal and J. J. Danaher, "A Tutorial on Test Power", Inter. Symposium on Low-Power Electronics and Design (ISLPED), 2008.

[10] J. Rabaey, A. Chandrakasan and B. Nikolic, "Digital Integrated Circuits", 2nd Edition, Pearson Education, 2002.

[11] E. A. Amerasekera and D. S. Campbell, "Failure Mechanisms in Semiconductor Devices," John Wiley & Sons, London, United Kingdom, 1987.

[12] B. W. Johnson, "Design and Analysis of Fault-Tolerant Digital Systems," Adison Welsey, 1989.

[13] V. P. Nelson, "Fault-tolerant computing: fundamental concepts," in Computer, vol. 23, no. 7, pp. 19-25, 1990.

[14] M. L. Bushnell and V. D. Agrawal, "Essentials of Electronic Testing for Digital Memory and Mixed-Signal VLSI Circuits," Kluwer Academic Publishers, 2000.

[15] T. W. Williams and N. C. Brown, "Defect level as a function of fault coverage," IEEE Transactions on Computers, vol. 30, no. 12, pp. 987-988, 1981.

[16] M. Sachdev and J.P. de Gyvez, "Defect-Oriented Testing for Nano-Metric CMOS VLSI Circuits," Springer, 2007.

[17] G. R. Case, "Analysis of actual fault mechanisms in CMOS logic gates," Proceedings Design Automation Conference (DAC), pp. 265-270, San Fransisco, 1976.

[18] J. P. Hayes, "Fault modeling," IEEE Design & Test, vol. 2, no. 2, pp. 88-95, 1985.

[19] J. A. Abraham and W. K. Fuchs, "Faults and error models for VLSI," Proceedings of the IEEE, vol. 74, no. 5, pp. 639-654, 1986.

[20] K. Bowman, et al., "Dynamic Variation Monitor for Measuring the Impact of Voltage Droops on Microprocessor Clock Frequency," Proceedings IEEE International Custom Integrated Circuits Conference (CICC), pp. 1-4, 2010.

[21] K. Bowman, et al., "Circuit techniques for dynamic variation tolerance," Proceedings ACM/IEEE Design Automation Conference (DAC), pp. 4–7, 2009.

[22] K.S. Jones, S. Prussin, and E. R. Weber, "Applied Physics A, Solids and Surfaces," Springer, vol. A 45, pp. 1-34, 1988.

[23] J.R. Black, "Electromigration - A brief survey and some recent results," IEEE Transactions on Electron Devices, vol. 16, no. 4, pp. 338–347, 1969.

[24] J.R. Black, "Electromigration Failure Modes in Aluminium Metallization for Semiconductor Devices," Proceedings of the IEEE, vol. 57, no. 9, pp. 1587–94, 1969.

[25] G. Gielen, P. de Wit, E. Maricau, J. Loeckx, J. Martin-Martinez, B. Kaczer and G. Groeseneken, "Emerging Yield and Reliability Challenges in Nanometer CMOS

Technologies," ACM/IEEE Design Automation and Test in Europe Conference, pp. 1322-1327, 2008.

[26] J. Lienig, "Introduction to Electomigration-Aware Physical Design," International Symposium on Physical Design, 2006.

[27] J.H. Stathis, S. Mahapatra, and T. Grasser, "Controversial issues in negative bias temperature instability," Microelectronics Reliability, vol. 81, pp. 244-251, 2018.

[28] N. Nicolici and B. Al-Hashimi, "Power-Constrained Testing of VLSI Circuits," Kluwer Academic Publications, Norwell, MA, 2003.

[29] D.K. Schroder and J.A. Babcock, "Negative Bias Temperature Instability: Road to Cross in Deep Submicron Silicon Semiconductor Manufacturing," Journal of Applied Physics, vol. 94, no. 1, pp. 1-18, 2003.

[30] S.P. Park, K. Roy and K. Kang, "Reliability Implications of Bias-Temperature Instability in Digital ICs," IEEE Design and Test of Computers, vol. 23, no. 6, pp. 8-17, 2009.

[31] V. Reddy, et al., "The Impact of NBTI on the Performance of Combinational and Sequential Circuits," Proceedings ACM/IEEE Design Automation Conference (DAC), pp. 364-369, 2007.

[32] K. Kang, S. Gangwal, S. Phil Park, and K. Roy, "NBTI Induced Performance Degradation in Logic and Memory Circuits: How Effectively Can We Approach a Reliability Solution?," Proceedings Asia /South Pacific Design Automation Conference (ASP-DAC), pp. 726-731, 2008.

[33] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao and S. Vrudhula, "Predictive modeling of the NBTI effect for reliable design," Proceedings IEEE International Custom Integrated Circuits Conference (CICC), pp. 189–192, 2006.

[34] T. Naphade, N. Goel, P. R. Nair and S. Mahapatra, "Investigation of stochastic implementation of reaction diffusion (RD) models for nbti related interface trap generation," IEEE International Reliability Physics Symposium (IRPS), pp. XT–5, 2013.

[35] M. Agarwal, V. Balakrishnan, A. Bhuyan, K. Kim, B. C. Paul, W. Wang, B. Yang, Y. Cao and S. Mitra, "Optimized circuit failure prediction for aging: Practicality and promise," Proceedings of IEEE International Test Conference, pp. 1-10, 2008.

[36] W. Wang, V. Reddy, A.T. Krishnan, R. Vattikonda, S. Krishnan and Y. Cao, "Compact Modeling and Simulation of Circuit Reliability for 65nm CMOS Technology," IEEE Transactions on Device and Material Reliability, vol. 7, no. 4, 2007.

[37] A. W. Strong, E. Y.Wu, R. P. Vollertsen, J. Sunea, G. L. Rosa, T. D. Sullivan and S. E. Rauch, "Reliability Wearout Mechanisms in Advanced CMOS Technologies," Institute of Electrical and Electronics Engineers, 2009.

[38] R. L. Wadsack, J. M. Soden, R. K. Treece, M. R. Taylor, and C. F. Hawkins, "CMOS IC Stuck-Open Fault Electrical Effects and Design Considerations," Proceedings of IEEE International Test Conference, pp. 423-430, 1989.

[39] J. B. Velamala, K. B. Sutaria, H. Shimizu, H. Awano, T. Sato, G. Wirth, and Y. Cao, "Compact Modeling of Statistical BTI under Trapping/Detrapping," IEEE Transactions on Electron Devices, vol. 60, no. 11, pp. 3645-3654, 2013.

[40] D. Patra, J. Zhang, R. Wang, M. Katoozi, E. H. Cannon, R. Huang, Y. Cao, "Compact Modeling and Simulation for Digital Circuit Aging," Proceedings IEEE International Custom Integrated Circuits Conference (CICC), 2018.

[41] F. Niklaus, H. Andersson, P. Enoksson, E. Kalvesten, G. Stemme, "Low temperature full wafer adhesive bonding of structured wafers," Sensors and Actuators A-Physical," Elsevier, vol. 92, no. 3, pp. 235-241, 2001.

[42] X. Zhou, H. Feng, J. K.O. Sin, "Hot Carrier Injection Effects in the Ultrashallow Body SONOS Gate Power MOSFET," IEEE Transactions on Electron Devices, vol. 60, no. 6, pp. 2008-2014, 2013.

[43] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, S. Vrudhula, "Predictive modeling of the nbti effect for reliable design," Proceedings IEEE International Custom Integrated Circuits Conference (CICC), pp. 189–192, 2006.

[44] A. Asenov, S. Kaya, J. Davies, S. Saini, "Oxide thickness variation induced threshold voltage fluctuations in decanano MOSFETs: a 3D density gradient simulation study", Superlattices and Microstructures, vol. 28, pp. 507-515, 2000.

[45] S.D. Provencher and A.R. Subramanian, "Method and system for stateful recovery and self-healing," US Patent 9 569 480 B2, 2017.

[46] G. Martinovic, I. Novak, "A combined architecture of biologically inspired approaches to self-healing in embedded systems," Proceedings IEEE International Conference on Smart Systems and Technologies (SST), pp. 17–22, 2017.

[47] H. Psaier, S. Dustdar, "A survey on self-healing systems: approaches and systems," Computing, Springer-Verlag, vol. 91, no. 1, pp. 43–73, 2011.

[48] K. Khalil, O. Eldash, A. Kumar and M. Bayoumi, "Self-healing hardware systems: A review," Microelectronics Journal, Elsevier, vol. 93, pp. 1–15, 2019.
[49] P. Franco et al., "On-line delay testing of digital circuits," Proceedings VLSI Test Symposium (VTS), pp. 167-173, 1994.

[50] M. Favalli et al., "Sensing circuit for on-line detection of delay faults," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 4, pp. 130-133, 1996.

[51] M. Nicolaidis, "Time redundancy based soft error tolerance to rescue nanometer technologies," Proceedings VLSI Test Symposium (VTS), pp. 86-94, 1999.

[52] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge, "Razor: A low-power pipeline based on circuit-level timing speculation," Proceedings 36th Annual IEEE/ACM International Symposium on Microarchitecture, pp. 7-18, 2003.

[53] Y. Tsiatouhas et al., "A sense amplifier based circuit for concurrent detection of soft and timing errors in CMOS ICs," Proceedings International On-line Testing Symposium (IOLTS), pp. 12-16, 2003.

[54] A. Paschalis et al., "Concurrent delay testing in totally self-checking systems," Journal of Electronic Testing: Theory and Applications, vol. 12, pp. 55-61, 1998.

[55] V. Huard, C. Partasarathy, C. Guerin, T. Valentin, E. Pion, M. Mammasse, N. Planes and L. Camus, "NBTI Degradation: From Transistor to SRAM Arrays," IEEE 46th Annual International Reliability Physics Symposium (PRS), pp. 289-300, 2008.

[56] T. Sato et al., "A simple flip-flop circuit for typical-case designs for DFM," Proceedings International Symposium on Quality Electronic Design, pp. 539-544, 2007.

[57] K. Bowman et al., "Circuit techniques for dynamic variation tolerance," Proceedings Design Automation Conference (DAC), pp. 4-7, 2009.

[58] F. Ahmed and L. Milor, "Online Measurement of Degradation Due to Bias Temperature Instability in SRAMs," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, no. 06, pp. 2184-2194, 2016.

[59] M. Choudhury, V. Chandra, K. Mohanram and R. Aitken, "TIMBER: Time Borrowing and Error Relaying for Online Timing Error Resilience," ACM/IEEE Design Automation and Test in Europe Conference, pp.1554-1559, 2010.

[60] M. Kurimoto et al., "Phase-adjustable error detection flip-flops with 2-stage hold driven optimization and slack based grouping scheme for dynamic voltage scaling," Proceedings Design Automation Conference (DAC), pp. 884-889, 2008.

[61] K. Hirose et al., "Delay-compensation flip-flop with in-situ error monitoring for low-power and timing-error-tolerant circuit design," Japanese Journal of Applied Physics, vol. 47, pp. 2779-2787, 2008.

[62] B. Jacob, Spencer W. NG, D.T. Wang, "Memory Systems, Cache, DRAM, Disk," Elsevier, 2008.

[63] B. Keeth and R.J. Baker, "DRAM Circuit Design: A Tutorial," IEEE Press, Series on Microelectronic Systems, 2001.

[64] A. K. Sharma, "Semiconductor Memories: Technology, Testing, and Reliability," IEEE Press, 1997.

[65] B. Prince, "Semiconductor Memories: A Handbook of Design, Manufacture and Application," John Wiley & Sons, 1996.

[66] K. Itoh, "VLSI Memory Chip Design," Springer-Verlag, Germany, 2001.

[67] A. E. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra, and M. A. Alam, "Recent issues in negative-bias temperature instability: Initial degradation, field dependence of interface trap generation, hole trapping effects, and relaxation," IEEE Transactions on Electron Devices, vol. 54, no. 9, pp. 2143–2154, 2007.

[68] P. Govind and K. Rambabu, "Statistical (M-C) and static noise margin analysis of the SRAM cells," Students Conference on Engineering and Systems (SCES), 2013.

[69] J. Singh, S. P. Mohanty, D. K. Pradhan, "Robust SRAM Designs and Analysis," Springer Science & Business Media, 2013.

[70] N. H. E. Weste, D. M. Harris, "CMOS VLSI Design," Addison-Wesley Publications, 2011.

[71] M. Nourani, A. R. Attarha, "Detecting signal-overshoots for reliability analysis in high-speed system-on-chips," IEEE Transactions on Reliability, vol. 51, no. 4, pp. 494-503, 2002.

[72] I. Agbo, M. Taouil, D. Kraak, S. Hamdioui, H. Kukner, P. Weckx, P. Raghavan and F. Catthoor, "Integral Impact of BTI, PVT Variation, and Workload

on SRAM Sense Amplifier," IEEE Transactions on VLSI Systems, vol. 25, no. 04, pp. 1444-1454, 2017.

[73] S. Pae, J. Maiz, C. Prasad, and B. Woolery, "Effect of BTI Degradation on Transistor Variability in Advanced Semiconductor Technologies," IEEE Transactions Device and Materials Reliability, vol. 8, no. 3, pp. 519–525, 2008.

[74] F. Cacho, S.K. Singh, B. Singh, C. Partasarathy, E. Pion, F. Argoud, X. Federspiel, H. Pitolet, D. Roy and V. Huard, "Hot Carrier Injection Degradation Induced Dispersion: Model and Circuit-Level Measurement," IEEE International Integration Reliability Workshop, pp. 137–141, 2011.

[75] S. Chakravarthi, A. Krishnan, V. Reddy, C. F. Machala and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," Proceedings IEEE International Reliability Physics Symposium, pp. 273–282, 2004.

[76] K. Kang, M.A. Alam and K. Roy, "Characterization of NBTI induced Temporal Performance Degradation in Nano-Scale SRAM array using I_{DDQ}," IEEE International Test Conference (ITC), p. 11.1, 2007.

[77] S.O. Toh, Z. Guo, T-J.K. Liu and B. Nikolic, "Characterization of Dynamic SRAM Stability in 45 nm CMOS," IEEE Journal of Solid-State Circuits, vol. 46, no. 11, pp. 2702-2712, 2011.

[78] J. Qin, X. Li and J.B. Bernstein, "SRAM Stability Analysis Considering Gate Oxide SBD, NBTI and HCI," International Integrated Reliability Workshop (IRW), pp. 33-37, 2007.

[79] T. Liu, C-C. Chen, J. Wu and L. Milor, "SRAM Stability Analysis for Different Cache Configurations Due to Bias Temperature Instability and Hot Carrier Injection," IEEE 34th International Conference on Computer Design (ICCD), pp. 225-232, 2016.

[80] S. Khan, I. Agbo, S. Hamdioui, H. Kukner, B. Kaczer, P. Raghavan and F. Catthoor, "Bias Temperature Instability Analysis of FinFET Based SRAM cells," Design and Test in Europe Conference (DATE), 2014.

[81] Y. Sfikas and Y. Tsiatouhas, "BTI and HCI Degradation Detection in SRAM Cells," International Conference on Modern Circuits and Systems Technologies (MOCAST), 2017.

[82] Y. Sfikas and Y. Tsiatouhas, "Variation Tolerant BTI Monitoring in SRAM Cells," IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 100-105, 2017.

[83] D. Kraak, M. Taouil, I. Agbo, S. Hamdioui, P. Weckx, H. Kukner, S. Cosemans and F. Catthoor, "Impact and Mitigation of Sense Amplifier Aging Degradation Using Realistic Workloads," IEEE Transactions on VLSI Systems, vol. 25, no. 12, pp. 3464-3472, 2017.

[84] I. Agbo, M. Taouil, S. Hamdioui, P. Weckx, S. Cosemans, P. Raghavan, F. Catthoor and W. Dehaene, "Quantification of sense amplifier offset voltage degradation due to zero- and run-time variability," IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 725–730, 2016.

[85] J. Kinseher, L. Heib and I. Polian, "Analyzing the Effects of Peripheral Circuit Aging of Embedded SRAM Architectures," Design Automation and Test in Europe Conference (DATE), pp. 852-867, 2017.

[86] S. Hamdioui, A.J. van de Goor, J.D. Reyes and M. Rodgers, "Memory test experiment: industrial results and data," IEEE Proceedings – Computers and Digital Techniques, vol. 153, no. 1, pp. 1-9, 2006.

[87] S. Hamdioui, Z. Alars and A.J. van de Goor, "Opens and Delay Faults in CMOS RAM Address Decoders," IEEE Transactions on Computers, vol. 55, no. 12, pp. 1630 – 1639, 2006.

[88] S. Khan, et al., "Impact of Partial Resistive Defects and Bias Temperature Instability on SRAM Decoder Reliability," 8th IEEE Design and Test Symposium (IDT), 2013.

[89] S. Khan, et al., "Bias temperature instability analysis in SRAM decoder," 18th IEEE European Test Symposium (ETS), 2013.

[90] S. Zafar, Y.H. Kim, V. Narayanan, C. Cabral, V. Paruchuri, B. Doris, J. Stathis, A. Callegari, M. Chudzik, "A comparative study of NBTI and PBTI in SiO2/HfO2 stacks with FUSI, TiN gates," Proceedings of VLSI Technology Symposium, 2006.

[91] M. A. Breuer, A. D. Friedman, "Diagnosis and Reliable Design of Digital Systems," Computer Science Press, 1976.

[92] E. J. McCluskey, "Logic Design Principles: With Emphasis on Testable Semiconductor Circuits," Prentice-Hall, Englewood Cliffs, 1986.

[93] C. Stroud, "A Designer's Guide to Built-In Self-Test," Kluwer Academic Publishers, 2002.

[94] N. Jha and S. Gupta, "Testing on Digital Systems," Cambridge University Press, 2003.

[95] L-T. Wang, Ch-W. Wu and X. Wen, "VLSI Test Principles and Architectures: Design for Testability," Morgan Kaufmann Publishers, 2006.

[96] B. Straka, H. Manhaeve, J. Vanneuville and M. Svajda, "A fully digital controlled off-chip IDDQ measurement unit," Proceedings Design, Automation and Test in Europe (DATE), pp. 495–500, 1998.

[97] J. M. Soden, C. F. Hawkins, R. K. Gulati and W. Mao, "IDDQ Testing: A Review," Journal of Electronic Testing: Theory and Applications, vol. 3, no. 4, pp. 5-17, 1992.

[98] D. Kraak, I. Agbo, M. Taouil, P. Weckx, S. Cosemans, F. Catthoor, W. Dehaene, and S. Hamdioui, "Mitigation of Sense Amplifier Degradation Using Input Switching," Design Automation and Test in Europe Conference (DATE), pp. 858-863, 2017.

[99] H-M. Dounavi, Y. Sfikas and Y. Tsiatouhas, "Aging Monitoring in SRAM Sense Amplifiers," International Conference on Modern Circuits and Systems Technologies (MOCAST), 2018.

[100] A. Ceratti, T. Copetti, L. Bolzani, F. Vargas and R. Fagundes, "An On-Chip Sensor to Monitor NBTI Effects in SRAMs," Springer Journal of Electronic Testing: Theory & Applications, vol. 30, pp. 159-169, 2014.

[101] X. Wang, J. Keane, T.T-H. Kim, P. Jain, Q. Tang and C.H. Kim, "Silicon Odometers: Compact In Situ Aging Sensors for Robust System Design," IEEE Micro, vol. 34, no. 6, pp. 74-85, 2014.

[102] M-C. Tsai, et al, "Embedded SRAM Ring Oscillator for In-Situ Measurement of NBTI and PBTI Degradation in CMOS 6T SRAM Array," International Symposium on VLSI Design, Automation and Test (VLSI-DAT), 2012.

[103] B. Alorda, C. Carmona, G. Torrens and S. Bota, "On-line Write Margin Estimator to Monitor Performance Degradation in SRAM Cores," IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 90-95, 2016.

[104] A. Gebregiorgis, M. Ebrahimi, S. Kiamehr, et al., "Aging mitigation in memory arrays using self-controlled bit-flipping technique," The 20th Asia and South Pacific Design Automation Conference, 2015.

[105] A. Haggag, G. Anderson, S. Parihar, D. Burnett, G. Abeln, J. Higman and M. Moosa, "Understanding SRAM High-Temperature-Operating-Life NBTI: Statistics and Permanent vs Recoverable Damage," IEEE 45th Annual International Reliability Physics Symposium (PRS), pp. 452–456, 2007.

[106] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "Impact of NBTI on SRAM Read Stability and Design for Reliability," 7th International Symposium on Quality Electronic Design, pp. 213–218, 2006.

[107] M. Rahma, Y. Chen, W. Sy, W-L. Ong, L-Y. Ting, S-S Yoon, M. Han and E. Terzioglou, "Characterization of SRAM Sense Amplifier Input Offset for Yield Prediction in 28nm CMOS," IEEE International Custom Integrated Circuits Conference (CICC), 2011.

[108] H. Santos, J. Semião, R. Cabral, A. Romão, M. B. Santos, I. C. Teixeira, J. P. Teixeira, "Aging and Performance Sensor for SRAM," Conference on Design of Circuits and Integrated Systems (DCIS), 2016.

[109] W. Needham, C. Prunty, and E.H. Yeoh, "High volume microprocessor test escapes, an analysis of defects our tests are missing," Proceedings International Test Conference 1998, pp. 25-34, 1998.

[110] A.J. van de Goor, S. Hamdioui, and R. Wadsworth, "Detecting faults in the peripheral circuits and an evaluation of SRAM tests," 2004 International Conference on Test, pp. 114–123, 2004.

[111] P.S. Hughes, "Detection of address decoder faults," U.S. Patent US20 090 037 782A1, 2009.

[112] R. Ramaraju and A.B. Hoefler, "Word line fault detection," U.S. Patent US8 379 468B2, 2013.

[113] D. Kraak, I. Agbo, M. Taouil, S. Hamdioui, P. Weckx S. Cosemans and F. Catthoor, "Hardware-Based Aging Mitigation Scheme for Memory Address Decoder," Proceedings 24th IEEE European Test Symposium (ETS), 2019.

[114] S. Hamdioui, G. Gaydadjiev, "Future challenges in memory testing", Proceedings 14th ProRISC Workshop on Circuits, Systems and Signal Processing, pp. 78-83, Netherlands, 2003.

[115] T-W. Tseng, J-F. Li, "A Shared Parallel Built-In Self-Repair Scheme for Random Access Memories in SOCs," IEEE International Test Conference, 2008.

[116] Y-Y. Hsiao, C-H. Chen and C-W. Wu, "A built-in self-repair scheme for NOR-type flash memory," 24th IEEE VLSI Test Symposium, 2006.

[117] S. Duan, B. Halak and M Zwolinski, "Cell flipping with distributed refresh for cache ageing minimization," Proceedings IEEE Asian Test Symposium (ATS), pp. 1–6, 2018.

[118] P. Dudek, S. Szczepanski, and J.V. Hatfield, "A High-Resolution CMOS Time-to-Digital Converter Utilizing a Vernier Delay Line," IEEE Journal of Solid-State Circuits, vol. 35, no. 2, pp. 240–247, 2000.

[119] V. G. Oklobdzija, "Digital Design and Fabrication," CRC Press, pp. 7-74, 2007.

[120] Y. Cao, J. Velamala, K. Sutaria, et al., "Cross-Layer Modeling and Simulation of Circuit Reliability," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 33, no. 1, pp. 8-23, 2014.

[121] E. Seevinck, F.J. List, J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells", IEEE Journal of Solid-State Circuits, vol. 22, no. 5, pp. 748-754, 1987.

AUTHOR'S PUBLICATIONS

- H-M. Dounavi, Y. Sfikas and Y. Tsiatouhas, "Aging Monitoring in SRAM Sense Amplifiers," International Conference on Modern Circuits and Systems Technologies (MOCAST), 2018.

- H-M. Dounavi, Y. Sfikas and Y. Tsiatouhas, "Periodic Aging Monitoring in SRAM Sense Amplifiers," IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 12-16, 2018.

- H-M. Dounavi, Y. Sfikas and Y. Tsiatouhas, "Periodic Monitoring of BTI Induced Aging in SRAM Sense Amplifiers," IEEE Transactions on Device and Materials Reliability, vol. 19, no.1, pp. 64-72, 2019.

- H-M. Dounavi, Y. Sfikas and Y. Tsiatouhas, "Aging Monitors for SRAM Memory Cells and Sense Amplifiers," Book chapter in Ageing of Integrated Circuits: Causes, Effects and Mitigation Techniques, Editor: B. Halak, Springer, pp. 181-210, 2019.

- H-M. Dounavi, Y. Sfikas and Y. Tsiatouhas, "Aging Prediction and Tolerance for the SRAM Memory Cell and Sense Amplifier," IET Circuits, Devices & Systems, 2019 *(under review)*.

- H-M. Dounavi and Y. Tsiatouhas, "Monitoring of BTI and HCI Aging in SRAM Decoders," IEEE European Test Symposium (ETS), 2020 *(accepted)*.

SHORT BIOGRAPHY

Dounavi Eleni-Maria received her Bachelor diploma from the Department of Computer Science (University of Ioannina, GR), in 2010, and the MSc. degree from the Department of Computer Science and Engineering of the same University in 2013. In 2016 she started pursuing her PhD in the same University and received scholarship from the Hellenic Foundation for Research and Innovation (HFRI) to complete her research. She is currently working as an academic fellow in the University of Ioannina, teaching computer science. Her main research interests include memory integrated circuit design and design for testability, integrated circuits testing and self-healing.

GRANT ACKNOWLEDMENT

I would like to thank the General Secretariat for Research and Technology (GSRT) and the Hellenic Foundation for Research and Innovation (HFRI) for this support and contribution with awarding me with a scholarship and co-funding this research.