Design and Test of Embedded SRAMs

by

Andrei S. Pavlov

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2005

I hereby declare that I am the sole author of this thesis.
I authorize the University of Waterloo to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Andrei S. Pavlov

I further authorize the University of Waterloo to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Andrei S. Pavlov

Abstract

Embedded SRAMs can occupy the majority of the chip area in SoCs. The increased process spreads of modern scaled-down technologies and non-catastrophic defect-related sensitivity to environmental parameters can compromise the stability of SRAM cells, which is quantified by a low Static Noise Margin (SNM). A Stability Fault (SF) can present itself in a cell whose SNM is so small that it can accidentally flip in the worst-case operating conditions. In this work, we conduct a comprehensive SRAM SNM sensitivity analysis and identify the major factors causing low SNM. Based on this study, we propose a Weak Cell Fault Model, which can be used in fault simulations to mimic an SRAM cell with a compromised SNM. Furthermore, we have derived an analytical expression for the SNM of the recently proposed loadless 4T SRAM cell.

Reading a 6T SRAM cell with bit lines precharged to V_{DD} may not detect several types of defects in the pull-up path of the cell. Such defects can cause the SFs. Regular SRAM March Tests are shown to have extremely limited ability to detect SRAM cells with potential SFs. The traditional Data Retention Test (DRT) is costly in terms of the test time and fails to detect open defects in the cell's pull-up path of less than $50M\Omega$, even if provided unlimited pause time at 150^0C ($0.13\mu m$ technology). These factors show the disparity between the existing SRAM test practices and the need for an economical and low PPM cell stability tests.

We introduce the SRAM Cell Stability Detection Concept explaining the mechanism of the weak cell detection. Based on this concept, we propose three novel programmable Design for Testability (DFT) techniques capable of detecting the SFs and replacing the DRT. For verification of the proposed techniques, we have designed two fully functional SRAM test chips: an asynchronous SRAM (CMOS $0.18\mu m$ technology) and a synchronous SRAM (CMOS $0.13\mu m$ technology). The simulation and measurement results have proven

the concept and shown the effectiveness of the proposed DFTs. Compared to the DRT, the proposed techniques offer superior defect coverage and flexibility, reduced test time and no high-temperature requirements. The programmability of the pass/fail threshold facilitates tracking of process modifications and/or quality requirements changes without the need for post-silicon design iterations. This allows the test engineer to strike a balance between the number of test escapees and the yield loss incurred due to excessively stringent testing.

The work presented in this thesis was conducted in collaboration with Philips Research Labs (Eindhoven, The Netherlands) and resulted in two patent applications and several publications.

Acknowledgements

I would like to take this opportunity to express my extreme gratitude to my research supervisor Professor Manoj Sachdev. Without his guidance and expertise, suggestions and insights this work would not have been possible.

I deeply appreciate the guidance and support from the scientists of Philips Research Labs (Eindhoven, The Netherlands) during my internships with them. My mentor José Pineda de Gyvez as well as Mohamed Azimane and Roelof Salters from Philips Reserach and Patrick van de Steeg from Philips Semiconductors and many others made my internships at Philips an excellent and productive experience.

Obtaining the proof of the proposed concepts and measurement results would not have been possible without the chip fabrication service provided by Canadian Microelectronics Corporation (CMC) and support from Phil Regier (University of Waterloo) and Rutger van Veen and Bram Kruseman (Philips Research Labs).

Thanks to all my fellow group mates for the thought-provoking discussions and fun coffee breaks. Especially, I appreciate David J. Rennie and Mohammad Sharifkhani for their willingness to help me out meeting the tight chip tape-out deadlines.

Special thanks go to my wife Natasha whose encouragement and support during my graduate studies have been invaluable.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Motivation

This chapter provides some basics of memory design and test. Section 1.1 gives a short introduction and motivation behind this research. Section 1.2 presents the place of SRAM among other types of memory. Section 1.3 presents the basics of SRAM test. Section 1.4 summarizes the Introduction and presents the chapter break-down of the thesis.

1.1 Introduction

The stability of embedded Static Random Access Memories (SRAMs) is a growing concern in the design and test community [7, 8, 9]. Maintaining an acceptable Static Noise Margin (SNM) in embedded SRAMs while scaling the minimal feature sizes and supply voltages of the Systems-on-a-Chip (SoC) becomes increasingly challenging. The increased process spreads of modern scaled-down technologies and non-catastrophic defect-related sensitivity to environmental parameters can cause stability degradation in SRAMs [10]. Moreover, the minimal feature sizes of SRAM cell transistors combined with the high packing density

of SRAM arrays, often involving relaxed design rules, aggravate this problem. The static noise margin (SNM) can serve as a figure of merit in evaluation of the stability of SRAM cells. Due to factors which are discussed in the following sections, the SNM of even defect-free cells is declining with scaling. In the presence of non-catastrophic defects such as poor vias and contacts, cell stability is degraded even further. However, such defective cells can still escape the standard functional tests (e.g., march tests).

International Technology Roadmap for Semiconductors (ITRS)-2003 [11, 1] predicts "greater parametric yield loss with respect to noise margins" for high density circuits such as SRAM arrays, which are projected to occupy more than 90% of the SoC area in less than ten years. In more severe cases, poor noise margins caused by weak opens in the cell can lead to a Data Retention Fault (DRF). Detection of DRF and stability faults (SF) has been a time consuming and expensive effort. In this thesis, cells causing such faults are referred to as *weak cells*. The traditional Data Retention (a.k.a. Delay or Pause) Test (DRT) has limited fault coverage with respect to the weak cell detection and is time consuming and may even require elevated temperatures. Moreover, for most of the stability faults, the DRT can be ineffective. If undetected, stability faults can manifest themselves under adverse circumstances not covered in the production test and may lead to expensive field failures. Since stability faults are caused by manufacturing defects, they can also indicate potential long-term reliability issues.

Detection of stability faults in SRAM cells requires sensitizing such cells by applying stress, which on one hand should be sufficient to flip the unstable cells and on the other hand insufficient to flip the healthy cells. A number of weak cell detection DFT techniques utilizing weak disturbs have been proposed in the field [9, 7, 8, 12, 13, 14, 15, 16, 17, 18]. However, most of these techniques lack flexibility in setting the detection threshold, which is defined by the stress applied during the test. In this thesis we introduce several new defect-

oriented DFT techniques for stability fault detection is proposed. The proposed techniques offer an easily changeable degree of test stress applied to the Cell Under Test (CUT), do not require any additional circuitry in SRAM array and can be easily incorporated into the memory structure.

1.2 SRAM in the Memory Hierarchy

Memory has been the driving force behind the rapid development of CMOS technology we have been witnessing in the past few decades. Starting from the first 1Kb DRAM chip developed by Intel in the seventies, nowadays DRAM capacities have reached beyond 1Gb.

The advent of the virtual memory in personal computers contributed to the hierarchical structure of various kinds of memory ranging from the small capacity, fast but more costly cache memories to large capacity, slower but more affordable magnetic and optical storage. The pyramid-like hierarchy of memory types in a personal computer, shown in Figure 1.1, reflects the growing speed and cost/bit as we move from the bottom Level 5 (L5) remote secondary storage to the topmost register Level (L0). The introduction of memory hierarchy is a fundamental consequence of maintaining the random access memory abstraction and practical limits on cost and power consumption.

The growing gap between the Micro Processor Unit (MPU) cycle time and DRAM access time necessitated the introduction of several levels of caching in modern data processors. In personal computer MPUs such levels are often represented by L1 and L2 on-chip embedded SRAM cache memories. As the speed gap between MPU, memory and mass storage continues to widen, deeper memory hierarchies have been introduced in high-end server microprocessors [19].

Depending on the amount of L2 cache, ITRS distinguishes the Cost-Performance MPU,

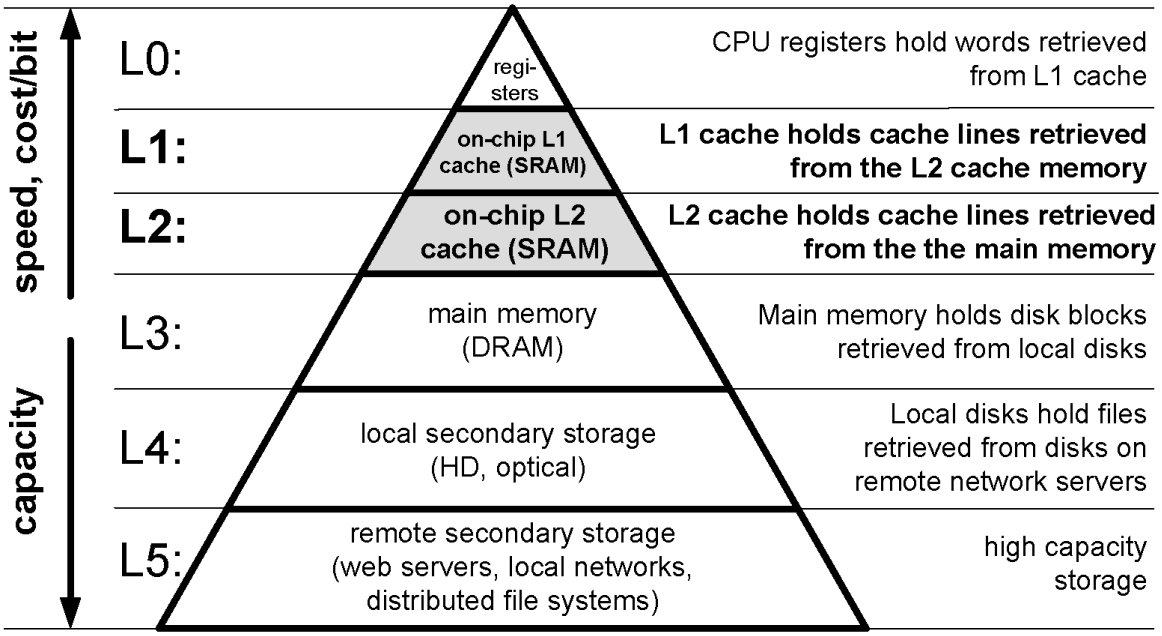
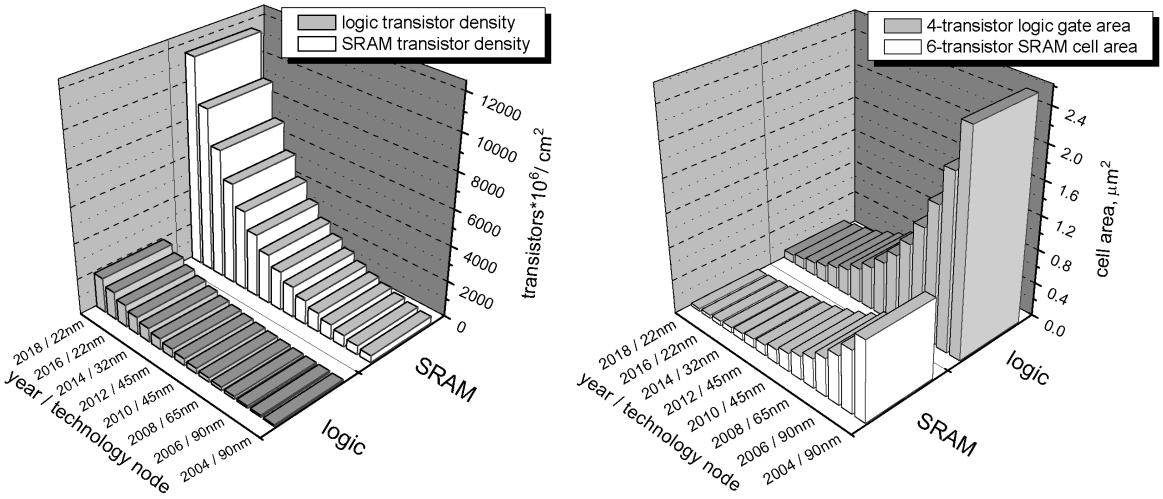


Figure 1.1 Computer memory hierarchy.

which is optimized for maximum performance, and the lowest cost by limiting the amount of on-chip SRAM Level-2 (L2) cache and the high-performance MPU optimized for maximum system performance by combining a single or multiple CPU cores with a large L2, and recently, L3 on-chip SRAM cache [19]. Logic functionality and L2 cache capacity typically doubles every technology generation by doubling the number of on-chip CPU cores and associated memory.

One of the ways to increase the on-chip cache sizes is to use the high-density dynamic RAM. An SoC with embedded DRAMs implemented in the standard logic process can benefit from fast low-V_{TH} transistors. However, the inherently high subthreshold leakage current complicates implementation of a 1T DRAM cell. Replacing 1T DRAM cells with alternative DRAM cell designs having a larger number of transistors results in an area penalty and undermines the cell area advantage that embedded DRAMs normally have



(a) Transistor density trends: six-transistor SRAM cell vs. four-transistor logic gate.

(b) Area trends: six-transistor SRAM cell area vs. a four-transistor logic gate.

Figure 1.2 High-volume microprocessor technology trends with respect to year/technology node (ITRS-2004 [1] prediction).

over embedded SRAMs. If DRAM process is utilized to fabricate embedded DRAM, a 1T DRAM cell can be used to achieve high packing density. However, the high-V_{TH} low-leakage DRAM process limits the performance of such an SOC [20],[21]. Therefore, using of embedded DRAMs may be justified in specialized SoC, requiring large embedded memory size and operating at relatively low to medium speeds.

Embedded SRAMs have been used to accelerate the performance of high-end microprocessors, network routers and switches. They use the regular fast logic process and do not require additional mask steps. The reduced supply voltages of the new technology nodes pose a higher risk of soft errors that can be especially problematic for the scaled-down embedded DRAMs. Due to the reduced charge stored on the DRAM capacitor and the absence of the latching feedback mechanism as in SRAMs, embedded DRAMs can be more susceptible to α-particles and cosmic rays. This is explained in more detail in

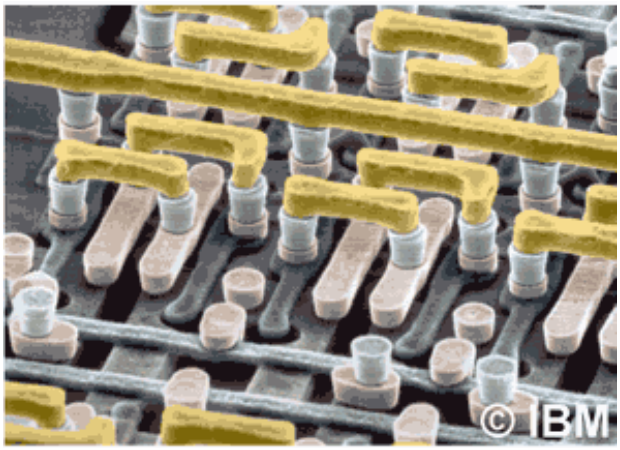


Figure 1.3 High packing density in an SRAM array.

Section 3.4.3 on page 69.

In order to double on-chip functionality every two years according to Moore's Law, technology-node scaling of 0.7 in linear size and 0.5 in area has to be carried out every three years; as well as an additional device/process design improvement of 0.8/(2 years) must be achieved. The advancement rate through the technology nodes determined by the ability to economically manufacture a chip based upon the best available leading-edge design and manufacturing process. The typical high-performance ASIC design is assumed to have the same average transistor density as high-performance MPUs, which mostly consist of SRAM transistors [1].

Every technology node the transistor density gap between the regular logic and embedded SRAMs is predicted to grow from approximately x5 in 2005 and to around x6 by 2018, as shown in Figure 1.2(a) on the previous page. At the same time, the six-transistor SRAM cell area now constitutes only 36% of the typical four-transistor logic gate area and is projected to further reduce to become 35% in year 2018 (Figure 1.2(b)). As a rule, SRAM cell size continues to scale ≈0.5x/generation driven by the need for high-performance processors

with larger caches.

Scaling of a memory cell leads to the increase in the critical area where a spot defect of a given size may cause a bridge or a break in the topology and damage the cell. Figure 1.3 on the preceding page illustrates the defect-prone high-density SRAM layer structure. Combined with the exponentially growing bit count of embedded SRAMs, design and test of embedded SRAMs are becoming more challenging with each technology generation. Special test methods are required to provide efficient and cost-effective testing for reliable screening and/or repairing of the defective cells. Several such test methods will be introduced in this thesis.

1.3 SRAM Test Basics

An efficient and economical memory test should provide the best defect coverage in the shortest time. Besides fault detection, a manufacturing memory test must also include diagnostic capability that allows to identify and possibly repair defective locations by applying redundant elements. In addition, the diagnostic capability can be instrumental if the manufacturing yield ramp-up by providing feedback to designers and process engineers.

1.3.1 March Tests

March tests can be classified as functional test techniques. Such tests verify the circuit functionality by conducting read and write operations on each memory location. A march test consists of a finite sequence of march elements, which in turn consist of the sequence of operations applied to every cell before proceeding to the next cell. March elements can be applied in increasing ($\Uparrow$), decreasing ($\Downarrow$) or arbitrary ($\Updownarrow$) address order. Each operation of the march element can consist of: writing a 1 (w1), writing a 0 (w0), reading an expected

Table 1.1 Example SRAM march tests.

Test	Sequence	N	Faults Detected			
			SAF	AF	TF	CF
MATS	⇳(w0)⇳(r0,w1)⇳(r1)	4N	+	+/-	-	-
MATS+	⇳(w0)⇑(r0,w1)⇓(r1,w0)	5N	+	+	-	-
MATS++	⇳(w0)⇑(r0,w1)⇓(r1,w0,r0)	6N	+	+	+	-
March Y	⇳(w0)⇑(r0,w1,r1)⇓(r1,w0,r0)⇳(r0)	8N	+	+	+	+/-
March C-	⇳(w0)⇑(r0,w1)⇑(r1,w0)⇓(r0,w1)⇓(r1,w0)⇳(r0)	10N	+	+	+	+

value 1 (r1) and reading an expected value 0 (r0). If reading does not produce the expected value, the memory location is deemed faulty. (⇑) and (⇓) can be reversed and data values inverted without affecting the fault coverage of a march test [22].

March tests are often designed to target a certain fault model, such as stuck-at, address, transition, coupling, pattern-sensitive, delay and data retention. The fault coverage of a march test varies and can range from 0% to 95-100% depending on the chosen test algorithm for every given fault. However, if the fault model poorly characterizes the real faults that may occur in the given circuit, then the developed tests can test for non-existing faults or some of the existing faults may escape the test [23]. Inductive Fault Analysis (IFA) can be used to determine which faults are more likely to occur by placing a physical defect of a given size into a particular circuit layout. Using IFA we conducted analysis of the SNM sensitivity to the resistive defects, which are likely to appear in an SRAM cell layout similar to shown in Figure 4.3 on page 91 (see Section 3.4.2).

Table 1.1 shows some of the examples of march tests [22]. The simplest practical march test, Modified Algorithmic Test Sequence (MATS), is the basic march test that verifies only the stuck-at faults (SAFs) by writing each data background once and reading it back. By

adding extra march operations/elements to the basic MATS march test, one can improve the test to cover more faults at the expense of the extra test time. The March C- algorithm is one of the most efficient march tests with respect to fault coverage/complexity [22]. It has been observed to provide the highest fault coverage detecting Stuck-At Faults (SAFs), Address Faults (AFs), Transition Faults (TFs) and unlinked idempotent Coupling Faults (CFs). More complex march test are capable of detecting Neighborhood Pattern Sensitive Faults (NPSF).

1.3.2 Data Retention Test

A typical Data Retention Test (DRT) is implemented as a pause of an order of 100ms between the march elements. DRT can detect a complete open in the pull-up path of an SRAM cell. In case of a symmetric defect, where the pull-up paths in both the inverters are open, the detection is not dependent on the data value stored in the cell. However, if a cell has an asymmetric defect, where only one of the inverters has an open in the pull-up path, the DRT will only detect an open in the pull-up path of the node storing a "1" [22]. This property of the DRT requires to run it for each of the two opposite backgrounds to cover both the asymmetric faults in the pull-up path of the cell.

Since running the DRT for each of the data backgrounds takes 100-200ms [7], applying the DRT will result in each chip spending an extra 200-400ms on the tester. Figure 1.4 illustrates the insertion locations of the Delay elements between the march elements March#1 and March#2, and March#3 and March#4 in March C- test. In many cases, the total impact of using the DRT on the test time and hence, the test cost is prohibitively high. Combined with the limited defect resistance coverage of the DRT and the high-temperature requirement to improve it, which is discussed in detail in Section 4.1.1, the disadvantages of using the DRT urge the industry to replace it with special Design for Testability (DFT)

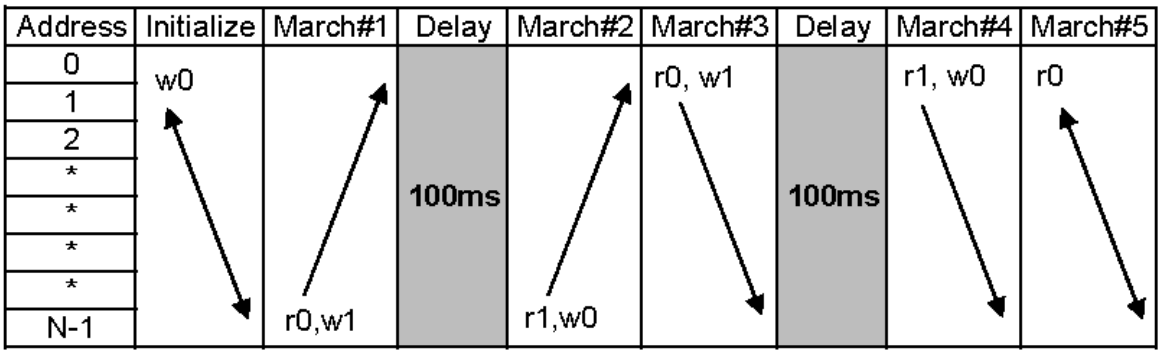


Figure 1.4 Inserting delay elements into the March C- test to include a Data Retention Test (DRT).

techniques (e.g., [24] and [13]).

In Chapter 5 we have investigated the efficiency of some of the march test in detection of the stability faults in SRAMs.

1.3.3 SRAM Test Economics

Memory testing has become more difficult over the past decades, while parametric faults are becoming more common leading to yield loss and reliability problems. While the number of I/O pins has increased by an order of magnitude, the number of transistors has increased by up to four orders of magnitude [25]. The increasing transistor/pin ratio which is projected to exceed 2.3 million/pin by 2016 [11] is limiting the controllability from the primary inputs and the observability of the faulty behavior at the primary outputs in embedded memories [2]. Moreover, striving to keep up with the increasing clock speeds of SoCs increases the cost of Automatic Test Equipment (ATE) so that the at-speed test of high-performance chips becomes problematic. The fastest available ATE is always slower than the chips it will test. As multi-million-dollar ATEs become commonplace, the cost of the tester time spent on every chip directly impacts the total cost of the chip.

Table 1.2 Test time as a function of the memory size (calculated for $t_{cycle} = 10ns$). Test algorithms of more than linear complexity are not economical for large memories.

Size	N	$6N$	$11N$	$Nlog_2N$	$N^{1.5}$	N^2
$1Mb$	0.010s	0.063s	0.115s	0.210s	10.7s	1.3d
$4Mb$	0.042s	0.252s	0.461s	0.923s	85.9s	20.4d
$16Mb$	0.168s	1.007s	1.8s	4.0s	19m	325.8d
$64Mb$	0.671s	4.0s	7.4s	17.5s	1.53h	14.3y
$256Mb$	2.7s	16.1s	29.5s	75.2s	12.2h	228.5y

A customer regards a product to be of high quality if the product is meeting their requirements at the lowest possible cost. Memory tests check conformance to the requirements, while the cost is reduced by improving the process yield. Quality can be expressed as the number of customer returns per million (parts per million):

$$Defect\ level = \frac{test\ escapes}{total\ number\ of\ shipped\ chips}\ (PPM) \tag{1.1}$$

It is easy to see that a 99% test fault coverage will result in 10000PPM, which is a long way from the industry's aim of sub-100 PPM [26].

Exhaustive functional memory test is economically unfeasible. For instance, exhaustive test of a 1Kb SRAM array will take 2^{1024} combinations to complete a full functional test. Such a test of an SRAM with the access time of 10ns will conclude in more than 10^{290} years! Attention was given to the fault modelling, structural testing and DFT techniques to ensure and maintain test cost effectiveness and low defect levels.

The test cost per chip, which can run up to a half of the product cost and is directly related to the test time, cannot increase significantly. However, the number of bits/chip is exponentially growing and fault sensitivity is increasing. Maintaining an acceptable defect

level in the upcoming scaled down generations will likely require more complicated and lengthy tests.

Table 1.2 on the previous page presents the test time required to test a memory using march algorithms of various complexity, where N is the number of tested memory locations. The access cycle time is assumed to be 10ns. As apparent from Table 1.2, early memory test methods with test times proportional to $t_{cycle} * Nlog_2N$ or even $t_{cycle} * N^2$ are now prohibitively expensive [25].

An ideal test algorithm(s) should have maximum fault coverage with minimum complexity, which is proportional to the test time and test cost. However, real test algorithms have limited fault coverage. To improve the fault coverage, several tests may have to be employed. Therefore, a test engineer faces a difficult choice between balancing the test cost and the defect level.

1.3.4 Design For Test Techniques

In general, DFT techniques seek to improve the testability by including additional circuitry. While the additional DFT circuitry can increase design cost, it is often offset by the resulting reduction of test development time and improvement of the quality level. DFT circuitry can be controlled directly by the external tester (ATE) or through a Built-In Self Test circuitry described below.

Built-In Self Test

Built-In Self Test (BIST) is a special type of DFT technique that facilitates internal test pattern generation and output response compaction [2]. A basic BIST architecture is shown in Figure 1.5 on the following page. Two essential parts of a BIST are the Test Pattern Generator (TPG) and the Output Response Analyzer (ORA). The TPG generates

a sequence of patterns for testing of the CUT, whereas the ORA compacts the circuit responses into some type of pass/fail indication. Test Controller provides control signal to all blocks of a BIST. The input isolation circuitry switches the input of the CUT from the normal system inputs to the TPG outputs.

Applying BIST strategies is often the only economical way to ensure at-speed testing and to overcome the limited observability and controllability of embedded SRAMs.

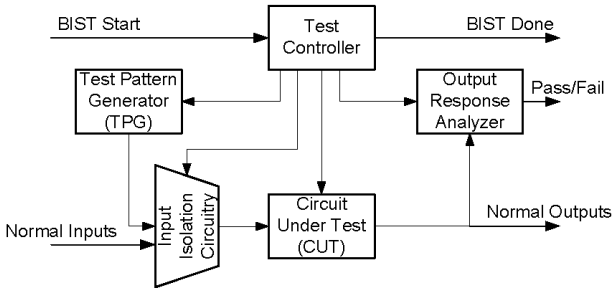


Figure 1.5 Basic BIST architecture [2].

Running at-speed test helps to identify delay faults in manufacturing and burn-in tests that otherwise might not be detected in a slower speed test provided by ATE. Moreover, BIST-equipped SoCs are less demanding to ATE's test vector memory and number of serviced pins. The extra design time incurred by including BIST is often offset by the savings in test development time and can expedite time-to-market in some cases. Disadvantages of incorporation of BIST circuitry include the area overhead, design effort and performance penalties of the circuit isolation MUXs in high-performance designs. However, in most cases the benefits of BIST overweigh its disadvantages.

1.3.5 Redundancy

Redundancy can be used to repair a certain number of failing cells by replacing the faulty rows and/or columns or single faulty bits with the spare ones. Applying row or column

redundancy inevitably introduces a performance penalty for the affected row or column. On the other hand, the bit redundancy when only one faulty bit is replaced by a spare bit incurs a speed penalty only for that one bit. The repair capability can significantly improve the overall manufacturing yield for large SRAM arrays. The exact redundancy yield benefit is determined by a complex relationship between the chip size, SRAM array real estate share, and the technology maturity.

One of the main objectives of any chip design is to minimize cost. Two major factors of the chip cost are the number of good chips per wafer (GCPW) and the test cost per chip (TCPC).

Decisions on whether to use redundancy is affected by several considerations and represents a trade-off between the cost and the benefits of using redundancy [27]:

- Smaller chip size can increase the number of chips per wafer (CPW).
- Large SRAM arrays typically have lower manufacturing yield for new, not matured technologies, decreasing GCPW.
- Redundancy allows some of the failing chips to be repaired, increasing the manufacturing yield and GCPW, especially for large SRAM arrays.
- Redundancy requires extra area (fuses and larger BIST controllers) decreasing the CPW.
- Redundancy requires extra test time for registering the failing cell addresses during test, address rerouting by blowing the fuses, which increases the TCPC.

There are practical limitations with respect to the amount of non-redundant memory that can be used on a chip. A chip can only contain less than 256Kb bits of non-redundant SRAM [28]. To maintain an acceptable yield for a larger number of bits in an SRAM array, redundancy becomes indispensable. Since a chip can be rendered faulty by any single failing

memory bit, the yield is determined by the total amount of embedded memory, not only by the largest instance.

During the initial manufacturing test, a map of faulty locations is stored using fuses. Memory arrays with redundancy are connected to a fuse decompression circuitry that reads the fuse values and decompresses them to the fuse shift-register chain under the control of the BIST circuit. During a power-on-reset sequence, the fuse information is decompressed and shifted through the fuse shift-register chain to the corresponding row and column decoders. Based on the fuse information, the decoders reroute the defective address values to the spare rows and columns. Once all the memories in SoC received the redundancy information, a "ready" flag is issued and the SoC begins to operate. A separate scan chain has to be used to restore the fuse information on power-up from the sleep mode, which adds to design complexity and requires co-design of SRAM and BIST for each instance.

Before redundancy can be applied, all the faulty memory locations have to be reliably identified. Detecting weak cells for each memory array using the traditional DRT can add to the test time of each memory array at least 200ms as was shown in Figure 1.4 on page 10. Furthermore, the DRT often requires high temperatures to provide higher defect resistance coverage. Often, the resulting impact of DRT on the test cost is prohibitive.

The novel DFT techniques presented in this work can successfully replace the DRT. The proposed DFT techniques exceed the DRT in test time and defect resistance coverage. Moreover, the flexibility of the pass/fail threshold provided by digital programmability of the applied test stress allows control over the balance between the test escapees and the defect level.

1.4 Summary and Thesis Outline

Simpler integration into the regular logic process and high operation speed often make embedded SRAMs a choice memory for bridging the performance gap between the high-speed microprocessors and the main system memory. The area of many SoCs is dominated by embedded SRAMs composed of minimal-size transistors with transistor densities over five times higher than that of regular logic [11].

The minimal-size transistors typically used in the constantly growing SRAM arrays represent tremendous defect-sensitive area, which often makes embedded SRAMs the yield limiters in SoCs. Subtle defects that are not detected by functional march tests are commonplace in modern SRAMs. Data retention tests in SRAMs can fail to detect many of such defects and they significantly add to the test time. Test time is a major contributor to the total cost of SoCs with large embedded SRAMs. The shortcomings of the Data Retention Test have been addressed by introducing a single-threshold DFT techniques such as the Weak Write Test Mode [7]. However, as the process technology continued to scale down, DFTs with a single pass/fail threshold could not be relied on anymore. The amount of test stress generated by these DFTs became difficult to control. This work is addressing this problem by developing DFT techniques with multiple programmable pass/fail thresholds.

Keeping the tester speed on par with that of high-end microprocessors is not economical. Therefore, the throughput of a multimillion-dollar tester is critical for economical efficiency of the test. Moreover, the growing transistor/pin ratio limits the controllability of the inputs and the observability of the outputs in embedded SRAMs. Built-In Self Test (BIST) is often the only way to overcome the controllability and observability limitations and ensure economical at-speed testing. Replacing the faulty memory locations with redundant ones can dramatically improve the yield of large SRAMs. DFT techniques developed in this work facilitate reliable and economical identification of such faulty unstable SRAM

cells.

This thesis is organized as follows: the introductory Chapter 1 is followed by a description of the main SRAM building blocks in Chapter 2. SRAM cell stability characterization and extensive SNM sensitivity analysis is presented in Chapter 3. Chapter 4 introduces the proposed stability fault model and detection concept followed by the classification and description of the existing industrial stability test methods. Chapter 5 investigates the stability fault detection capabilities of march tests and introduces a march sequence for coupling fault detection. Chapter 6 presents the main contributions of this work – three novel DFT techniques for Stability and Data Retention Faults detection in SRAM cells. Chapter 7 summarizes the findings and contributions of this work.

Chapter 2

SRAM Design and Operation

In this chapter the main SRAM building blocks are discussed. Various implementation variations relevant design tradeoffs based on technology and performance metrics. Some of the presented blocks will be used in the test chips designed during the course of this work to prove the concept of the proposed SRAM cell stability test techniques. Section 2.1 explains the basic SRAM block structure. Section 2.2 presents various types of SRAM cells and their design principles. Section 2.3 describes sense amplifiers and precharge/equalization circuits. Write drivers are described in Section 2.4. Row and column decoding schemes are discussed in Section 2.5. Section 2.6 is dedicated to control signal timing generation. The Address Transition Detector for asynchronous SRAMs is described in Section 2.7. Section 2.8 summarizes this chapter.

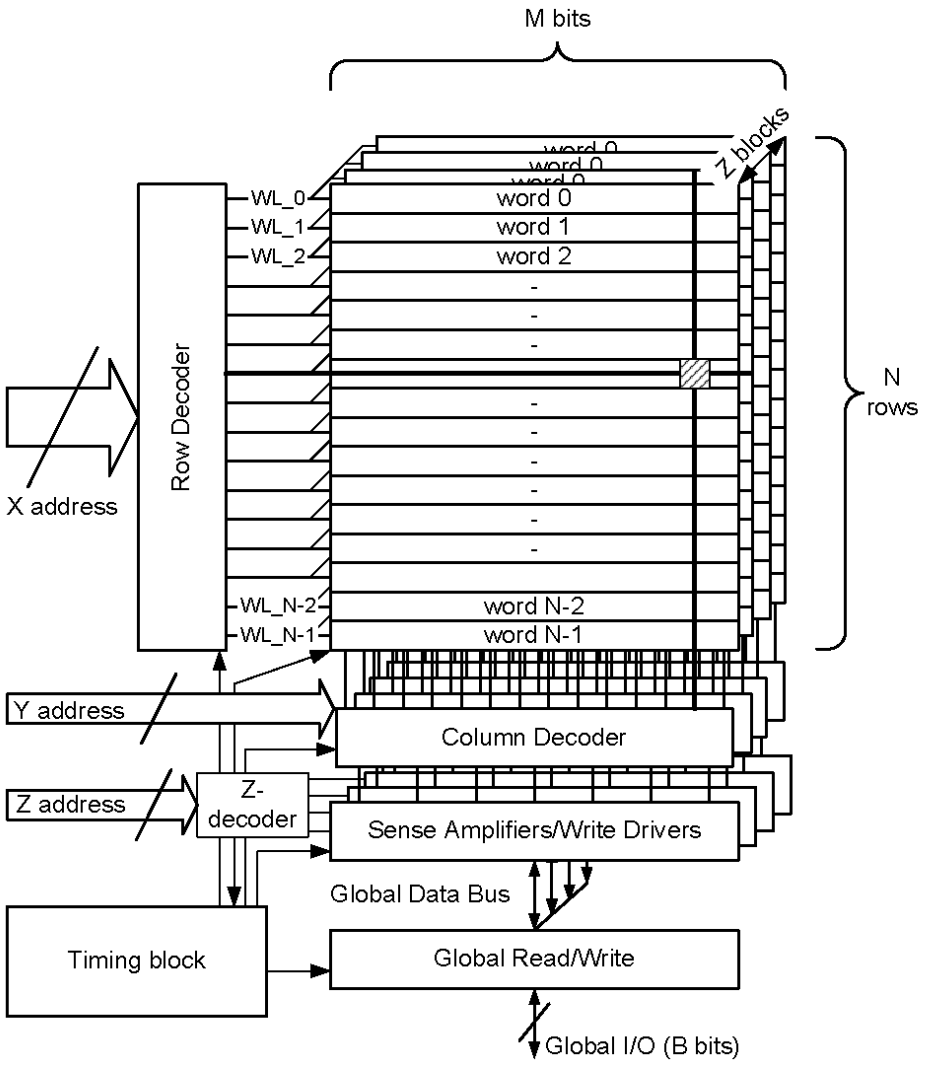


Figure 2.1 SRAM block diagram.

2.1 SRAM Block Structure

Figure 2.1 shows an example of the basic SRAM block structure. A row decoder gated by the timing block decodes X row address bits and selects one of the word lines WL_0–WL_N-1. If an SRAM array of N rows and M bits is arranged in a page manner, an additional

Z-decoder activates the accessed page. Figure 2.1 shows an example with four pages of NxM arrays with the corresponding I/O circuitry.

Memories can be bit-oriented or word-oriented. In a bit-oriented memory, each address accesses a single bit. Whereas in a word-oriented memory, a word consisting of n (where the popular values of n include 8, 16, 32 or 64) bits is accessed with each address. Column decoders or column MUXs (YMUX) addressed by Y address bits are often used to allow sharing of a single sense amplifier among 2, 4 or more columns. Most of modern SRAMs are self-timed, i.e. all the internal timing is generated by the timing block within an SRAM instance. An additional Chip Select (CS) signal, introducing an extra decoding hierarchy level, is often provided in multi-chip architectures.

The main SRAM building blocks will be described in more detail in the following sections.

2.2 The SRAM Cell

Memory cells are the key components of any SRAM serving for storage of binary information. A typical SRAM cell is comprised two cross-coupled inverters forming a latch and access transistors. Access transistors enable read and write access to the cell and cell isolation for the not-accessed state. An SRAM cell has to provide non-destructive read access, write capability and infinite storage (or data retention) time provided the power is supplied to the cell. Hierarchically, memory cells are arranged in cores, which can be further divided into blocks and arrays depending on the system speed and power requirements.

We will consider three of the more recent SRAM cells: a resistive load four-transistor (4T) SRAM cell, a six-transistor (6T) CMOS SRAM cell and a loadless 4T SRAM cell. We will then discuss their advantages and disadvantages. The cell design considerations

represent a tradeoff between cell area, robustness, speed and power. Cell size minimization is one of the most important design objectives. A smaller cell allows the number of bits per unit area to be increased and thus, decreases cost per bit. Reduced cell size can indirectly improve the speed and power consumption due to the reduction of the associated capacitances. However, the cell area might have to be traded off for high-performance or low-power, radiation hardness or special functionality requirements.

2.2.1 4T SRAM Cell with Polysilicon Resistor Load

The main advantage of static 4T cells with polysilicon resistor load (PRL) (Figure 2.2) is the approximately 30% smaller area as compared to 6T SRAM cells. Due to the higher electron mobility ($\mu_n/\mu_p = 1.5 - 3$), all transistors in a PRL cell are normally NMOS. The load resistors serve to compensate for the off-state leakage of the pull-down devices. On one hand, the values of R_L must be as high as possible to retain a reasonable noise margin NM_L, i.e., to limit the "0" level rise and reduce the static power consumption. On the other hand, a high R_L severely increases the low-to-high propagation delay if $V_{DD}/2$ precharge is used and it also increases the cell size. Furthermore, precharging the bit lines to $V_{DD}/2$ can compromise the cell stability with scaling of the V_{DD}/V_{TH} ratio. Precharge of bit lines to full V_{DD} can alleviate the requirement for the low-to-high cell transition current at the cost of the additional precharge time required for a full-V_{DD} precharge and the associated power consumption. The upper resistance limit on R_L is put by the requirement to provide a pull-up current of at least two orders of magnitude larger than the leakage current [3]. The lower limit on R_L is put by the required noise immunity and power consumption requirements. The technological variations of R_L caused by the limitations of doping and annealing techniques pose another constraint on the increase of R_L.

Historically, 4T polysilicon resistor load cells are the remnants of the pre-CMOS tech-

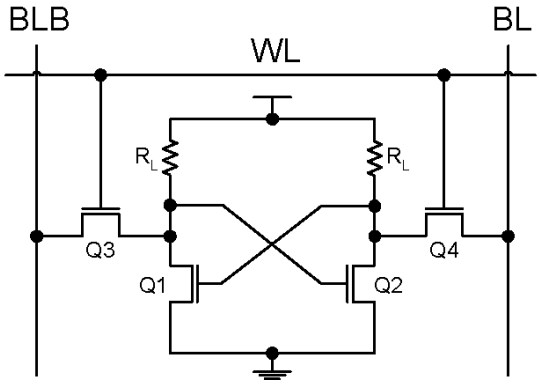


Figure 2.2 Four-transistor (4T) SRAM cell with polysilicon resistor load.

nologies. Ratioed inverters comprising the cell have lower gain in the transition region and produce inherently less steep VTCs, which reflects on the SNM values and the recovery time from a metastable state [29] of such cells. The reduction of V_{DD} from the standard 5V to 3.3V, 2.5V and so on, i.e., the switch from constant-voltage scaling to constant-field scaling to combat the short-channel effects, revealed non-satisfactory low-voltage stability of the PRL cells. Moreover, the extra technological steps of forming high-resistivity polysilicon are not a part of the standard logic technological process. Insufficient tolerance to soft errors, which is directly linked to the SNM, adds to the list of disadvantages of a PRL cell. These factors prohibit using the PRL SRAM cells in Systems-on-a-Chip (SoCs) traditionally implemented using a standard full CMOS process. All the mentioned factors excluded the PRL cell from the current mainstream scaled-down technologies. The PRL cells will not be considered in this thesis.

2.2.2 6T CMOS SRAM Cell

The mainstream six-transistor (6T) CMOS SRAM cell is shown in Figure 2.3. Similarly to one of the implementations of an SR latch, it consists of six transistors. Four transistors

($Q1 - Q4$) comprise cross-coupled CMOS inverters and two NMOS transistors $Q5$ and $Q6$ provide read and write access to the cell. Upon the activation of the word line, the access transistors connect the two internal nodes of the cell to the true (BL) and the complementary (BLB) bit lines.

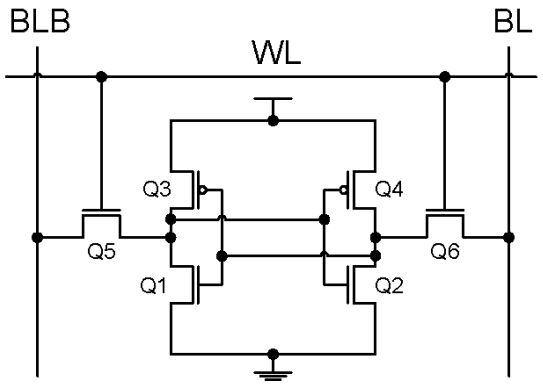


Figure 2.3 Six-transistor (6T) CMOS SRAM cell.

Since in this work we refer to the 6T SRAM cell (except for Section 3.5), we will discuss the operation and the transistor sizing constraints of the 6T SRAM cell in more detail.

As was mentioned earlier, an SRAM cell has to provide a non-destructive read and a quick write – the two opposing requirements, which impose constraints on the cell transistor currents governed by their transistor sizing.

Read Operation

The read operation is stared by enabling the word line (WL) and connecting the precharged bit lines, BL and BLB, to the internal nodes of the cell.

Upon read access, the bit line voltage V_{BL} remains at the precharge level equal V_{DD}. The complementary bit line voltage V_{BLB} is discharged through transistors $Q1$ and $Q5$ connected in series (Figure 2.4). Effectively, transistors $Q1$ and $Q5$ form a voltage divider

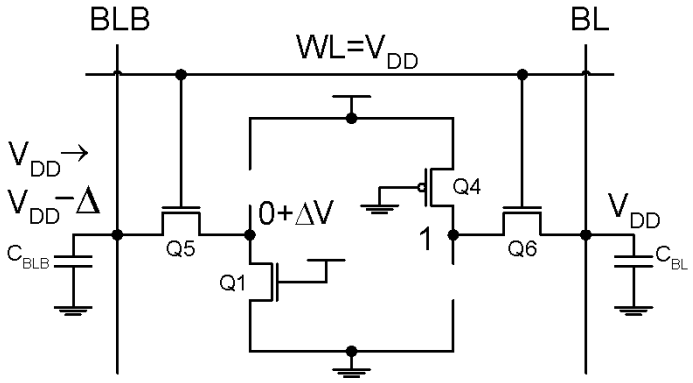
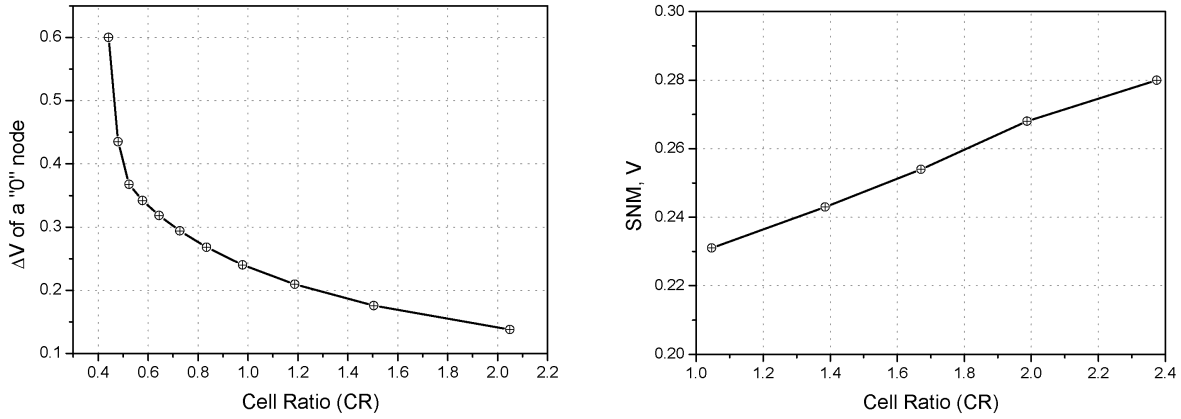


Figure 2.4 Simplified model of a 6T CMOS SRAM cell during a read operation.



(a) (b)

Figure 2.5 The rise ΔV of the "0" node (a) and the SNM (b) as a function of the Cell Ratio (CR) $\left(CR = \frac{W_1}{L_1}/\frac{W_5}{L_5} = \frac{W_2}{L_2}/\frac{W_6}{L_6}$ in Figure 2.3$\right)$ in a 6T CMOS SRAM cell (simulated in CMOS 0.13μm technology, V_{DD}=1.2V).

whose output is connected to the input of inverter $Q2-Q4$ in Figure 2.3. Sizing of $Q1$ and $Q5$ should ensure that inverter $Q2-Q4$ does not switch causing a read upset. In other words, $0+\Delta V$ should be less than the switching threshold of inverter $Q2-Q4$ plus some *safety margin* or *Noise Margin*.

Ignoring the short-channel and body effects, the maximum allowed value $0+\Delta V$ of the "0" node during read access can be expressed as [3]:

$$\Delta V = \frac{V_{DSATn} + CR(V_{DD} - V_{THn}) - \sqrt{V_{DSATn}^2 (1 + CR) + CR^2 (V_{DD} - V_{THn})^2}}{CR} \quad (2.1)$$

where CR (a.k.a. β) is the cell ratio defined as

$$CR = \frac{W_1/L_1}{W_5/L_5} \quad (2.2)$$

Since the cell is fully symmetrical, the CR is the same for $Q2$ and $Q6$.

The dependence of ΔV on the CR is shown in Figure 2.5(a). Normally, in order to ensure a non-destructive read and an adequate noise margin, CR has to be greater than one and can be varied depending on the target application of the cell from approximately 1 to 2.5. Larger CRs provide higher read current I_{read} (and hence – the speed) and SNM (see Figure 2.5(b)) at the expense of larger area taken by the driver transistors $Q1$ and $Q2$. Whereas smaller CRs make for a more compact cell with moderate speed and noise margins. Both for ensuring cell stability and reducing the leakage current of the access transistors, a preferred sizing solution is to use a minimum width with a slightly larger than minimal length access transistors and a larger than minimal width with a minimal length driver transistors.

Once the complementary bit line voltage V_{BLB} has been discharged to a certain $V_{DD}-\Delta$ sufficient for reliable sensing by the sense amplifier, the sense amplifier is enabled and amplifies the small differential voltage between the bit lines to the full-swing CMOS level.

Write Operation

The write operation is similar to resetting an SR latch. One of the bit lines, e.g., BL in Figure 2.6, is driven from precharged value (V_{DD}) to the ground potential by a write driver

through transistor $Q6$. If transistors $Q4$ and $Q6$ are properly sized, then the cell is flipped and its data is effectively overwritten. Note that the write operation is applied to the node storing a "1". This is necessitated by the non-destructive read constraint that ensures that a "0" node does not exceed the switching threshold of inverter $Q2 - Q4$. The function of the pull-up transistors is only to maintain the high level on the "1" storage node and prevent its discharge by the off-state leakage current of the driver transistor during data retention and to provide the low-to-high transition during overwriting.

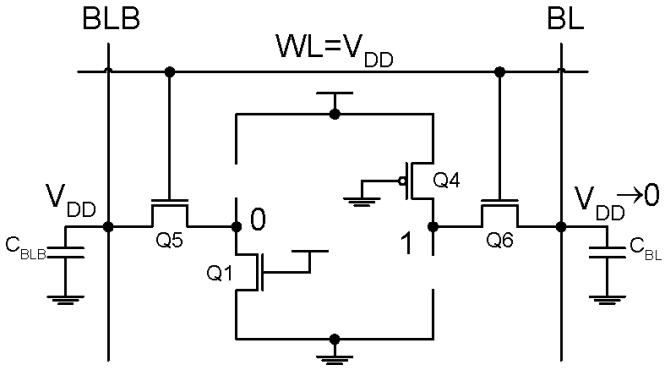


Figure 2.6 Simplified model of a 6T CMOS SRAM cell during a write operation.

Assuming that the switching will not start before "1" node is below V_{TH_Q1}, a simplified overwrite condition can be expressed as [3]:

$$V_{\text{“1”}} = V_{DD} - V_{THn} - \sqrt{(V_{DD} - V_{THn})^2 - 2\frac{\mu_p}{\mu_n} PR \left((V_{DD} - |V_{THp}|) V_{DSATp} - \frac{V_{DSATp}^2}{2} \right)} \tag{2.3}$$

where the pull-up ratio of the cell, PR, is defined as:

$$PR = \frac{W_4/L_4}{W_6/L_6} \tag{2.4}$$

The $V_{\text{“1”}}$ requirement is easily met using minimal-sized access and pull-up transistors only due to μ_n/μ_p ratio. Simulation results shown in Figure 2.7 demonstrate that for a normal

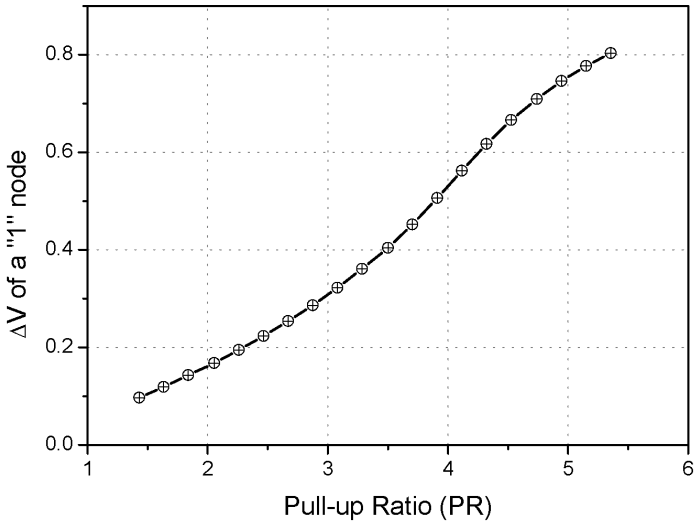


Figure 2.7 **The voltage drop at node $V_{\text{“1”}}$ during write access as a function of the Pull-Up ratio (PR) $\left(CR = \frac{W_4}{L_4}/\frac{W_6}{L_6} = \frac{W_5}{L_5}/\frac{W_3}{L_3}$ in Figure 2.3$\right)$ of a 6T CMOS SRAM cell (simulated in CMOS 0.13μm technology, V_{DD}=1.2V).**

write operation, i.e., to pull the $V_{\text{“1”}}$ node below V_{THn}, the W/L of the pull-up transistor has to be less than 3-4 W/L of the access transistor. The exact maximum allowed PR is defined by the V_{THn} process option and by the switching threshold of inverter $Q1 - Q3$ in Figure 2.3. Normally, to minimize the cell area and hence, increase the packing density, the sizes of the pull-up and access transistors are chosen to be minimal and approximately the same. However, stronger access transistors and/or weaker pull-up transistors may be needed to ensure a robust write operation under the worst process conditions e.g., in the fast PMOS and slow NMOS corner.

Despite the larger number of transistors compared to the other discussed cells, 6T CMOS SRAM offers superior stability and packing density provided the same performance and environmental tolerance.

2.2.3 4T Loadless SRAM Cell

Recently, the new loadless 4T CMOS SRAM cell shown in Figure 2.8 was proposed by NEC [30] and [31] for ultra-high density SRAM macros [32] and [33]. A 4T SRAM cell comprises minimal size NMOS (Q1, Q2) and PMOS (Q3, Q4) transistors. Data retention without the need for refresh is provided if the leakage current of the PMOS transistors is higher than the leakage current of the NMOS transistors. This condition is normally provided by using dual-V_{TH} process with V_{TH_p} <V_{TH_n}. In a 4T SRAM cell, PMOS transistors Q3 and Q4 serve as access transistors as opposed to NMOS access transistors Q3 and Q4 in a 6T SRAM cell. Due to the mobility ratio of the NMOS to the PMOS transistors μ_n/μ_p which is normally around two to three, all the transistors in a 4T SRAM cell can be of minimal size. Recall that for a 6T SRAM cell to guarantee a non-destructive read operation, NMOS driver transistors Q1 and Q2 have to be 1.5-2.5 times larger than NMOS access transistors Q3 and Q4 in Figure 2.8. This fact, in addition to the larger number of transistors in a 6T SRAM cell, makes a 4T SRAM cell an area-efficient choice.

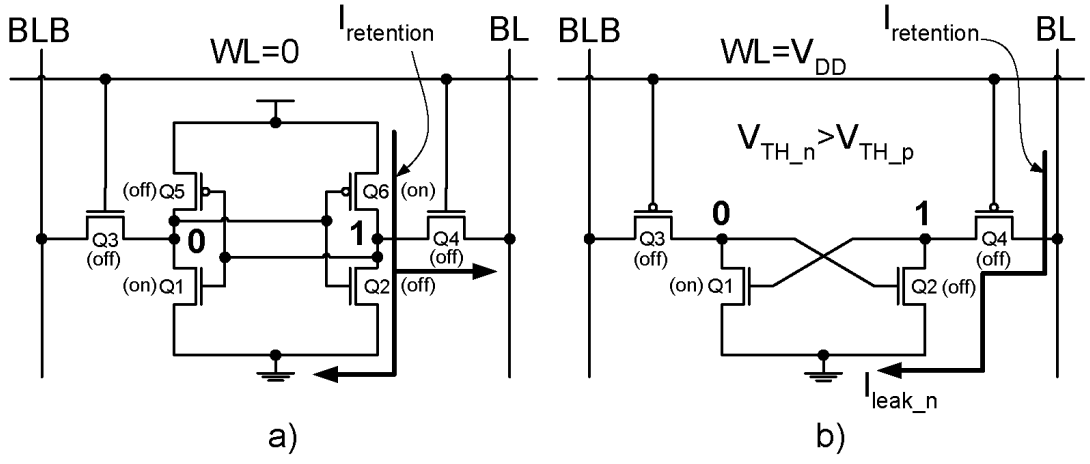


Figure 2.8 Six-transistor (a) and four-transistor (b) CMOS SRAM cells.

For the same access speed and comparable SNM, the 4T SRAM cell area is 50-65% smaller than the area of a conventional 6T SRAM cell. Since memory blocks occupy considerable chip area in SoCs, SRAM cell area is a critical factor in the SoC design. The area savings offered by the 4T SRAM cells have been one of the main driving forces behind the 4T CMOS SRAM cell development.

However, 4T CMOS SRAM cells are not without drawbacks. Reliable data retention in a 4T cell can be guaranteed only if I_{leak_p} is significantly larger than I_{leak_n} under the worst case PVT conditions.

In [34] and [35], the SNM for conventional 6T SRAM and 4T PRL SRAM cells have been expressed analytically and investigated by computer simulations. In Section 3.5 I have derived an analytical expression for SNM calculation of a 4T loadless SRAM cell that is suitable for stability estimation of such cells.

2.3 Sense Amplifier and Precharge-Equalization

Sense amplifiers (SA) represent an important component in memory design. The choice and design of an SA defines the robustness of bit line sensing, impacts the read speed and power. Due to the variety of SAs in semiconductor memories and the impact they have on the final specs of the memory, the sense amplifiers have become a separate class of circuits.

The primary function of an SA in SRAM is to amplify a small analog differential voltage developed on the bit lines by a read-accessed cell to the full swing digital output signal thus greatly reducing the time required for a read operation. Since SRAM do not feature data refresh after sensing, the sensing operation has to be nondestructive, as opposed to a destructive sensing of a DRAM cell. Having an SA allows the storage cells to be small, since each individual cell need not fully discharge the bit line.

Design constraints of an SA are defined by the minimum differential input signal amplitude, the minimum gain A, and tolerance to the environmental conditions. The gain A, is a function of the initial bit line voltage (precharge level). Gain A influences the sense delay t_{sense}, however a high A does not necessarily reduce t_{sense}. Usually, t_{sense} is traded off for reduced power consumption, layout area and for better tolerance of environmental effects [27].

Special attention is given to the area taken by the SA. Depending on the SRAM array column architecture, the area requirements for the SA may vary. Architectures using column multiplexing can share a single SA among the multiplexed columns such that only one column is connected to the sense amplifier at any given time. The total area available for the SA is defined by a multiple n of the bit line pitch values, where n can normally be from 1 to 16. In turn, the bit line pitch is defined by the size of a memory cell. This example illustrates the complexity of SRAM design and layout planning. The choice of the cell size, number of columns, number of cells per column, the minimum differential swing, the choice of the SA architecture and size are all the factors taken into consideration when designing an SRAM to comply with the target power, speed and reliability.

Generally, the parameters characterizing a sense amplifier include:

- Gain $A = V_{out}/V_{in}$
- Sensitivity $S = V_{in\_min}$ - minimum detectable signal
- Offsets V_{offset} and I_{offset} - the difference at outputs with the common mode signal at the inputs
- Common Mode Rejection Ratio $CMRR = A_{diff}/A_{cm}$ - ratio of amplification for a differential and a common mode signals

- Rise time t_{rise}, fall time t_{fall} - 10% to 90% of the signal transient
- Sense delay $t_{sense} = t_{50\%\_WL} - t_{50\%\_V\_out}$ - where $t_{50\%\_WL}$ - the 50% point of the word line enable signal and $t_{50\%\_V\_out}$ - the 50% point of SA output transient

The optimization of the above parameters is a difficult task involving balancing the circuit complexity, layout area, reliability, power, speed and environmental tolerance. The process spread of the modern DSM technologies can add to the complexity of the SA design by introducing significant parameter mismatches, asymmetry and offsets. Practical SA design is an iterative procedure that has to pay close attention to the fabrication process parameters and their variations in the target technology. The choice of circuit, transistor sizing, operating point, gain and transient response has to be done based on the timing and layout constraints of the particular memory system. To alleviate short-channel effects and the atomistic dopant distribution effects in the channel [36], SA often employ devices with non-minimum length and width. That helps to reduce the asymmetry resulting from transistor V_{TH} and geometry spread and thus mitigate the SA offset caused by the inherent variation in the parameters of the fabricated transistors.

The bit line differential voltage, the reliability and the power consumption of an SA are directly linked. Moreover, the minimal bit line differential is a factor in defining the total read access time and thus, the speed of an SRAM. On the one hand, a larger bit line differential is beneficial for more reliable sensing. However, the resulting better tolerance to the process and environmental fluctuations comes at the cost of the extra read access time and the power spent on the discharging and precharging of the bit lines.

The differential sensing, widely utilized in SRAMs, allows rejection of the common-mode noise that may present on both the bit lines. Noise sources, such as power spikes, capacitive coupling between the bit lines and between the word line and the bit lines, can inject common-mode noise to both SA inputs. The common mode noise is then attenuated

by the value of ratio CMRR and the true differential signal is amplified.

A classical current-mirror differential SA with active load is shown in Figure 2.9. The sensing operation begins with setting the SA operation point by precharging and equalization of both the inputs of the SA (which are the bit lines BL and BLB in Figures 2.9, 2.10) to the identical *precharge* voltage level. Once both BL and BLB are precharged and equalized, the precharged levels are stored in the bit line capacitance C_{BL}. Next, the decoded word line WL of a read-accessed cell is activated starting the build-up of the differential voltage on the bit lines BL and BLB (around 100-200 mV). Once the differential voltage has exceeded the sensitivity S of the SA, a Sense Amplifier Enable (SAE) signal is issued and the SA amplifies the differential voltage on the bit lines to the full-swing output level *out*. Then, SAE and WL are disabled and the read operation is complete.

For reliable operation, current mirror SAs require biasing power to set up in the high-gain region. The minimum biasing current is limited by the minimum required SA gain, CMRR and sensing delay, whereas the maximum biasing current is limited by the power dissipation and the layout area. The gain of a current-mirror SA is given by Equation 2.5, It is typically set around ten. The gain can be increased by widening $Q1$ and $Q2$ or by increasing the biasing current [3]:

$$A = -g_{m\_Q1}(r_{o2}||r_{o4}) \tag{2.5}$$

where g_{m_Q1} is the transconductance of $Q1$, and r_{o2} and r_{o4} are small-signal output resistance of $Q2$ and $Q4$, respectively.

The output resistance of the current-mirror SA is given by $R_{out} = r_{o2}||r_{o4}$. Parameters of g_m and r can be modified by the proper transistor sizing. To ensure high initial amplification, the sizing of $Q1 - Q4$ is chosen such that the bit line precharge level corresponds to the high-gain region of the SA's transfer characteristic. However, the gain A is also a function of the V_{GS} of transistors $Q1$ and $Q2$, and hence a function of the precharge

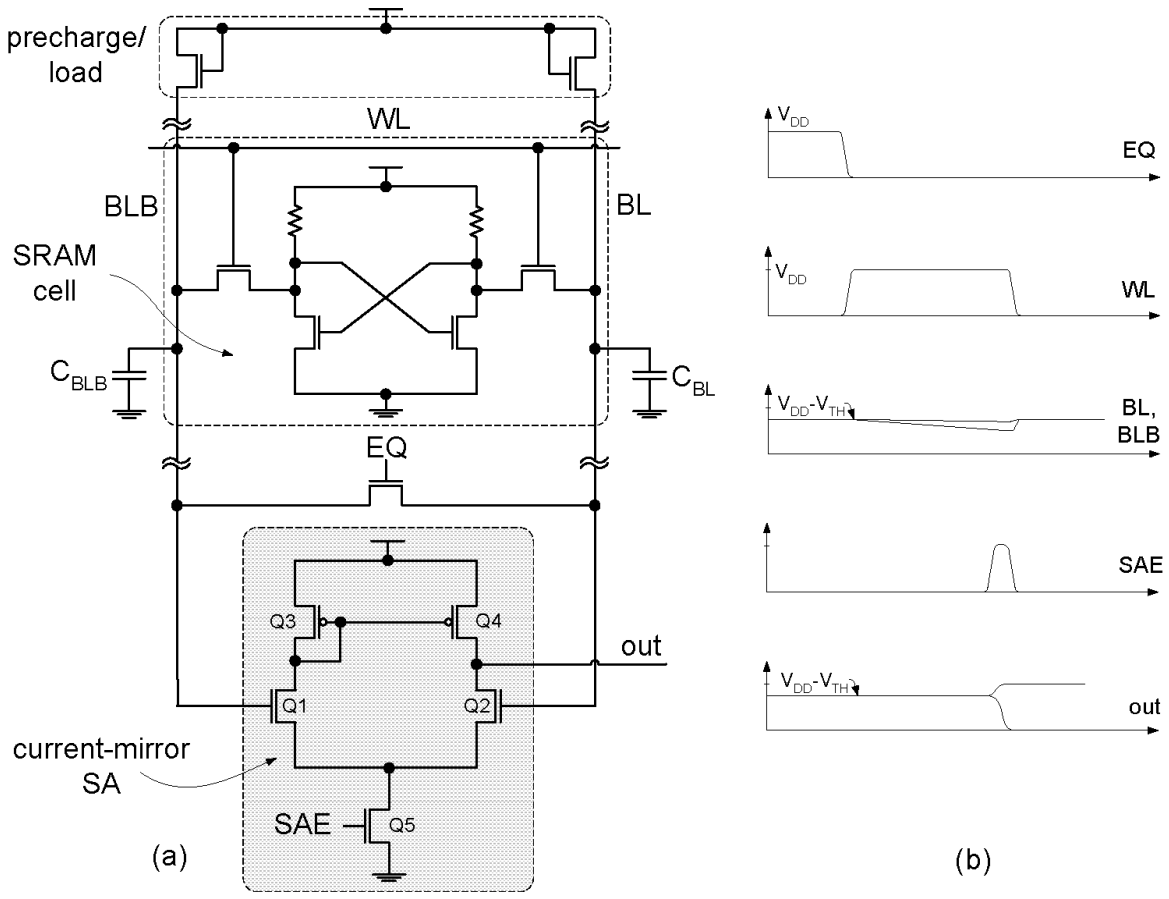


Figure 2.9 A typical circuit with a current-mirror type sense amplifier, a PRL SRAM cell and precharge/load transistors (a); signal waveforms during a read operation (b).

level. The precharge NMOS transistors used to statically precharge the bit lines should be sized so that their contention with the pull-down transistors does not flip the cell. That puts a sizing constraint on the precharge/load transistors. The sizing of these transistors determines the bit line recovery speed, which is especially critical after a write operation when the bit line is completely discharged. High sensitivity to transistor mismatches in a current-mirror SA causes increased offsets. To compensate for possible offsets and main-

tain reliable sensing, the minimum differential voltage has to be increased, slowing down the sensing. Combined with the sizable power consumption and special precharge conditions, this causes the usage of the current-mirror type SA to decline in the scaled-down low-voltage technologies. The circuit in Figure 2.9 does not require SA isolation as the bit lines are connected to the transistor gates and are isolated from the output. The voltage-divider action of the serially-connected driver, access and the precharge/load transistors prevents the complete discharge of the bit lines. Thus, the word line deactivation timing requirements can be relaxed as the bit line discharge will stop at the potential defined by the relative sizing of the precharge/load, access and driver transistors.

A latch-type SA shown in Figure 2.10. This type of a SA is formed by a pair of cross-coupled inverters, much like a 6T SRAM cell. The sensing starts with biasing the latch-type SA in the high-gain metastable region by precharging and equalizing its inputs. Since in the latch-type SA the inputs are not isolated from the outputs, transistors $Q5$ and $Q6$ are needed to isolate the latch-type SA from the bit lines and prevent the full discharge of the bit line carrying a "0", which costs extra power and delay. Due to the presence of the column MUX/isolation transistors, two precharge/equalize circuits are needed to ensure reliable sensing: global precharge/equalization *pre* for the column and a local precharge/equalization *lpre* for the inputs of the SA (Figure 2.10 (a)).

When a cell accessed by the word line WL has discharged the bit lines BL and BLB to a sufficient voltage differential (see Figure 2.10 (b)), the SA is enabled by a high-to-low transition of SAE pulse. Shortly after that, the column MUX/isolation transistors $Q6$ and $Q7$ are turned off, isolating the highly capacitive bit lines from the SA latch and preventing the complete discharge of C_{BL} and C_{BLB}. Then, the positive feedback of the cross-coupled inverters $Q1 - Q3$ and $Q2 - Q4$ quickly drives the low-capacitance outputs *out* and *out*$'$ to the full swing complementary voltages. Note that in the typical circuit presented in

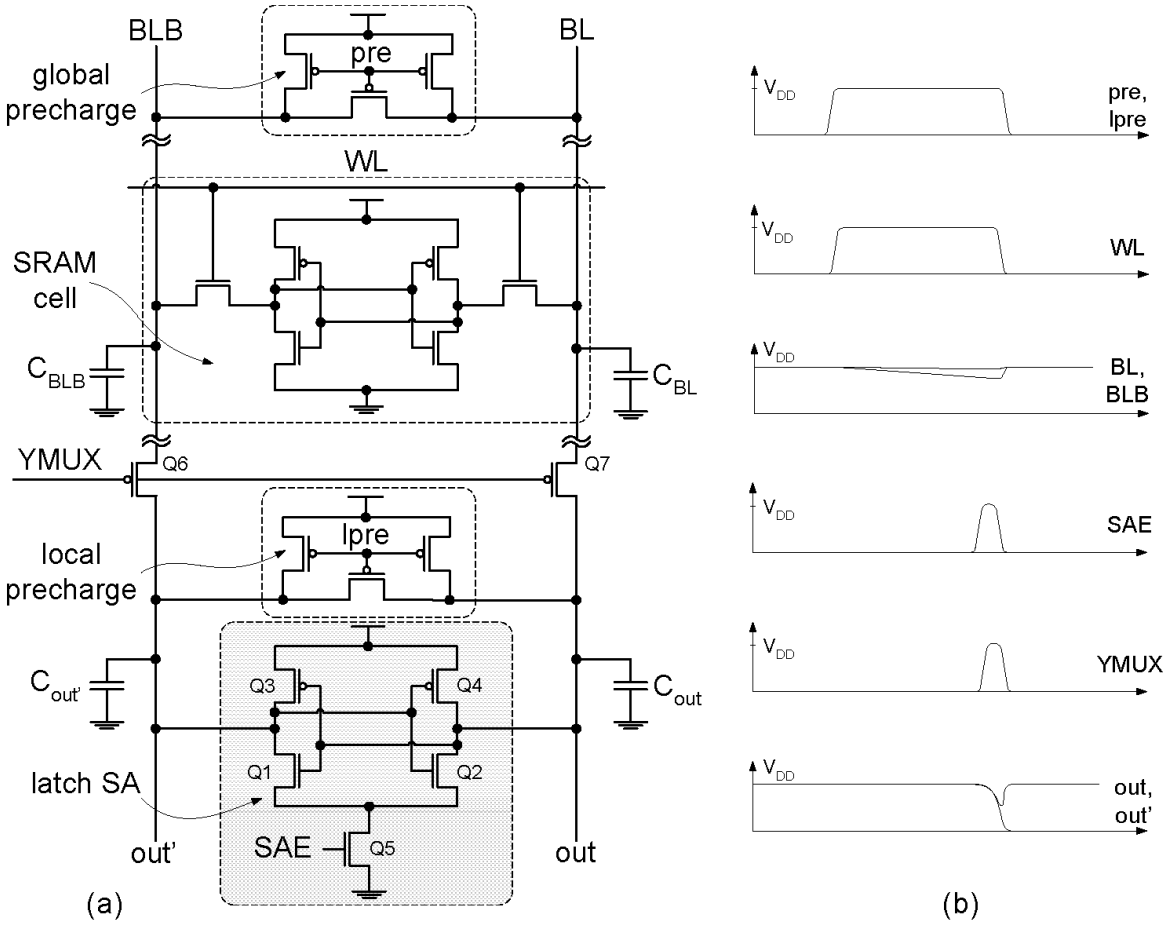
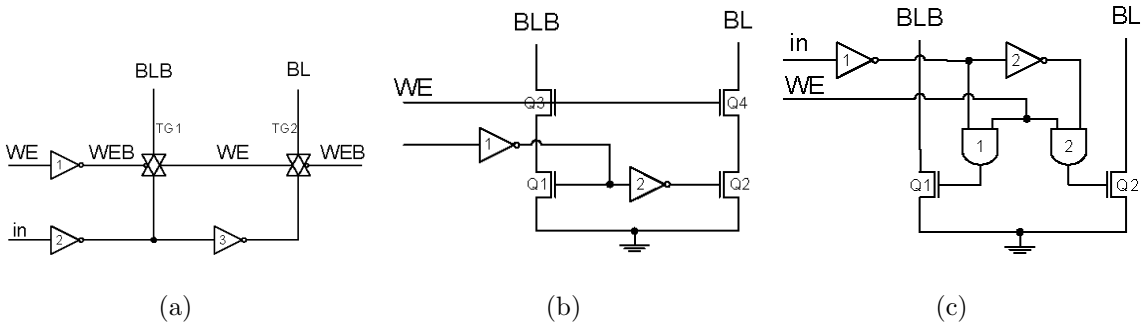


Figure 2.10 **A typical circuit with a latch-type sense amplifier, a full CMOS 6T SRAM cell, column mux and precharge (a); signal waveforms during a read operation (b).**

Figure 2.10, the local and global precharge/equalize circuits are clocked to save power. To improve the noise robustness in low-voltage operation and to ensure a non-destructive read under process variations and minor defects in the cell in modern DSM technologies, the precharge level is typically set to full V_{DD}.

2.4 Write Driver

The function of the SRAM write driver is to quickly discharge one of the bit lines from the precharge level to below the write margin of the SRAM cell. Normally, the write driver is enabled by the Write Enable (WE) signal and drives the bit line using full-swing discharge from the precharge level to ground. The order in which the word line is enabled and the write drivers are activated is not crucial for the correct write operation.



(a) (b) (c)

Figure 2.11 Write driver circuits.

Some of the typical write driver circuits are presented in Figure 2.11. The circuit in Figure 2.11(a) writes the input data *in* and its complement buffered by inverters 2 and 3 to the bit lines BL and BLB through two transmission gates $TG1$ and $TG2$. WE and its complementary WEB are used to activate $TG1$ and $TG2$ and discharge BL or BLB through the NMOS transistors in inverter 2 or 3. The write driver presented in Figure 2.11(b) uses two stacked NMOS transistors to form two pass-transistor AND gates using transistors $Q1, Q3$ and $Q2, Q4$. The sources of NMOS transistors $Q1$ and $Q2$ are grounded. When enabled by WE, the input data *in* enables, through inverters 1 and 2, one of the transistors $Q1$ or $Q2$ and a strong "0" is applied by discharging BL or BLB from the precharge level to the ground level. Another implementation of the write driver is presented in Figure 2.11(c). When WE is asserted, depending on the input data *in*,

inverters 1 and 2 activate one of two two-input AND gates 1 and 2 to turn on one of the pass-transistors $Q1$ or $Q2$. Then, the activated transistor discharges the corresponding bit line to the ground level.

Even though a greater discharge of the highly capacitive bit lines is required for a write operation, a write operation can be carried out faster than a read operation. Only one write driver is needed for each SRAM column. Thus, the area impact of a larger write driver is not multiplied by the number of cells in the column and hence the write driver can be sized up if necessary.

2.5 Row Address Decoder and Column MUX

Address decoders allow the number of interconnects in a binary system to be reduced by a factor of $\log_2 N$, where N is the number of independent addressed locations. The memory address space is defined as the total number of the address bits required to access a particular memory cell or word for bit-oriented memories and word-oriented memories, respectively. For instance, the total address space in a 1Mb bit-oriented SRAM will be 20 (2^{20}=1Mb) address bits A0...A19. On the other hand, in a 1Mb word-oriented SRAM with a 32-bit (2^5) word width, which can be organized in 32 blocks each of which has 256 rows and 128 columns, the address space reduces to 15 ($2^{(20-5)} = 2^{15}$) address bits A0...A14.

The SRAM row decoder can be of a single- or multi-stage architecture. In a single-stage decoder all decoding is realized in a single block. The multiple-stage decoding utilizes several hierarchically-linked blocks. Normally, the most significant bits are decoded ($pre-decoded$) in the first decoder stage, effectively selecting the array that is to be accessed by providing enable signals for the subsequent decoder stage(s) that enable a particular word line. The number of outputs of the last decoding stage corresponds to the number

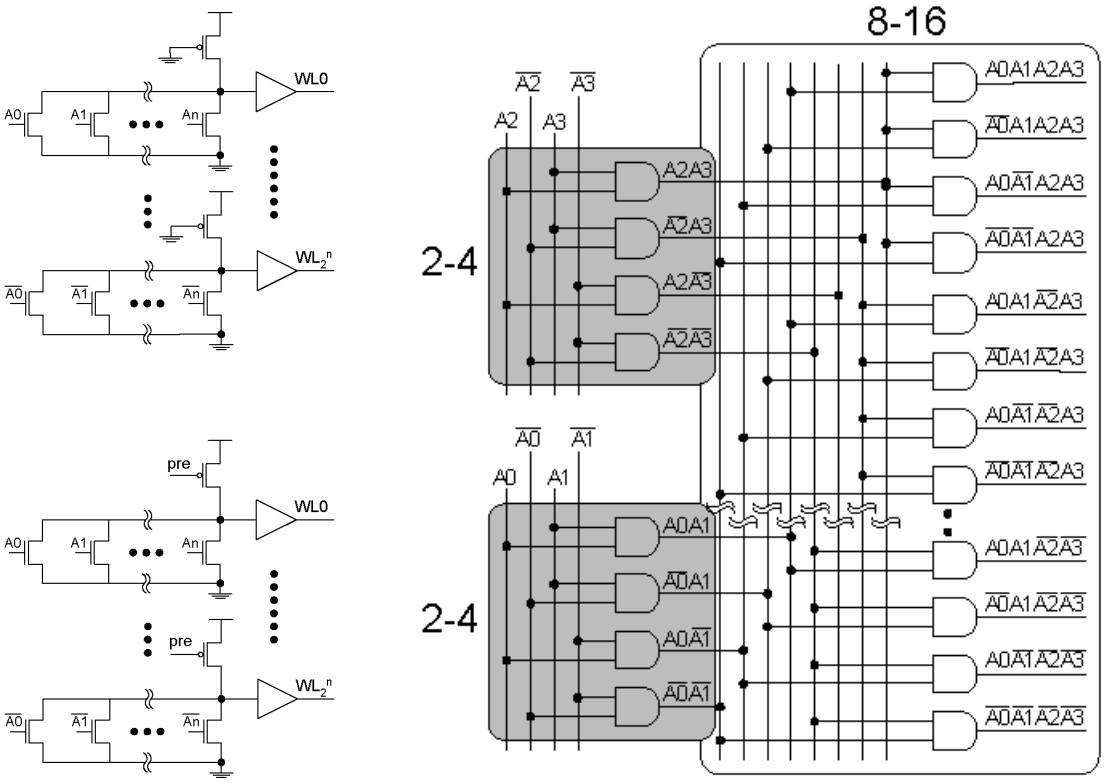
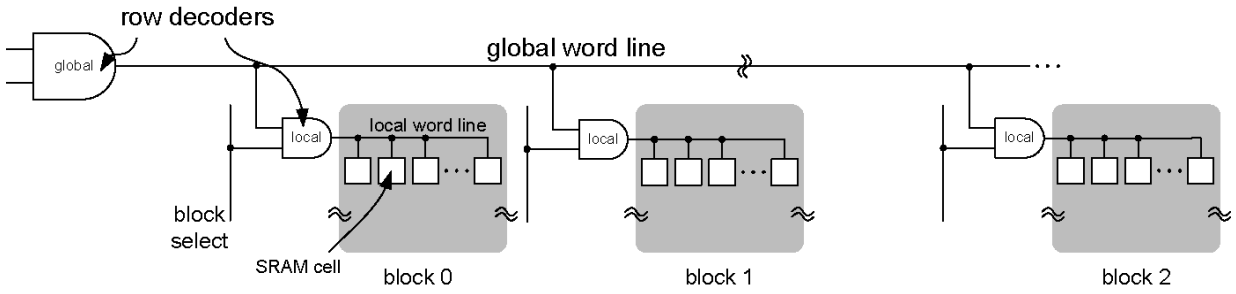


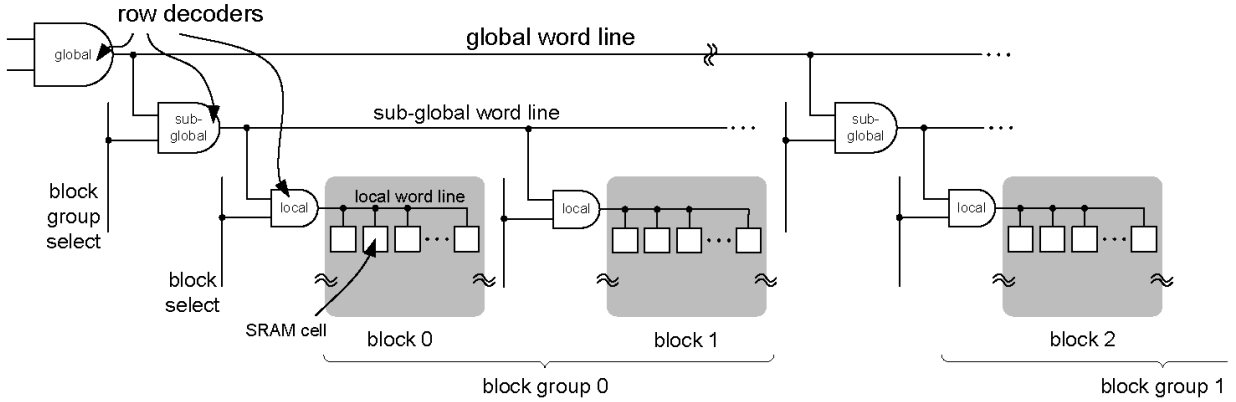
Figure 2.12 (a) Single-stage static (top) and dynamic (bottom) decoders; (b) Multi-stage static 4-16 decoder.

of rows (word lines) to be decoded. An example of single-stage static decoders is shown in Figure 2.12(a). Each word line is decoded in a single stage by a wide NOR gate with a fan-in equal to the number of row address bits. To simplify the circuit and reduce the layout area, such decoders are often designed using a static PMOS transistor load (top circuit in Figure 2.12(a)). Another variation of this decoder can use a clocked precharge PMOS transistor (bottom circuit in Figure 2.12(a)).

Single-stage row decoders are attractive for use with small single-block memories. How-



(a) Divided Word Line (DWL) row decoder architecture.



(b) Hierarchical Word Decoding (HWD) [37] row decoder architecture.

Figure 2.13 Multi-stage row decoder architectures.

ever, most memories today split the row address space into several blocks decoded by separate decoder stages. This approach is proven to be more power efficient and fast for large memories consisting of multiple arrays. The example of multi-stage 4–16 decoder in Figure 2.12(b) uses two 2-4 decoders and a 8–16 stage. Address bits A0,A1 and A2,A3 are predecoded separately by the 2–4 decoders, which drive eight address lines. A set of 16 AND gates further decodes the predecoded combinations of A0,A1 and A2,A3 into 16 outputs.

All large decoders are realized using at least a two-stage implementation [3]. The

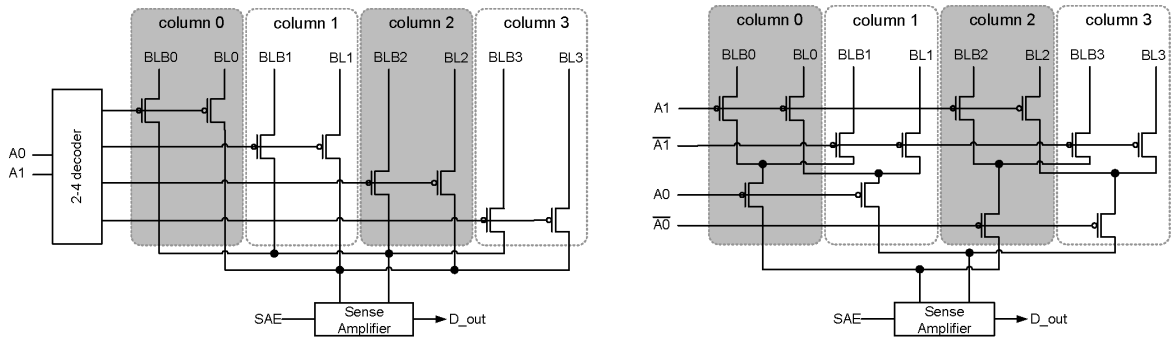


Figure 2.14 (a) 4-1 pass-transistor column decoder with a predecoder; (b) 4-1 tree-based column decoder.

conventional Divided Word Line (DWL) structure shown in Figure 2.13(a) partitions the SRAM into blocks. A local (block) word line is activated when both the global word line and the block select are asserted. Since only one block is activated, the DWL structure reduces both the word line delay and the power consumption. An additional decoding level, coined Hierarchical Word Decoding (HWD) architecture, has been proposed for larger than 4Mb SRAMs (Figure 2.13(b)) to cope with the growing delay time and power consumption [37]. The HWD offers $\approx 20\%$ delay and $\approx 30\%$ total load capacitance reduction over the DWL decoder architecture.

A column MUX in an SRAM comprises a 2^K-input multiplexer for each of the bit lines, where K is the size of column address word. A column MUX allows several columns to be connected to a single SA and thus, relax the area constraints on the SA design. An example of a typical column MUX using pass-transistors and a 2–4 predecoder is shown in Figure 2.14(a). PMOS transistors enabled by one of the outputs of the 2–4 predecoder pass the read differential voltage from the bit lines of one out of four columns to the inputs of a differential SA. The simpler version of a column MUX shown in Figure 2.14(b) uses

a binary tree decoder formed of PMOS pass transistors. This column MUX requires no predecoding and utilizes fewer transistors. However, since the propagation delay increases quadratically with the number of sections, a large tree-based column MUX introduces extra delay and its usage may be prohibitively slow for large decoders [3].

2.6 Timing Control Schemes

The timing control block controls precharge, word line, sense amplifier clocking and write driver activation to ensure the correct write and read operations. Technology scaling poses extra challenges for accurate timing generation. As the gate overdrive voltage is reduced with every generation of DSM process [38], V_{TH} fluctuations and process variability across the process corners are growing [39].

The key aspect of the precharged SRAM read cycle is the timing relationship between the RAM addresses, the precharge deactivation, enabling of the row and column decoders and SA activation. If the asserting of the word line precedes the end of the precharge cycle, SRAM cells on the activated word line will see both the bit lines pulled high and the accessed cells may flip state. Another timing hazard may arise if the address changes before the read operation is complete i.e., when precharge is deactivated. In this case more than one SRAM cell will be discharging the bit lines which may lead to reading erroneous data. If the SA is enabled during the write operation, a "write through" can occur and the data being written will appear at the output without an intended read operation. Fundamentally, the signal path delay should match to the clock path delay for correct, fast and power-efficient SRAM operation. Typically, the delay variations are dominated by the bit line delay since the minimal-size transistors in SRAM cells are more susceptible to process variations. The timing control block should provide sufficient timing margins

to account for the worst-case process conditions.

Basic timing control methods employed in SRAMs include:

1. Timed by the clock phase (direct clocking) [40].

2. Delay line using a multitude of inverters to define the timing intervals [41].

3. Self-timed replica (dummy) loop mimicking the signal path delay [38].

Timing method using the direct clocking of the WL and SA has limited operation speed due to the larger timing margins necessary for reliable operation. The delay line allows faster operation than the direct clocking method. However, the delay of the delay loop may not track the delay variations caused by the process variations in modern DSM technologies. The self-timed replica (dummy) loop method has proven itself to be the most robust and precise in tracking the process variations and in maintaining tighter timing margins for faster operation. Since Delay-Line based (2) and Replica-Loop based (3) timing methods have been used in this work to design SRAM test chips, we will discuss them in more detail.

2.6.1 Delay-Line Based Timing Control

A functional diagram of a delay line timing loop is shown in Figure 2.15(a) on the following page. A control signal $ctrl\_in$ sets the FSM. The timing loop is defined by the total delay through the delay elements $t_{delay1} - t_{delay\_n}$ in the FSM reset path. Typically, delay elements are based on serially-connected inverters. The delay time can be extended by using non-minimal length devices in the delay inverters or by utilizing current-starved inverters. The timing intervals formed by the delay elements $t_{delay1} - t_{delay\_n}$ with the complementary logic are used to generate the control signals for the read/write timing. The delay-line timing

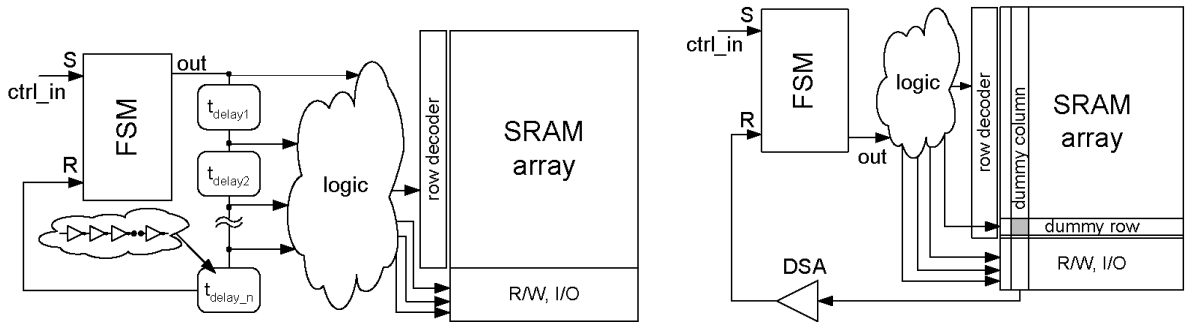


Figure 2.15 (a) Delay line timing loop; (b) Replica timing loop.

technique has been implemented in a SRAM test chip fabricated in $0.18\mu m$ CMOS TSMC technology we will discuss in more detail in Chapter 6.

2.6.2 Replica-Loop Based Timing Control

The replica-loop based timing method provides a tighter tracking of the bit line discharge delay. A replica (dummy) column and row containing the same number of SRAM cells as in the main array are used as delay elements (Figure 2.15(b)). The replica delay path mimics the capacitive loads of the real signal delay path and provides more precise timing for deactivation of the word line and activation of the SA. Similarly to the delay-line based method, control signal *ctrl_in* sets the FSM. The output signal *out* initiates activates the word lines both in the decoded row and in the dummy row. Once the dummy bit line has been discharged to the dummy SA (DSA) switching threshold, the DSA flips and resets the FSM. Next, the sense amplifier enable (SAE) signal is issued and the differential voltage on the active column is amplified to the full swing. The time it takes to discharge the dummy bit line to the switching threshold of the DSA is designed to be the same as the time required for a worst-case SRAM cell to develop sufficient differential voltage on the

active bit lines. Thus, bit line discharge is stopped and the SAE signal is issued as soon as a reliable read operation is guaranteed. The resulting read power savings and access time shortening make the replica-loop based timing technique popular in DSM SRAM designs. The power dissipation overhead associated with the switching of the dummy column on each memory access is inversely proportional to the number of simultaneously accessed columns.

The replica timing technique has been implemented in an SRAM test chip fabricated in $0.13\mu m$ CMOS IBM technology and we will discuss it in more detail in Chapter 6, Section 6.4.

2.7 Address Transition Detector

In an asynchronous SRAM a read or a write operation is initiated by an address change or chip enable signal, whereas in synchronous SRAMs a read or write operation is initiated by the system's master clock. The terms asynchronous and synchronous relate to the memory-system interface rather than to internal chip operation.

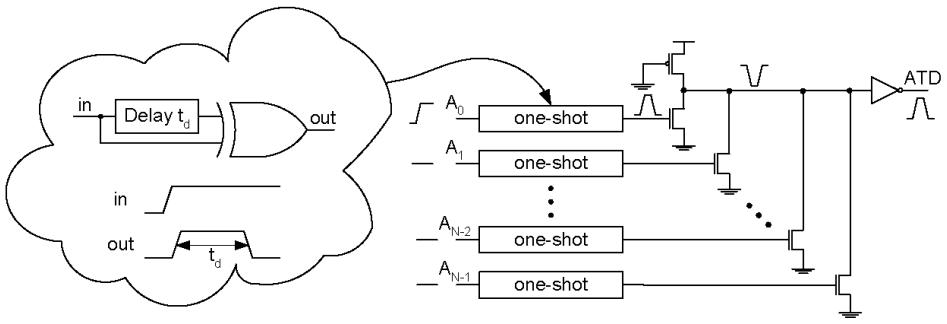


Figure 2.16 Address Transition Detector (ATD) [3].

An asynchronous SRAM features an Address-Transition Detector (ATD) that produces a pulse t_d of a controlled duration on every address transition. An ATD is typically

implemented as a wide-OR gate comprising a set of one-shot circuits as shown in Figure 2.16 on the preceding page. Every transition of address bus $A_0 - A_{N-1}$ produces a pulse on the output, which initiates a read or a write operation.

Although traditionally an asynchronous timing interface was used in stand-alone SRAM chips while embedded SRAMs were predominantly synchronous, recent large on-die caches seem to have reverted to an asynchronous interface with the processor core. The asynchronous interface that has been employed in the dual-core Itanium processor with 24MB on-die L3 cache totalling 1.47 billion transistors [19] is reported to reduce the data latency from 8 to 5 clock cycles. The switch to an asynchronous interface in such large caches helps to combat the delay associated with the clock skew over a large die area, the latch delay, and the margins in each cycle. In addition, significant margin must be added to the SRAM cell access cycle to account for slow, marginal cells that are statistically probable in large caches. Besides the performance benefits, an asynchronous design eliminates all clock distribution and associated latch power.

2.8 Summary

In this chapter, an overview of the main SRAM building blocks is presented. We discussed polysilicon resistor (PCL) and full CMOS six-transistor SRAM cells, the current mirror and the latch sense amplifiers with the corresponding precharge and equalization circuits, several types of write drivers, row and column decoders and architecture strategies, timing control based on delay elements and replica row/column and, finally, the address transition detector found in asynchronous SRAMs. This chapter serves to provide the background for the following chapters. Some of the discussed circuits and circuit techniques have been implemented in the two test chips designed to prove the concept of the proposed SRAM

cell stability DFT techniques.

The following chapter introduces the concept of static noise margin and provides insight into the sensitivity of the six-transistor SRAM cell's SNM to PVT and defect resistance.

Chapter 3

SRAM Cell Stability Characterization

In this chapter SRAM cell noise margin and stability and the factors affecting them are discussed. The motivation in Section 3.1 is followed by an introduction to the concept of the Static Noise Margin in Section 3.2. Section 3.3 presents SNM definitions found in the literature. A comprehensive SNM sensitivity investigation to PVT and defect resistance is conducted on a Philips 0.12μm SRAM cells is presented in Section 3.4. An analytical SNM expression for a loadless four-transistor SRAM cell using the α-power law model in Section 3.5 is derived. And finally, the results are summarized in Section 3.6.

3.1 Motivation

The semiconductor industry is being constantly pushed to improve the performance and reliability and decrease the cost per function of the products. One of the solutions is

to scale down the physical dimensions of circuit components. Constant-voltage scaling (CVS) dominated VLSI design down to $0.8\mu m$ CMOS technology. Velocity saturation due to the increasing electric field in short-channel transistors created a situation when further constant-voltage scaling did not give a performance advantage over the constant-field scaling [3]. Contrarily, since the drain current in short-channel transistors is no longer a quadratic function of drain voltage (see Section 3.5.1), the gain of an increased drain current has become less important than the penalty of a higher voltage causing the increased power dissipation. Constant-field scaling (CFS) requires a decrease of supply voltage and, as will be shown later in this chapter, leads to a decrease in noise margins of scaled-down SRAM cells. Moreover, other effects of elevated electric fields in scaled-down transistors such as gate oxide breakdown and leakage, and the hot-carrier effect have contributed to making constant-voltage scaling unsuitable for the deep sub-micron modern CMOS technologies. Reducing physical dimensions of scaled-down transistors, such as gate length, oxide thickness, atomistic distribution of dopant atoms in the channel, have introduced a significant degree of uncertainty in the prediction of electrical properties of manufactured transistors. Combined with reduced supply voltage, technological process spread can compromise the noise robustness of circuits.

The reduced metal routing width and pitches of traditional aluminum metallization with the tungsten via plugs in the scaled technologies has exacerbated the delay contribution of the interconnects [4]. A tighter routing pitch also brought attention to the electromigration caused by the increased inter-metal electric fields. To alleviate these issues, starting from the CMOS $0.18\mu m$ process node, the semiconductor foundries have adopted copper dual-damascene metallization. However, copper processing issues and the smaller contact and via dimensions have led to increasing problems with unreliable resistive interconnects. Such non-catastrophic resistive defects pose an additional threat to SRAM reliability, which will

be addressed in this work.

In this chapter we will discuss noise margin definitions applied to SRAM cells and conduct a comprehensive SNM sensitivity analysis case study on a 6T SRAM cell in CMOS $0.13\mu m$ technology. We will show how process and environmental parameters, as well as non-catastrophic defects, adversely impact the SNM of an SRAM cell. We will also propose an analytical expression for the noise margin calculation of a four-transistor loadless SRAM cell.

3.2 Introduction

The noise margin can be defined using the input voltage to output voltage transfer characteristic, a.k.a. Voltage Transfer Characteristic (VTC). In general, the Noise Margin (NM) is the maximum spurious signal that can be accepted by the device when used in a system while still maintaining the correct operation [42]. If the consequences of the noise applied to a circuit node are not latched, such noise will not affect the correct operation of the system and can thus be deemed tolerable. It is assumed that noise is present long enough for the circuit to react, i.e. the noise is "static" or *dc*. A Static Noise Margin is implied if the noise is a *dc* source. In the case when a very long noise pulse is applied, the situation is quasi-static and the noise margin asymptotically approaches the SNM (Figure 3.1) [43].

An ideal inverter would have tolerated a change in the input voltage (V_{in}) without any change in the output voltage (V_{out}) until the input voltage reaches the switching point. The switching point is presented in Figure 3.2(a) as $|\partial V_{out}/\partial V_{in}| = 1$. The switching point of an ideal inverter is equidistant from the logic levels. At the switching point, an ideal inverter demonstrates an absolutely abrupt change in V_{out} such that $|\partial V_{out}/\partial V_{in}| = \infty$. In other words, it has an infinite slope (gain) in the transition region. Thus, the valid logic

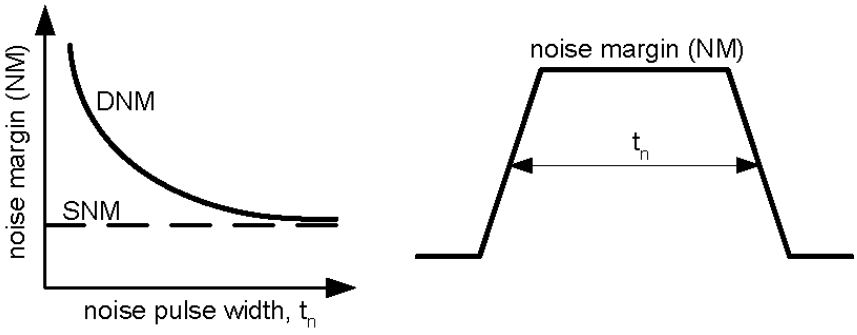


Figure 3.1 The reduction of the Dynamic Noise Margin (DNM) with the increase of the noise pulse width t_n. For very high t_n DNM approaches its minimum (SNM).

levels (noise margins) of an ideal inverter span from the power rails to the asymptotical proximity of the transition point.

However, in a real inverter the switching point is not equidistant from the logical levels and the transition region is characterized by a finite slope $\infty > |\partial V_{out}/\partial V_{in}| > 1$, as shown in Figure 3.2(b). The finite slope $\partial V_{out}/\partial V_{in}$ in the transition region of a real inverter creates uncertainty as to what point on the VTC should be used in determining the noise margin value. This is overcome by considering a chain of inverters rather than a single inverter [42],[44].

Lohstroh et al. [44] have shown that an infinite chain of inverters is identical to a flip-flop, which forms the basis of an SRAM cell. The advantages and disadvantages of the existing noise margin definitions and their applicability to the SRAM cells noise immunity analysis are discussed in Section 3.3.

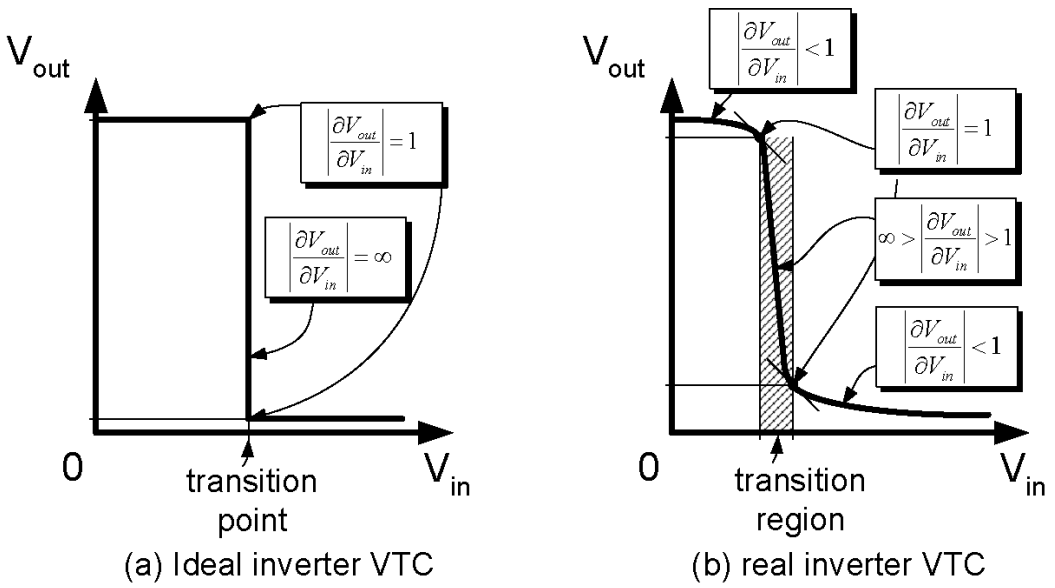


Figure 3.2 Voltage Transfer Characteristic (VTC) of an ideal (a) and a real (b) inverter. Values of $|\partial V_{out}/\partial V_{in}|$ represent inverter gains depending on the input voltage.

3.3 SNM Definitions

3.3.1 Inverter V_{IL}, V_{IH}, V_{OL} and V_{OH}

Several definitions of the SNM can be found in the current textbooks and one standard commonly used definition is yet to achieve universal acceptance. In textbooks [45],[46],[47], the noise margin high and noise margin low are defined as 3.1 and 3.2, respectively:

$$NM_H = V_{OH} - V_{IH} \tag{3.1}$$

$$NM_L = V_{IL} - V_{OL} \tag{3.2}$$

where V_{IL} is the maximum input voltage level recognized as logical "0", V_{IH} is the minimum input voltage level recognized as a logical "1", V_{OL} is the maximum logical "0" output voltage, V_{OH} is the minimum logical "1" output voltage as illustrated in Figure 3.3.

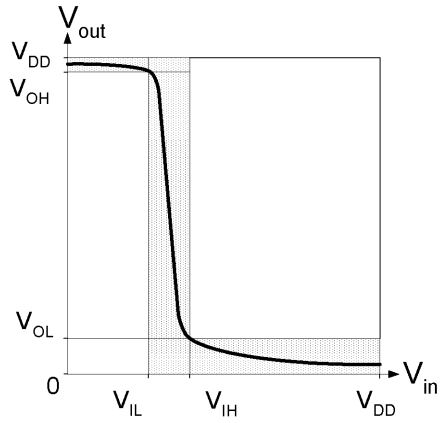


Figure 3.3 Definition of V_{IL}, V_{IH}, V_{OL} and V_{OH} in Equations 3.1 and 3.2. The inverter transfer curve must lie within the shaded region to be a member of the logic family.

Any inverter transfer curve, which falls into the shaded area, will have noise margins at least as good as given by the equations above. This approach specifies the compatibility of the logic levels of circuits in the same logic family. To provide the correct signal interaction without the need to employ the level conversion circuitry, the output logic levels of one circuit have to be compatible with the input logic levels of the next stage. For instance, input voltage in the range of $V_{IL} < V_{in} < V_{IH}$ may not be properly recognized by the gate and may cause a logic error.

3.3.2 Noise Margins NM_H and NM_L with V_{OL} and V_{OH} defined as stable logic points.

This is one of the ways to define V_{OH} and V_{OL}. In this approach, V_{OH} and V_{OL} are represented as the stable voltage states of a bistable inverter pair. The resulting values of NM_L and NM_H are larger than for other definitions. Defining V_{OH} and V_{OL} like this runs into trouble with the very basic and simple concepts discussed in connection with

logic level definitions depicted in Figure 3.3. The transfer characteristic does not lie within the required shaded area and thus cannot represent a set of valid logic level definitions from which any meaningful noise margins can be calculated. Thus, this approach must be rejected as a valid noise margin approach in spite of the fact that it is used by several highly respected digital electronics textbooks [42],[48].

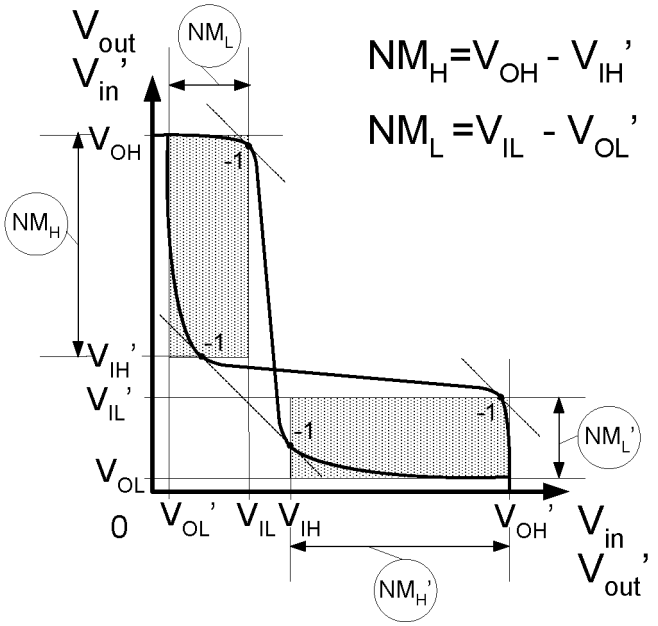


Figure 3.4 Graphical representation of SNM with V_{OH} and V_{OL} in Equations 3.1 and 3.2 as stable logic state points of a bistable inverter pair.

3.3.3 Noise Margins NM_H and NM_L with V_{OL} and V_{OH} defined as -1 slope points.

This is another way to represent the noise margins. In this approach V_{OH} and V_{OL} are represented as the stable points where the $dVout/dVin = -1$ of a bistable inverter pair and coincide with V_{IL} and V_{IH} respectively. That makes more sense since V_{OL} is defined

as the maximum output voltage level with the gate at "0" logical state, and V_{OH} is the minimum output voltage with "1" logical level. The resulting values of NM_L and NM_H are less than for the definition in Section 3.3.2.

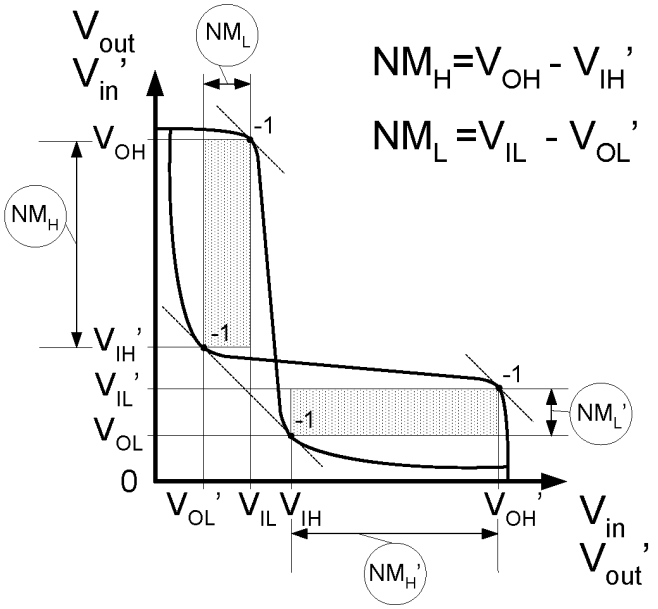


Figure 3.5 Graphical representation of SNM with V_{OH} and V_{OL} in Equations 3.1 and 3.2 as -1 slope points of a bistable inverter pair.

This technique represents more of a legacy from considering the noise margins of a inverter chain than a bistable inverter pair, when the mirrored transfer characteristics were not considered on the same coordinate system. The V_{OH} and V_{OL} levels have theoretical justification in this approach. The sum of NM_L and NM_H is maximal if the V_{OH} and V_{OL} points are chosen to be at the -1 slope points [44]. However, the application of this approach requires imposing some restrictions on the shape of the transfer curves as it maximizes the sum of the noise margins and not the individual noise margins. This theoretically can lead to a situation, when the maximal sum occurs at points where one of the noise margins is zero or even negative [48]. For the transfer curves of an SRAM

cell, however, this criterion produces reliable results. Another advantage of this approach is that it allows analytical calculation of the noise margins of an SRAM cell.

3.3.4 SNM as a Side of the Maximal Square Drawn Between the Inverter Characteristics

The approach was first described by Hill [42] in 1968. An important advantage of this method is that it can be automated using a DC circuit simulator, which to a great extent extends its practical usefulness. In this approach an SRAM cell is presented as two equivalent inverters with the noise sources inserted between the corresponding inputs and outputs Figure 3.6. Both series voltage noise sources (V_n) have the same value and act together to upset the state of the cell, i.e. have an "adverse" polarity to the current state of each inverter of the cell. Applying the adverse noise sources polarity represents the worst-case equal noise margins [48]. This method is only applicable to circuits with $R_{in} \gg R_{out}$, and CMOS inverters of an SRAM cell comply with this condition.

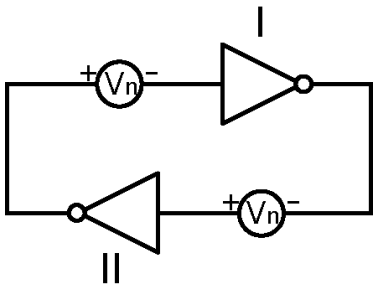


Figure 3.6 Flip-flop with two noise sources with adverse polarities.

Having two adverse noise sources applied to the input of each inverter of an SRAM cell makes the value of the obtained SNM to be the worst-case SNM as shown in Figure 3.7 on the following page [44]. In contrast, the best-case SNM would be obtained if only one

noise source was applied or the polarities of the noise sources were not adverse. However, one is rarely interested in such an idealized case.

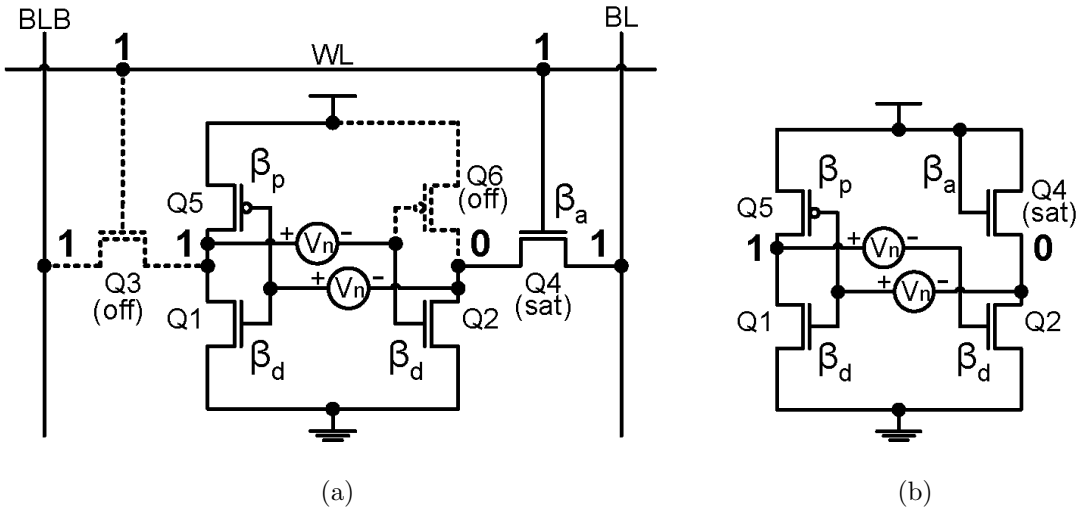


Figure 3.7 Read-accessed SRAM cell with inserted adverse polarity static noise sources V_n (a) and its equivalent circuit (b).

Figure 3.8 shows the superimposed normal inverter transfer curve and its mirrored with respect to $x = y$ line counterpart of a read-accessed 6T SRAM cell in a $x - y$ coordinate system. The $u - v$ system of coordinates is rotated around the same origin by 45^0 counter clockwise with respect to $x-y$ system. This is a convenient arrangement since knowing the diagonals of the maximum embedded squares we can calculate the sides, and the v axis is parallel to the sought diagonals. The dashed curve in the $u-v$ coordinate system represents the subtraction of the normal and mirrored inverter transfer curves in the $x-y$ coordinate system. Since the squares have maximum size when the lengths of their diagonals D_1 and D_2 are maximal, the extremes of this curve correspond to the diagonals of the maximal embedded squares. Generally, due to the process spread, $D_1 \neq D_2$. Suppose that $D_1 > D2$. Then $D_1/\sqrt{2}$ yields the SNM of the flip-flop.

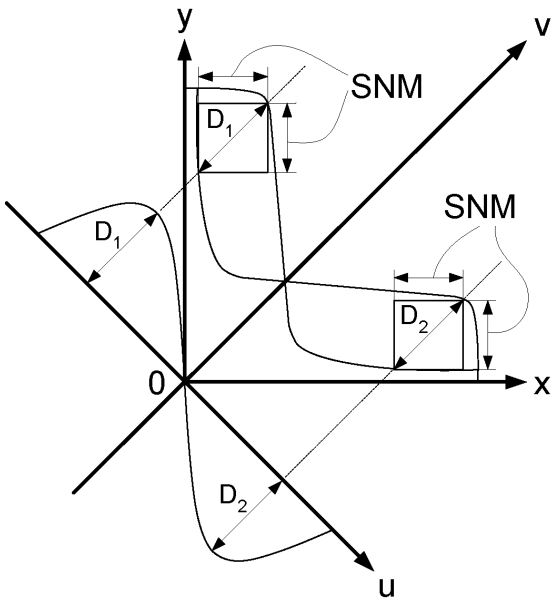


Figure 3.8 SNM estimation based on "maximum squares" in a 45^0 rotated coordinate system. The voltage transfer characteristics (VTCs) of both inverters comprising an SRAM cell are ideally symmetrical.

The above algorithm can be expressed mathematically in the following manner. Assume that the normal and mirrored inverter characteristics are defined as functions $y = F_1(x)$ and $y = F_2'(x)$, where the latter is the mirrored version of $y = F_2(x)$. To find F_1 in terms of u and v, the $x - y$ coordinate system has to be transformed as follows [34]:

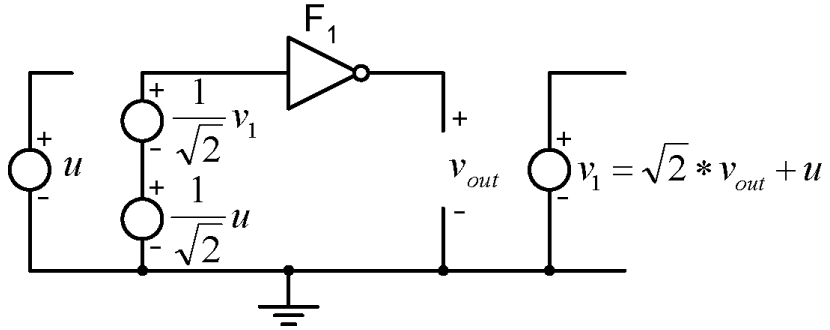
$$x = \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v \tag{3.3}$$

$$y = -\frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v \tag{3.4}$$

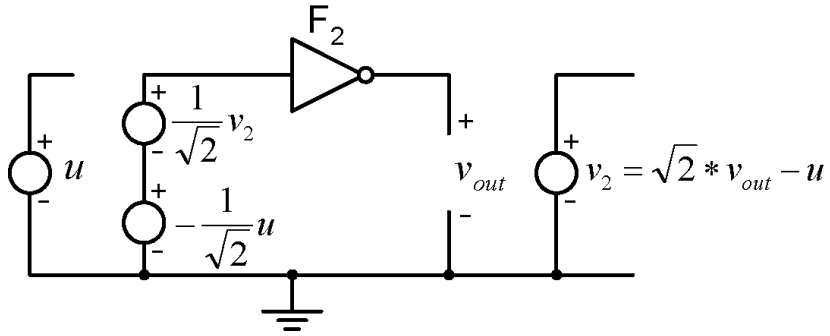
substitution of Equations 3.3 and 3.4 in $y = F_1(x)$ gives:

$$v = u + \sqrt{2}F_1\left(\frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v\right) \tag{3.5}$$

For F_2', first F_2 is mirrored in the $x - y$ system with respect to line $x = y$ (v axis) and



(a)



(b)

Figure 3.9 Circuit implementation of Equations 3.5 (a) and 3.8 (b) for finding the diagonal of the square embedded between the direct and mirrored SRAM flip-flop inverter curves.

then transformed using the same technique as in Equations 3.3 and 3.4 but with x and y interchanged, which produce:

$$x = -\frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v \tag{3.6}$$

$$y = \frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v \tag{3.7}$$

substitution of 3.6 and 3.7 in $y = F_2(x)$ gives:

$$v = -u + \sqrt{2}F_2\left(-\frac{1}{\sqrt{2}}u + \frac{1}{\sqrt{2}}v\right) \tag{3.8}$$

Equations 3.5 and 3.8 explicitly express v as a function of u. Solutions of Equations 3.5 and 3.8 can be found using a standard HSPICE-like DC circuit simulator by translating the equations into circuits with voltage-dependent voltage sources in a feedback loop [34].

The solutions of equations 3.5 and 3.8 are represented by v_1 and v_2 in Figure 3.9. The difference between the two solutions v_1 and v_2 is represented by the sine-like curve in Figure 3.8 on page 57. The absolute values of the extremes of this curve (D) where $dD/du = 0$ represent the lengths of the diagonals of the squares embedded between the direct and mirrored SRAM flip-flop inverter curves. Multiplication of the smaller of the two by $1/\sqrt{2}$ yields the worst-case SNM of an SRAM cell.

3.4 Stability Sensitivity Study

A 6T SRAM cell and its corresponding voltage transfer characteristic (VTC) for a non-ideal cell is depicted in Figure 3.10.

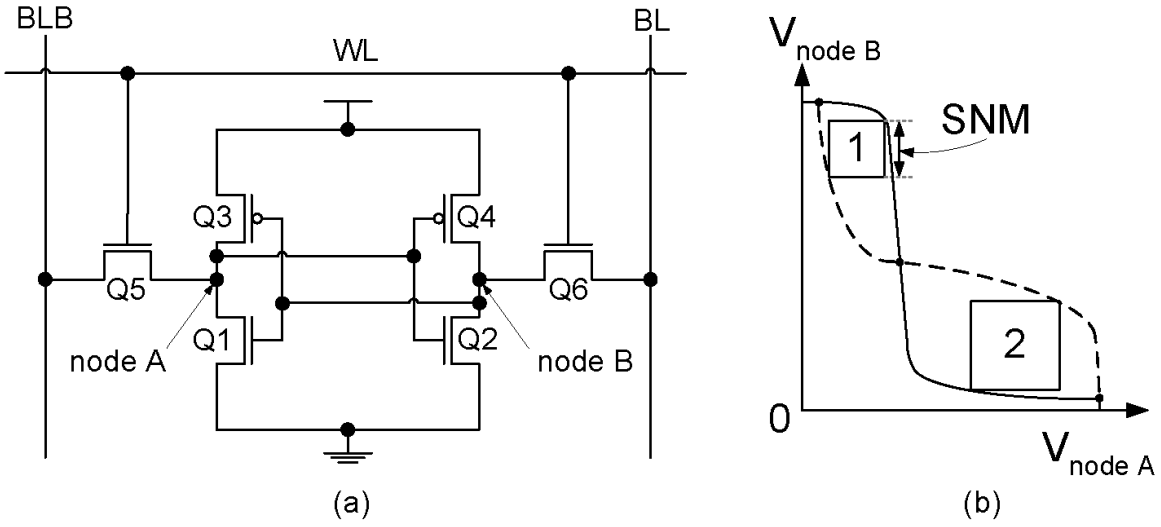


Figure 3.10 A 6T SRAM cell (a) and its SNM definition (b).

In an ideal SRAM cell, the VTC of both halves of such a cell would be perfectly symmetrical and the squares 1 and 2 between the VTC curves (Figure 3.10(b)) would be equal. However, in reality, process spreads and non-catastrophic resistive defects can change the shape of the VTC curves. We define the SNM as the side of the smaller of the two squares that can be fit in the eyes of the VTC curves, as shown in Figure 3.10(b). In our SRAM SNM sensitivity analysis, all measurements were taken in a *read-accessed* cell. That gives the worst case SNM [34] as the saturated access transistor $Q5$ is effectively shunting $Q3$, and $Q6$ is shunting $Q4$, which degrades the stored low-level state and reduces the SNM as shown in Figure 3.11 on the following page. Effectively, a $CMOS$ inverter $Q2 - Q4 - Q6$ is turned into a *ratioed* inverter. Consequently, the ideal logic "0" level of a CMOS inverter turns into the non-ideal "0" level above the ground potential.

We investigated the SNM sensitivity to the process spread (V_{TH}, L_{EFF}, and W_{eff}), the presence of non-catastrophic defects (resistive bridges and breaks), and the variation of operating voltages (V_{DD}), (V_{BL}) and (V_{WL}) for a 6T SRAM cell in 0.13 μm CMOS technology with V_{DD}=1.2 V using special SRAM transistor models and more relaxed design rules. The presented data are normalized with respect to the typical case (typical process corner, room temperature, typical voltages) according to Equation 3.9.

$$SNM_{relative} = \frac{SNM_X - SNM_{TYP}}{SNM_{TYP}} \cdot 100\% \tag{3.9}$$

To obtain a deeper insight into the SNM sensitivity, we applied variations of the process and environment variables. Defect resistances inserted into the cell netlist were swept from 1kΩ to 100GΩ, which covers the whole range of possible defect resistances in an SRAM cell. In the subsequent sections we used the following transistor notation (Figure 3.10(a)): $Q1$ and $Q2$ – driver transistors, $Q3$ and $Q4$ – load transistors, and $Q5$ and $Q6$ – access transistors. The signal notation: BL – bit line, BLB – bit line bar, WL – word line, node A and node B – the internal nodes of an SRAM cell.

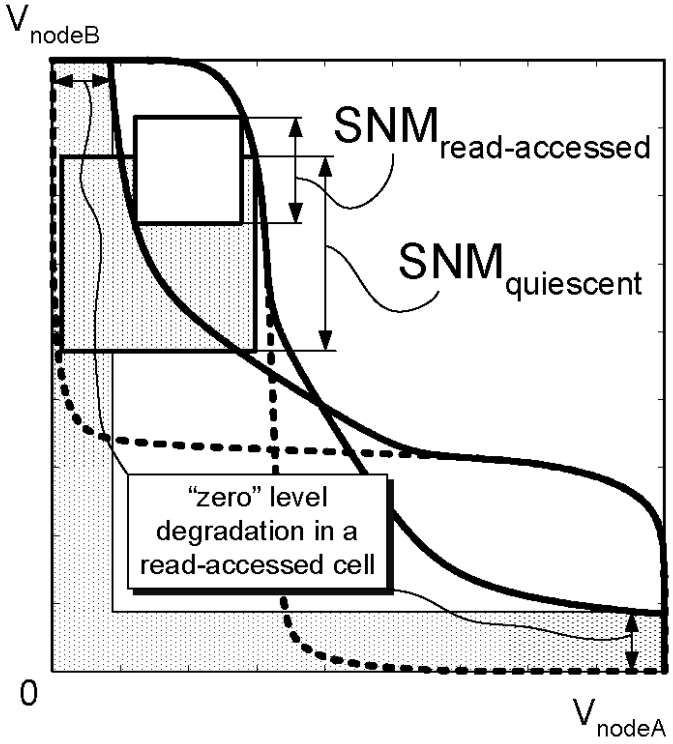


Figure 3.11 **Simulated VTCs of a 6T SRAM cell in retention (quiescent) and in the read-accessed modes (CMOS 0.18μm). Note that the read-accessed SNM is about a half of that in the retention mode.**

3.4.1 SRAM SNM and Process Variations

Process variations in modern CMOS DSM technologies pose an ever-growing threat to SRAM cell robustness. Threshold voltage (V_{TH}) spreads over ten percent of the typical are not unusual anymore [49],[4]. Such variations can dramatically reduce the SNM and thus, the stability of an SRAM cell, which is also demonstrated by our simulation results.

SRAM yield is correlated with the SNM spread. It is reported that, $\mu - 6\sigma$ of SNM is required to exceed $0.04 \cdot V_{DD}$ to reach a 90% yield on 1MB SRAM [50]. Poor transistor matching (increased $A_{\Delta V_{TH}}$) leads to a reduction in $\mu - 6\sigma$ and increases the number of unstable SRAM cells, impacting SRAM yield. Typically, this translates into a requirement

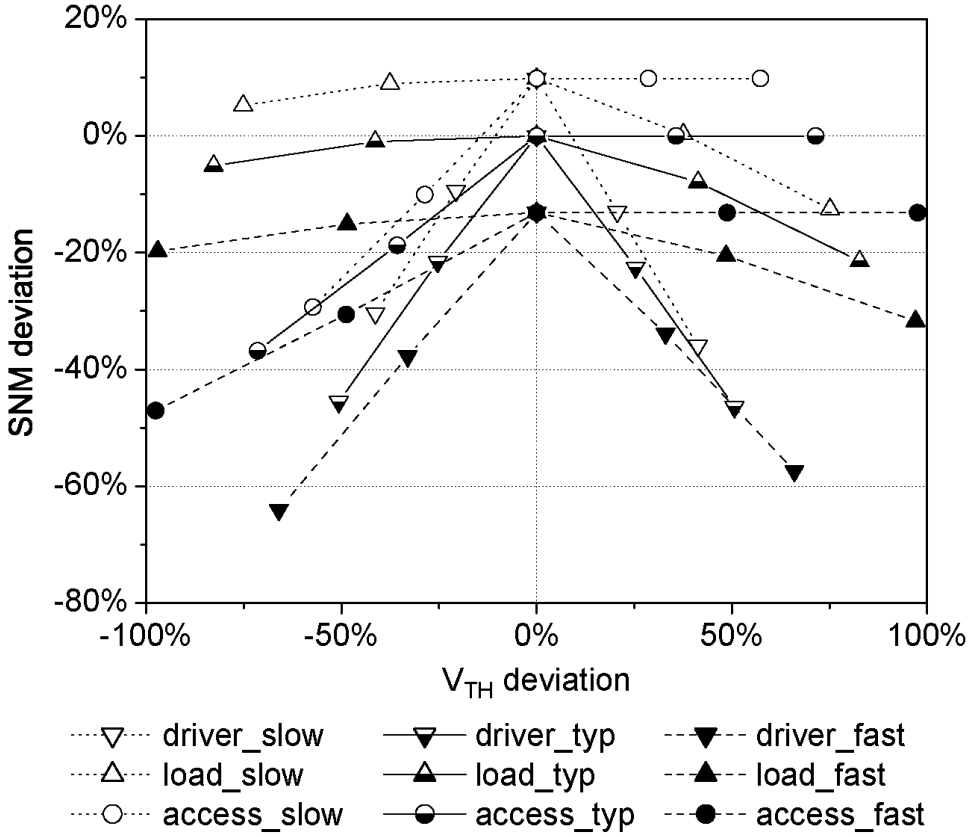


Figure 3.12 6T SRAM cell SNM deviation vs. threshold voltage deviation of one of the transistors.

that $\text{SNM}_{min} \geq 20\% \text{SNM}_{typ}$.

SNM dependence on V_{TH} variations for slow, typical and fast process corners is shown in Figure 3.12.

V_{TH} of only one transistor at a time was swept while keeping the V_{TH} of the other transistors typical. By sweeping V_{TH} of one of the transistors, we introduced a mismatch between the two halves of the SRAM cell. This essentially changes the shape of the transfer characteristics (see Figure 3.10(b)) and thus can adversely affect the SNM. The "0" point on the x-axis corresponds to the typical value of V_{TH} of a corresponding transistor in the

corresponding process corner.

V_{TH} variation of the driver transistor has the largest impact on the VTC shape and thus SNM due to its larger W/L ratio compared to other transistors in SRAM cell. Decreases in the V_{TH} of the access transistor also has a strong negative impact on the SNM. Since the measurements were taken in a read-accessed SRAM cell, the access transistors are effectively connected in parallel with the load transistors. Thus, reducing the V_{TH} of the access transistor compromises the low level stored in the cell, which in turn reduces the SNM. On the other hand, the V_{TH} variation of the PMOS load transistor has the least impact on the SNM due to its weaker drive and typically smaller W/L ratio.

Note that the SNM deviation is zero if the V_{TH} deviation of all transistors is zero (symmetrical cell), except only for the case of increasing the V_{TH} of the access transistor, which does not affect the SNM as its shunting action on the load transistor decreases.

If more than one V_{TH} is affected at a time, the SNM degradation can be stronger. Figure 3.13 presents several cases of the SNM vs. V_{TH} dependencies when V_{TH} of more than one transistor in the SRAM cell is not at its typical value (typical process corner). For instance, $Q2$_($Q1$=-25%) in Figure 3.13 represents the dependence of the SNM on the V_{TH} of $Q3$ provided that V_{TH} of $Q1$ is below its typical value by 25%. This dependence has its maximum at the point where V_{TH_Q1}=V_{TH_Q2}=-25% (i.e., where the cell is symmetrical). SNM vs. V_{TH} of $Q5$_($Q1$=-25%, $Q2$=$+25\%$, $Q3$=$+40\%$, and $Q4$=-40%) represents one of the worst cases of the SNM degradation due to the asymmetry of V_{TH} of the cell's transistors. SRAM cell SNM dependence on L_{EFF} or W_{EFF} variations in a single transistor under typical conditions is shown in Figure 3.14. The SNM decrease is insignificant when the transistor's effective length and width variation remains within 20% of the typical values. Regardless of the direction of the transistor geometry variation, the SNM is maximal at the typical (symmetrical) transistor sizes. This is due to the fact that

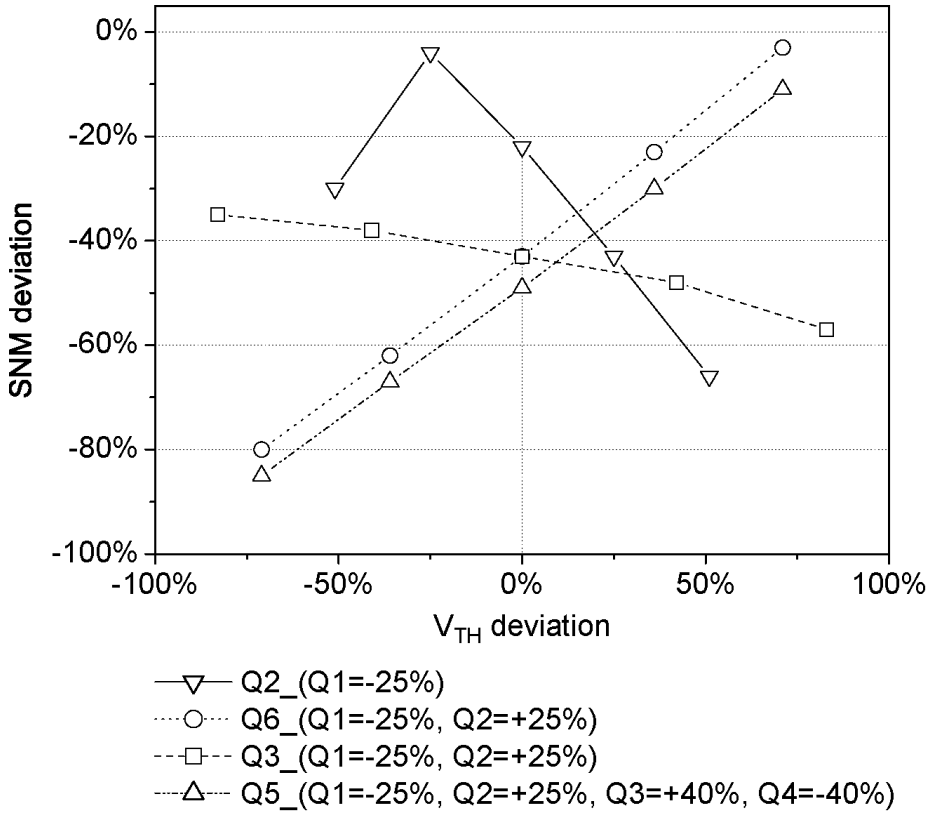


Figure 3.13 SRAM cell SNM vs. threshold voltage deviation of more than one transistor.

the variation of the transistor geometry in only one of the halves of the SRAM cell causes mismatch, which leads to the reduction of the SNM. Analysis of Figure 3.14 shows that for a weaker driver transistor (smaller W/L ratio) the SNM decreases, whereas a weaker access transistor improves the SNM. Deviations in W/L of the load transistor just slightly degrade the SNM.

From Figure 3.12, Figure 3.13 and Figure 3.14 it is apparent that the SNM of an SRAM cell is maximal if the effective driving strength of both halves ($Q1 - Q3 - Q5$ and $Q2 - Q4 - Q6$) of the cell is symmetrical with respect to their threshold voltages and W/L

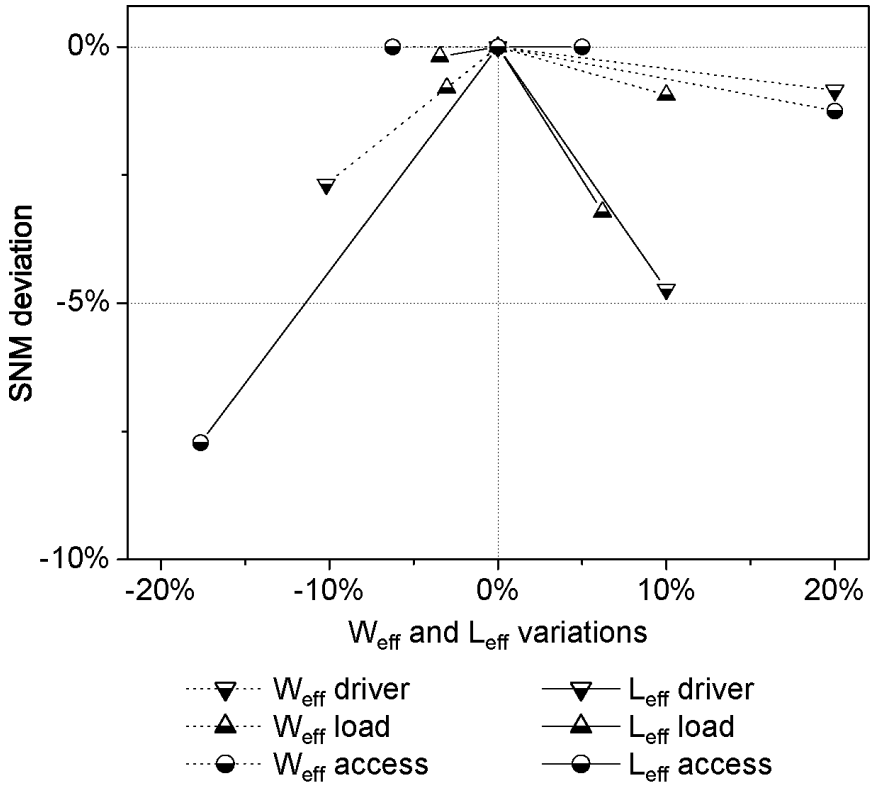


Figure 3.14 SRAM cell SNM deviation vs. L_{EFF} and W_{EFF}.

ratios.

As the technology is scaled into the decananometer region, V_{TH} spread becomes one of the main contributors to wider distribution of the SNM figures even in a defect-free chip. The granularity of the electric charge and the atomicity of matter begin to introduce substantial variation in number and position of MOSFET channel dopant atoms, t_{ox} becomes equivalent to several atomic layers with one to two atomic layers roughness [51]. For sub-100nm CMOS SRAMs, 6σ deviations of the SNM only due to intrinsic device fluctuations are projected to exceed the nominal SNM [49]. Randomness of channel dopant distribution will become a major source for the SNM reduction. To maintain reasonable

SNM and yield in future bulk CMOS SRAMs, cell ratios may have to be increased from the conventional $r = \frac{(W/L)_{driver}}{(W/L)_{access}} = 2$ towards the higher ratios [36], which counter-balances the scaling advantages of deep sub-micron technologies with respect to the area of the embedded SRAM cores.

3.4.2 SRAM SNM and Non-Catastrophic Defects

SNM vs. Non-Catastrophic Breaks and Bridges

Most catastrophic defects in SRAM cells cause drastic reduction of the SNM causing functional faults and, thus, tend to be easily detected by the regular march tests. However, SRAM cells with non-catastrophic defects can can have a non-zero SNM and escape standard tests, while degrading the cell stability and posing potential long-term reliability issues if not detected. In order to investigate SRAM SNM degradation based on non-catastrophic defects, we utilized the Carafe Inductive Fault Analysis (IFA) [23] tool to introduce *resistive* defects in the layout [52]. Carafe works by widening and shrinking the layout geometries and finds possible intersections of conductors in various process planes to determine how a spot defect of a certain size can affect the layout. We simulated defects with radii of 0.12, 0.24 and $0.36\mu m$. Based on the layout sensitivity analysis, Carafe generates a fault list. In this work we studied the SNM sensitivity of the cell layout similar to the one shown in Figure 4.3 on page 91.

We modelled the obtained faults as parallel and series resistors for bridges and breaks (opens) respectively and simulated the faulty netlists with the SPICE-like circuit simulator PSTAR. Defects were injected into the layout consisting of an array of 2×2 SRAM cells to generate the list of the most likely faults. Since an SRAM cell has a symmetrical structure, certain defects can appear in either half of the cell. The probability for such defects will

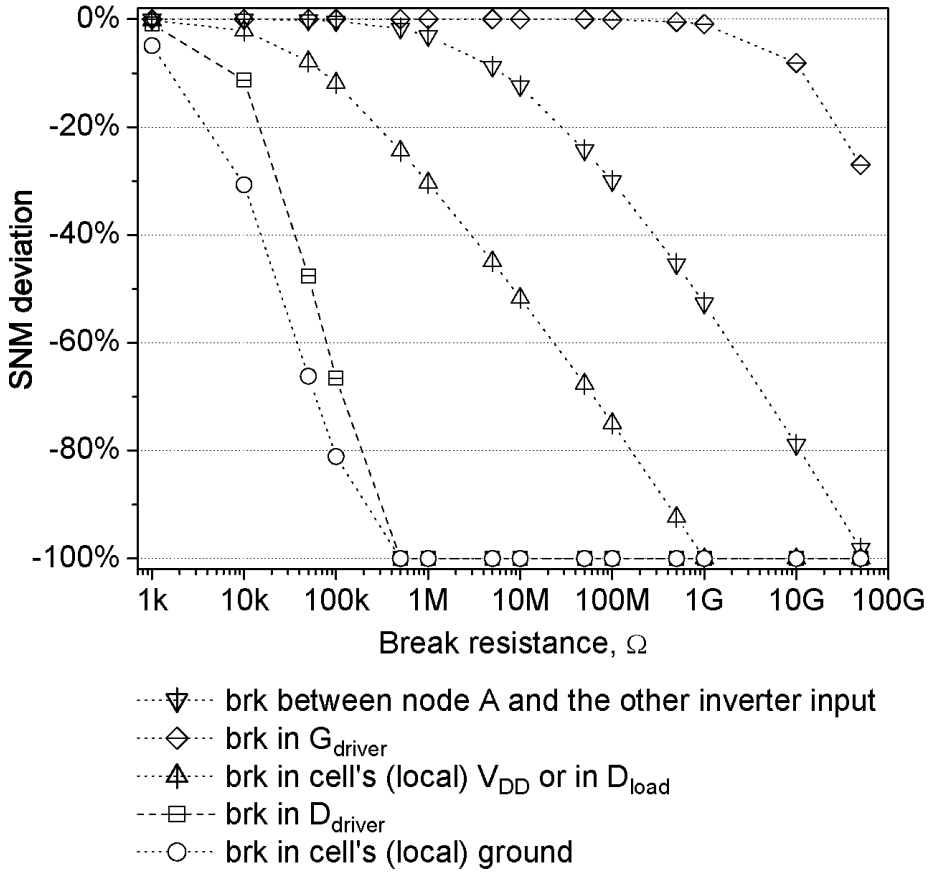


Figure 3.15 SRAM cell SNM deviation vs. break (resistive open) resistance.

be doubled. We simulated only one defect type at a time, while all other conditions were kept typical. The defect resistance values were swept from 1 $k\Omega$ to 50 $G\Omega$ for both the breaks and the bridges.

For the most likely breaks (resistive opens) and bridges (resistive shorts), the SNM deviation as a function of their resistance for typical conditions is presented in Figure 3.15 and Figure 3.16, respectively.

As is evident from Figure 3.15 and Figure 3.16, increasing the resistance of the bridges

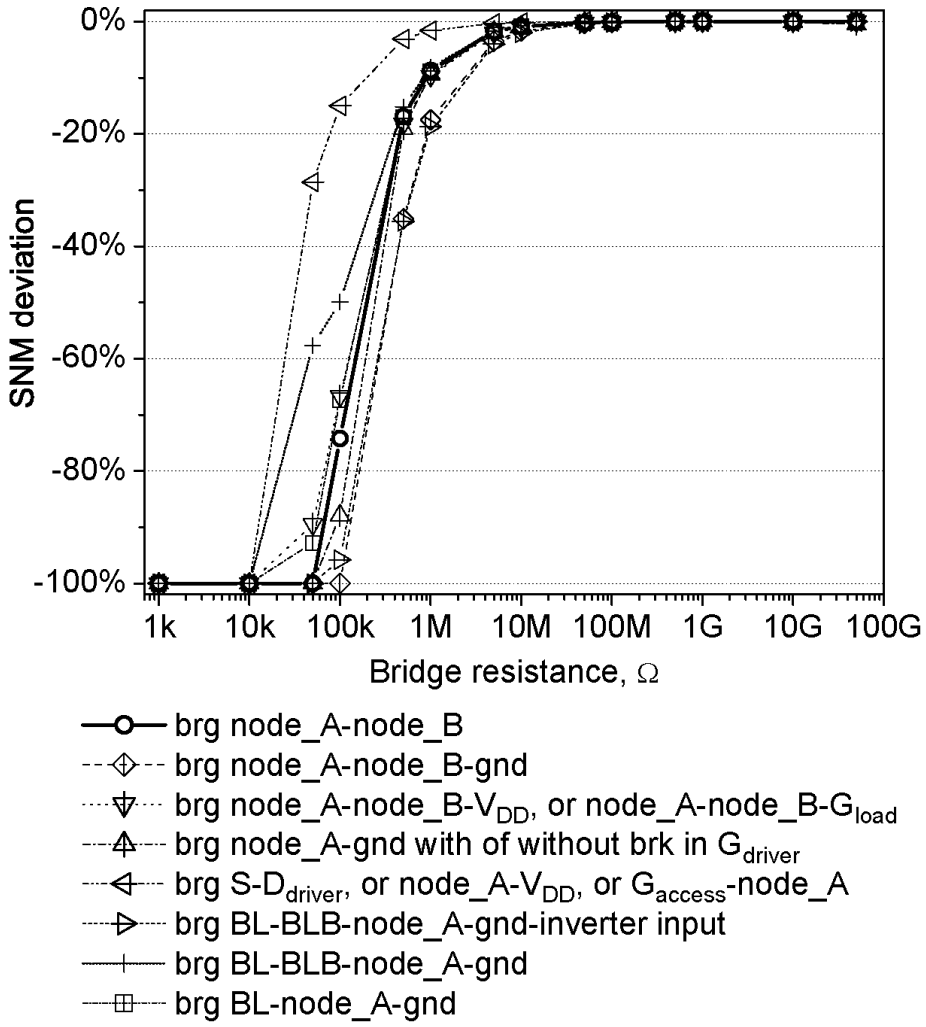


Figure 3.16 SRAM cell SNM deviation vs. bridge (resistive short) resistance.

and breaks has an opposite effect on an SRAM cell's SNM. When resistance is increasing, most resistive opens (breaks) linearly degrade the SNM and above certain resistance values, they cause the SNM to become a zero. Resistive open defects are likely to appear in place of poor or absent contacts, vias or silicide [7],[8]. Note that different breaks have different impact on the SNM. A break in the local ground contact of SRAM cell has the strongest

negative impact on the SNM. Break in drain of a driver transistor (D_{driver}) also causes a severe reduction of the SNM. A break in a cell's (local) V_{DD} or in the drain of a single load transistor (D_{load}), which can cause Data Retention Faults (DTFs), has a medium impact on the SNM. In contrast, resistive breaks in transistor gates do not cause a noticeable SNM degradation unless the break resistance exceeds approximately 1 $G\Omega$.

Figure 3.16 shows that unlike the case of resistive breaks, a reduction in the resistance of most bridges causes a very similar degree of SNM degradation. Bridges with resistance below $10 - 100$ $k\Omega$ reduce the SNM to be near zero and cause catastrophic functional failures, which are easily detected by the regular march tests. The SNM increases almost linearly for most of the bridges having resistances between 100 $k\Omega$ and 1 $M\Omega$. Bridge defects with a resistance of more than 10 $M\Omega$ show no impact on the SNM. Due to the cross-coupled layout of the SRAM cell, in the chosen layout the most likely resistive bridge is the bridge between the internal nodes of the cell. In Figure 3.16, the SNM dependence on this bridge (brg node_A-node_B) resistance is shown in a bold solid line.

3.4.3 Soft Errors and Defects in the Pull-up Path of a Cell

Single Event Upsets (SEUs) cause voltage spikes on the storage nodes of SRAM cells and last about 100ps [53]. The main sources of SEUs in SRAMs are [54]:

- α-particles from eutectic Pb-based solder bumps, packaging (mold compound, flip-chip underfill), semiconductor materials
- high-energy cosmic neutrons
- neutron-induced ^{10}B fission producing α-particles

If the voltage disturbance on a cell storage node is smaller than the noise margin of that node, the cell will continue to operate properly maintaining its data integrity. However,

if the *noise margin of a cell is not sufficient* to withstand the disturbance caused by a SEU, a "soft" error will result. While a SEU can cause a data error, the circuit itself is undamaged. Originally a concern for DRAMS, soft errors have recently become a growing issue in ultra high-density large embedded SRAMs operating at low voltages. Unlike in some combinational logic circuits which may mask SEUs, soft errors in SRAMs are always latched and thus lead to a data error. The only way to correct such data error is to rewrite the correct data into the cell.

The critical charge Q_{crit} required to upset a data node and flip a cell is given as [55],[56]:

$$Q_{crit} = \int_0^{\tau_{flip}} I_D dt = C_{node} V_{node} + I_{restore} \tau_{flip} \tag{3.10}$$

Where C_{node} and V_{node} (V_{node}=V_{DD} for a CMOS SRAM cell) are the capacitance and the voltage of the struck node respectively, $I_{restore}$ is the restore current provided by the pull-up path (PMOS transistors in case of a 6T SRAM cell), and τ_{flip} is the time required for the feedback mechanism to take over from the ion's current and flip the cell. The term $I_{restore}\tau_{flip}$ in Equation 3.10 represents the restoring force or active feedback of an SRAM cell. Note that the $I_{restore}$ of the node storing a "1" is weaker than the $I_{restore}$ of the node storing a "0" due to the typical sizing ratio of the driver and the pull-up transistors $\frac{W_{driver}}{L_{driver}} / \frac{W_{driver}}{L_{pull-up}} \approx 2$ as well as the mobility difference of electrons and holes $\mu_n/\mu_p \simeq 2-3$. Hence, the Q_{crit} of the node storing a "1" is smaller than the Q_{crit} of the node storing a "0". Thus, the node of an SRAM cell storing a "1" is more susceptible to soft errors.

Now consider a cell with a resistive defect in the pull-up path. Resistive defects in the pull-up path similar to the presented in Figure 3.15 reduce $I_{restore}$ of the cell and thus, its Q_{crit}. Therefore, a presence of a non-catastrophic resistive defect in the pull-up path reduces the SNM and at the same time increases the probability of a soft error while remaining undetected by the regular memory test methods. In the following chapters we will address the existing DFT detection techniques and propose several new detection

methods capable of detecting such weak SRAM cells.

3.4.4 SRAM SNM and Operating Voltages Variation

Variations in the operating voltages, such as the supply (V_{DD}), bit line (V_{BL}) or word line (V_{WL}) voltages, strongly impact an SRAM cell's SNM. The worst case SNM is typically observed for the fast process corner and high temperature; the best case SNM occurs for the slow process corner and low temperature. The results for all other temperature/process corner combinations fall in between the best and the worst cases. We swept V_{DD}, V_{BL} and V_{WL} one at a time from 0 to $1.5V$ and measured the corresponding SNM.

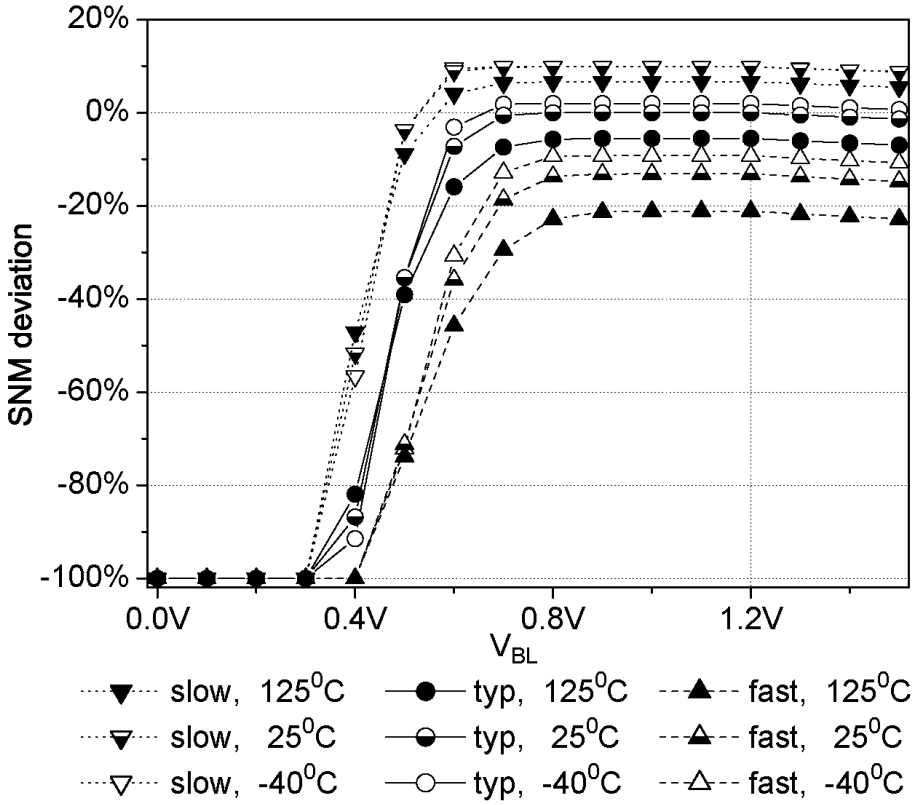


Figure 3.17 SRAM cell SNM deviation vs. bit line voltage

Figure 3.17 depicts the SNM dependence on V_{BL} while the V_{DD}, V_{WL}, and V_{BLB} are at the typical 1.2 V. The situation when one of the bit lines is driven from V_{DD} to the ground corresponds to a "write" operation. Overwriting the data stored in an SRAM cell becomes possible when the SNM is made zero. It can be seen from Figure 3.17 that the SNM becomes zero at $V_{BL} < 0.3\ V$ for the typical process corner. Note that the SNM does not decrease immediately once V_{BL} starts decreasing. The reduction of V_{BL} begins to reduce the SNM once $|V_{BL} - V_{WL}| > V_{TH_{access}}$ and the access transistor enters the linear mode. Since in the slow process corner transistors have higher V_{TH}, with the reduction of V_{BL} the cell SNM stays constant for a longer period than its counterparts from the typical and the fast process corners.

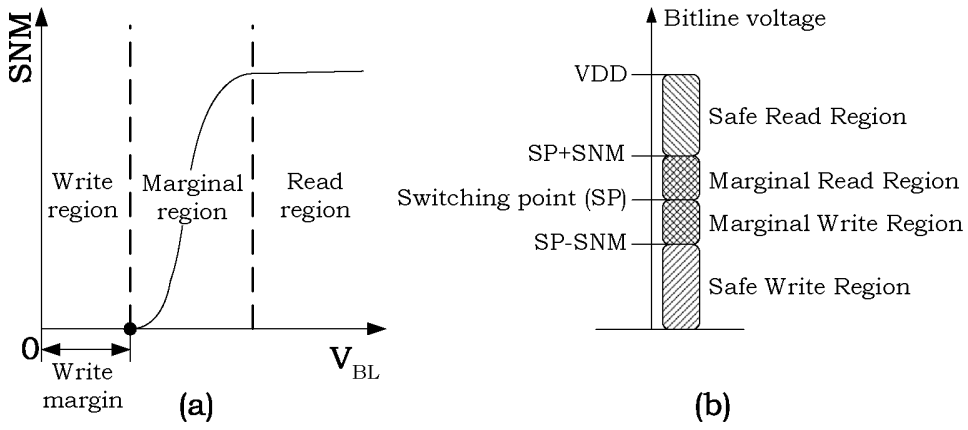


Figure 3.18 Read and write safe and marginal regions of an SRAM cell

Figure 3.18 (a) shows the read and write regions of an SRAM cell as a function of the bit line voltage. A write operation is possible in the region where the bit line voltage is at or below the voltage point where the SNM is zero. This voltage region is called the *write margin*.

The write margin is an important design parameter as it also defines the cell stability

to various disturbances. A balance between cell stability (SNM), cell area and access speed (read current) must be found, which may not maximize the cell stability.

Four regions can be identified for the bit line voltage between the ground and the precharge value: Safe Read Region, Marginal Read Region, Marginal Write Region and Safe Write Region (Figure 3.18 (b)) [57]. The Safe Read Region is defined from the switching point (SP) plus the SNM to the V_{DD}, whereas the Safe Write Region is defined from the SP minus the SNM to the ground. The regions between the SP and the Safe Read Region and between the SP and the Safe Write Region are described as the Marginal Read or Write regions, respectively.

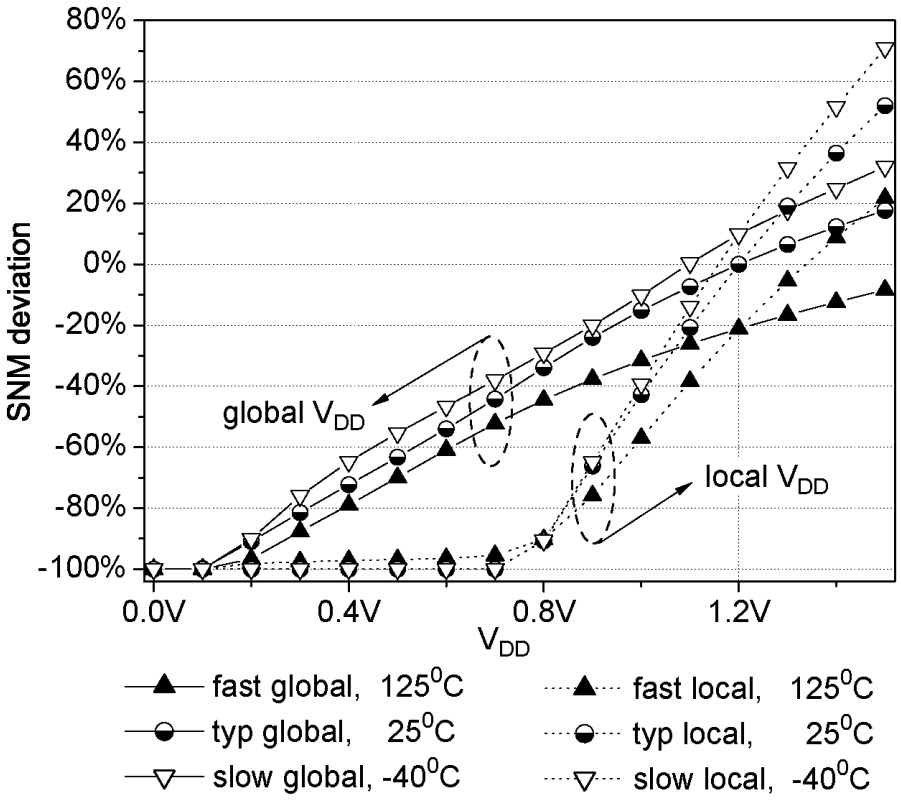


Figure 3.19 SRAM cell SNM deviation vs. V_{DD}

Figure 3.19 shows SNM deviation as a function of the global and local V_{DD}. By global V_{DD} variation, we mean the situation when V_{WL}, V_{BL}, V_{BLB}, and SRAM cell supply voltages vary all at the same time, which could model battery discharge in a mobile device. By local V_{DD}, we imply only the variation of SRAM cell supply voltage, while V_{WL}, V_{BL} and V_{BLB} are all at the typical values. Local V_{DD} variation mimics a faulty via in the supply voltage grid of SRAM cell. SNM shows strong dependence on both the local and the global V_{DD} variations. However, the SNM dependence on the local variation is stronger since in this case the word line and the global bit lines are at full V_{DD}, which causes the access transistor to shunt the pull-up transistors more strongly and degrade the low state of the cell. If we continue to raise the local V_{DD}, we can observe a significant increase of the SNM because the drive of the access transistors of the read-accessed cell is becoming weaker while the power supply of the cross-coupled inverters rises.

Figure 3.20 shows the dependence of the SNM on the word line voltage. All other operating voltages are at typical values. The SNM does not decrease if the V_{WL} is below the V_{TH} of the access transistor. Once $V_{WL}>V_{TH}$ of the access transistor, the SNM starts to deteriorate as the access transistor starts to shunt the load transistor and pull higher the node storing a logic zero. Note that if the word line exceeds V_{DD}, the SNM continues to deteriorate as the shunting action of the access transistor strengthens.

If all voltages are kept at their typical values and the temperature is varied from $-40°$C to 125°C, the SNM demonstrates rather weak temperature dependence from 5% for the slow process corner to 12% for the fast process corner, as is apparent from Figure 3.21. The SNM tends to decrease at the elevated temperatures. However, compared to other parameters, contributing to the SNM degradation, the temperature factor alone is likely to be negligible.

In the previous paragraphs, we described the impact of a single process parameter

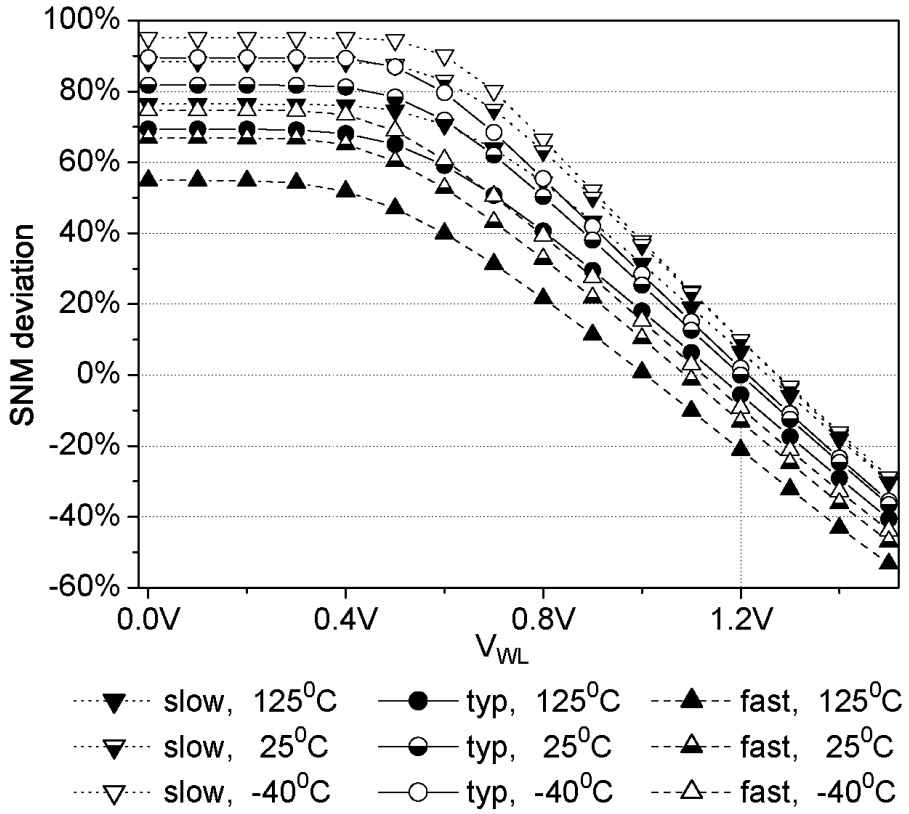


Figure 3.20 SRAM cell SNM deviation vs. word line voltage

variation on the SNM. However, in real life, several process parameters may change simultaneously. In case more than one process parameter (especially V_{TH}) departs from its typical value, the impact on the SNM is dramatic, often reducing the SNM to a very low value. Such SRAM cells are prone to stability faults, which may escape standard tests. Stability faults can thus potentially manifest themselves as long-term reliability problems.

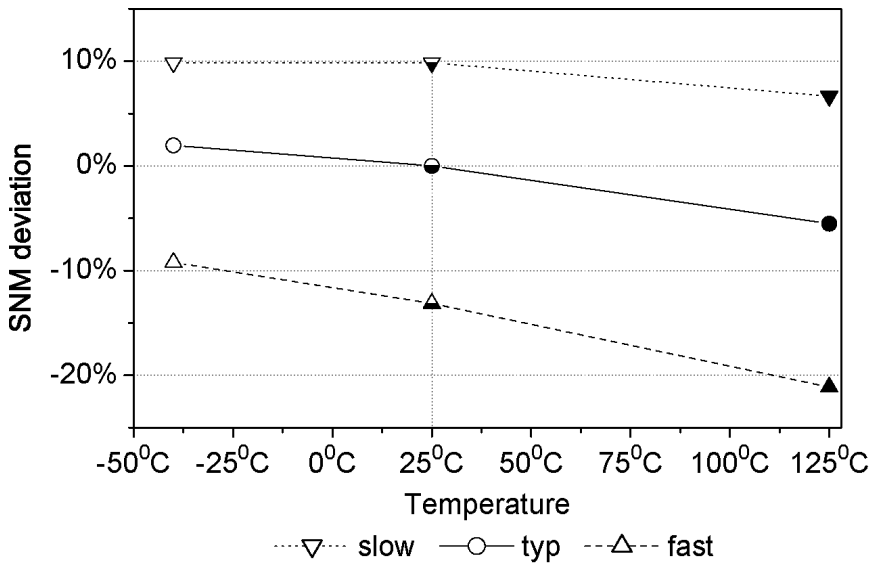


Figure 3.21 SRAM cell SNM deviation vs. temperature

3.5 Analytical SNM Expression for a Loadless Four-Transistor SRAM Cell Using the Alpha-Power Law Model

3.5.1 Alpha-Power Law Model

Shockley's MOSFET model, represented by Equation 3.11 has been used to analytically calculate circuit parameters for long-channel transistors. However, the Shockley model is increasingly inaccurate in describing the behavior of the modern short-channel transistors. Short-channel effects, such as the carrier velocity saturation, have to be taken into account

for accurate analytical characterization of sub-micron MOSFET transistors.

$$I_D = \begin{cases} 0, \text{ for } V_{GS} \leq V_{TH} \ - \text{ cutoff region} \\ K((V_{GS} - V_{TH})V_{DS} - 0.5V_{DS}^2), \text{ for } V_{DS} \leq V_{DSAT} \ - \text{ linear region} \\ 0.5K(V_{GS} - V_{TH})^2, \text{ for } V_{DS} \geq V_{DSAT} \ - \text{ saturation region} \end{cases} \tag{3.11}$$

where $V_{DSAT} = V_{GS} - V_{TH}$ is the drain saturation voltage, V_{TH} is the threshold voltage, K is the drivability factor (Equation 3.12) defined as follows:

$$K = \mu(\varepsilon_{ox}/t_{ox})(W/L_{EFF}) \tag{3.12}$$

where μ is the effective mobility, ε_{ox} and t_{ox} are the gate oxide dielectric constant and thickness respectively, W and L_{EFF} - transistor channel width and effective channel length respectively.

In scaled-down transistors, Shockley's square-law dependence does not hold. The shift of V_{DSAT} and discrepancies in the saturation region called for the Alpha-Power Law (APL) proposed in [58]. Drain current in the APL is then proportional to $(V_{GS} - V_{TH})^{\alpha}$, where α is the velocity saturation index. While in the Shockley model $\alpha = 2$, the measured α values can range from one to two. The APL model is defined as shown in Equation 3.13.

$$I_D = \begin{cases} 0, \text{ for } V_{GS} \leq V_{TH} \ - \text{ cutoff region} \\ (I_{D0}^{'}/V_{D0}^{'})V_{DS}, \text{ for } V_{DS} \leq V_{D0}^{'} \ - \text{ linear region} \\ I_{D0}^{'}, \text{ for } V_{DS} \geq V_{D0}^{'} \ - \text{ saturation region} \end{cases} \tag{3.13}$$

where

$$I_{D0}^{'} = I_{D0}\left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\alpha}, \tag{3.14}$$

$$V_{D0}^{'} = V_{D0}\left(\frac{V_{GS} - V_{TH}}{V_{DD} - V_{TH}}\right)^{\alpha/2}, \tag{3.15}$$

V_{D0} is the drain saturation voltage at $V_{GS} = V_{DD}$ and I_{D0} is the drain current at $V_{GS} = V_{DS} = V_{DD}$.

The MOSFET drain current equations can also be rewritten as in [59]:

$$I_D = \begin{cases} 0, \text{ for } V_{GS} \leq V_{TH} \ - \text{ cutoff region} \\ K_L(V_{GS} - V_{TH})^{\alpha/2} V_{DS}, \text{ for } V_{DS} < V'_{D0} \ - \text{ linear region} \\ K_S(V_{GS} - V_{TH})^{\alpha}, \text{ for } V_{DS} \geq V'_{D0} \ - \text{ saturation region} \end{cases} \tag{3.16}$$

where

$$K_L = \frac{I_{D0}}{V_{D0}(V_{DD} - V_{TH})^{\alpha/2}} \tag{3.17}$$

$$K_S = \frac{I_{D0}}{(V_{DD} - V_{TH})^{\alpha}} \tag{3.18}$$

We will use Equation 3.16 to derive an analytical expression for the SNM calculation in four-transistor loadless SRAM cells (Section 3.5.2).

3.5.2 Analytical SNM Expression Derivation

Assuming that both the inverters comprising a four-transistor loadless SRAM cell are equivalent, we used the following equivalent circuits to derive the SNM expression (Figure 3.22). Since we are interested in the worst-case SNM, we will consider the cell in the read-accessed mode, i.e. with the activated word line. In the case of a four-transistor loadless SRAM cell, which is using PMOS transistors as both the access and the load, the read-accessed mode corresponds to V_{WL}=0, i.e. when the gate of $Q4$ is grounded. Since many of the parameters in the APL model are technology-dependent, we will differentiate between n and p transistors as well as between the linear and saturated modes of the transistors. For instance, the velocity saturation index α, the threshold voltage V_{TH} and the saturation voltage V_{D0} will vary with the transistor type and operating mode. We will explain this in more detail in Section 3.5.2.

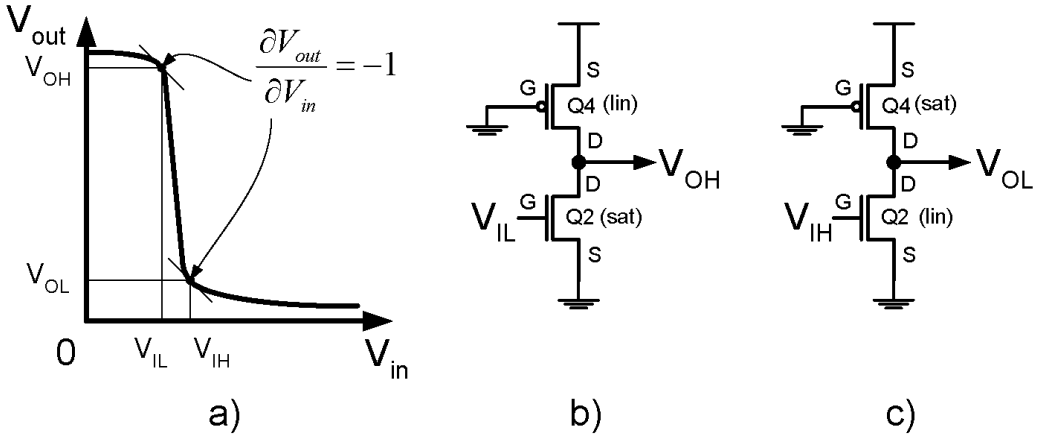


Figure 3.22 **Definitions of V_{OH}, V_{IL}, V_{IH} and V_{OL} (a); Equivalent circuit of a 4T loadless SRAM half cell showing the transistor modes for V_{IL} and V_{OH} (b) and for V_{IH} and V_{OL} (c).**

Finding V_{OH} and V_{IL}

Throughout the derivation we will use the following convention for $Q2$ and $Q4$ in Figure 3.22 (b) and (c). $Q2$ will be represented by a subscript n and $Q4$ - by a subscript p.

Analyzing Figure 3.22(a) we can conclude that the NMOS transistor $Q2$ in Figure 3.22(b) is in the saturation region (denoted as *sat*) and PMOS transistor $Q4$ is in the linear region (*lin*). Using Equation 3.16 and the KCL, we can equate the I_D equations for NMOS transistor in the saturation region and for the PMOS transistor in the linear mode. Using the APL model results in:

$$K_{s\_n(sat)}\left(V_{GS\_n}-V_{TH\_n(sat)}\right)^{\alpha\_n(sat)}=-K_{l\_p(lin)}\left(\left|V_{GS\_p}-V_{TH\_p(lin)}\right|\right)^{\alpha\_p(lin)/2}V_{DS\_p} \quad (3.19)$$

Note that $V_{GS\_P}=-V_{DD}$, $V_{GS\_N}=V_{IL}$, $V_{DS\_N}=V_{OH}$ and $V_{DS\_P}=V_{OH}-V_{DD}$. Due to

$V_{D0} < 0$ for PMOS transistor $Q4$, $K_p < 0$. Then, Equation 3.19 can then be rewritten as:

$$K_{s\_n(sat)} \left(V_{IL} - V_{TH\_n}\right)^{\alpha\_n(sat)} = -K_{l\_p(lin)} \left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\alpha\_p(lin)/2} \left(V_{OH} - V_{DD}\right) \tag{3.20}$$

V_{OH} can be expressed as:

$$V_{OH} = -\frac{K_{s\_n(sat)} \left(V_{IL} - V_{TH\_n(sat)}\right)^{\alpha\_n(sat)}}{K_{l\_p(lin)} \left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\alpha\_p(lin)/2}} + V_{DD} \tag{3.21}$$

We defined V_{OH} as the point where $\delta V_{OH}/\delta V_{IL} = -1$:

$$\frac{\partial V_{OH}}{\partial V_{IL}} = \left[-\frac{K_{s\_n(sat)}}{K_{l\_p(lin)}} \left(V_{IL} - V_{TH\_p(lin)}\right)^{\alpha\_n(sat)} K_{l\_p(lin)} \left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{-\alpha\_p(lin)/2}\right]' = -1 \tag{3.22}$$

Solving Equation 3.22 for V_{IL}, we obtain:

$$V_{IL} = \left(\frac{K_{l\_p(lin)}}{K_{s\_n(sat)}} \frac{1}{\alpha_{n(sat)}}\right)^{\frac{1}{\alpha_{n(sat)}-1}} \left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\frac{\alpha_{p(lin)}}{2\left(\alpha_{n(sat)}-1\right)}} + V_{TH\_n(sat)} \tag{3.23}$$

Finding $\mathbf{V}_{OL}$ and $\mathbf{V}_{IH}$

Now, let us consider the point where $V_{in} = V_{IH}$ and $V_{out} = V_{OL}$ on the VTC of the four-transistor loadless SRAM half-cell. Analyzing Figure 3.22 (a) we can conclude that the NMOS transistor $Q2$ in Figure 3.22 (c) is in the linear region and the PMOS transistor $Q4$ is in the saturation region. Using Equation 3.16 and the KCL, we can equate the I_D equations for NMOS transistor in the linear mode and for the PMOS transistor in the saturation mode. Using the APLM results in:

$$K_{l\_n(lin)} \left(V_{GS\_n} - V_{TH\_n(lin)}\right)^{\alpha\_n(lin)/2} V_{DS\_n} = K_{s\_p(sat)} \left(V_{GS\_p} - V_{TH\_p(sat)}\right)^{\alpha\_p(sat)} \tag{3.24}$$

Note that $V_{GS\_P} = -V_{DD}$, $V_{GS\_N} = V_{IH}$, $V_{DS\_N} = V_{OL}$. Equation 3.24 can be rewritten as:

$$K_{l\_n(lin)} \left(V_{IH} - V_{TH\_n(lin)}\right)^{\alpha\_n(lin)/2} V_{OL} = K_{s\_p(sat)} \left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\alpha\_p(sat)} \tag{3.25}$$

V_{OL} can be expressed as:

$$V_{OL} = \frac{K_{s\_p(sat)} \left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\alpha\_p(sat)}}{K_{l\_n(lin)} \left(V_{IH} - V_{TH\_n(lin)}\right)^{\alpha\_n(lin)/2}} \tag{3.26}$$

We defined V_{OL} as the point where $\delta V_{OL}/\delta V_{IH} = -1$:

$$\frac{\partial V_{OL}}{\partial V_{IH}} = \left[\frac{K_{s\_p(sat)}}{K_{l\_n(lin)}} \left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\alpha\_p(sat)} \left(V_{IH} - V_{TH\_n(lin)}\right)^{-\alpha\_n(lin)/2}\right]' = -1 \tag{3.27}$$

Solving Equation 3.22 for V_{IH}, it can be expressed as:

$$V_{IH} = \left(\frac{K_{l\_n(lin)}}{K_{s\_p(sat)}} \frac{2}{\alpha_{n(sat)}}\right)^{-\frac{2}{\alpha_{n(sat)}+2}} \left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\frac{2\alpha_{p(sat)}}{\alpha_{n(lin)}+2}} + V_{TH\_n(lin)} \tag{3.28}$$

SNM Expression for 4T Loadless SRAM Cell

Following Equations 3.1 and 3.2 and substituting V_{OH}, V_{IL}, V_{OL} and V_{IH} from Equations 3.21, 3.21, 3.26 and 3.28 respectively, we can find the Noise Margin High (NM_H):

$$\begin{aligned} NM_H = V_{OH} - V_{IH} = V_{DD} - \frac{K_{s\_n(sat)}\left(V_{IL} - V_{TH\_n(sat)}\right)^{\alpha\_n(sat)}}{K_{l\_p(lin)}\left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\alpha\_p(lin)/2}} - \\ - \left(\frac{K_{l\_n(lin)}}{K_{s\_p(sat)}} \frac{2}{\alpha_{n(sat)}}\right)^{-\frac{2}{\alpha_{n(sat)}+2}} \left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\frac{2\alpha_{p(sat)}}{\alpha_{n(lin)}+2}} + V_{TH\_n(lin)}, \end{aligned} \tag{3.29}$$

and the Noise Margin Low (NM_L):

$$\begin{aligned} NM_L = V_{IL} - V_{OL} = \left(\frac{K_{l\_p(lin)}}{K_{s\_n(sat)}} \frac{1}{\alpha_{n(sat)}}\right)^{\frac{1}{\alpha_{n(sat)}-1}} \left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\frac{\alpha_{p(lin)}}{2\left(\alpha_{n(sat)}-1\right)}} + \\ + V_{TH\_n(sat)} - \frac{K_{s\_p(sat)}\left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\alpha\_p(sat)}}{K_{l\_n(lin)}\left(V_{IH} - V_{TH\_n(lin)}\right)^{\alpha\_n(lin)/2}}. \end{aligned} \tag{3.30}$$

NM_H and NM_L represent the sides of a rectangle embedded between the two VTCs of the half-cells (Figure 3.23). For analytical calculations, we can express the SNM of a loadless

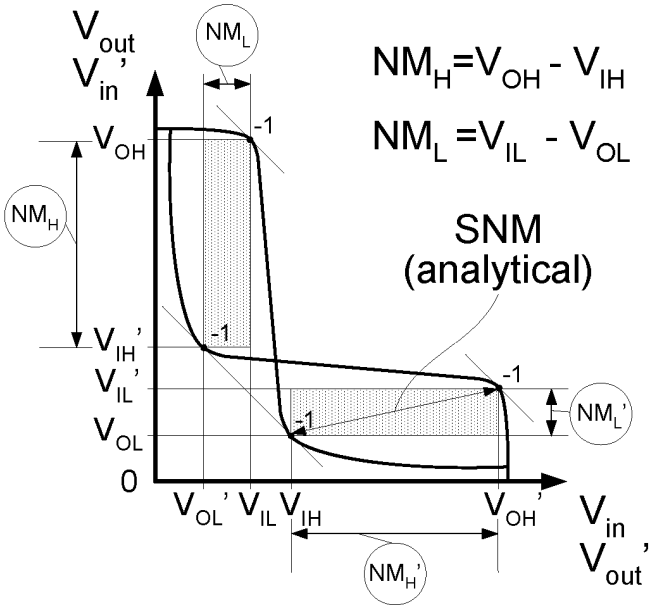


Figure 3.23 SNM definition utilized in the analytical SNM expression for a four-transistor loadless SRAM cell (Equation 3.31).

four-transistor SRAM cell as the diagonal of the rectangle with the sides equal to NM_H

and NM_L. The final expression can be presented as:

$$SNM_{4T\_loadless\_SRAM} = \sqrt{NM_H^2 + NM_{LH}^2} =$$

$$= \sqrt{\begin{array}{l} \left(\begin{array}{l} V_{OH} - V_{IH} = V_{DD} - \frac{K_{s\_n(sat)}\left(V_{IL} - V_{TH\_n(sat)}\right)^{\alpha\_n(sat)}}{K_{l\_p(lin)}\left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\alpha\_p(lin)/2}} - \\ \\ - \left(\frac{K_{l\_n(lin)}}{K_{s\_p(sat)}} \frac{2}{\alpha_{n(sat)}}\right)^{-\frac{2}{\alpha_{n(sat)}+2}} \left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\frac{2\alpha_{p(sat)}}{\alpha_{n(lin)}+2}} + V_{TH\_n(lin)} \end{array} \right)^2 + \\ \\ + \left(\begin{array}{l} \left(\frac{K_{l\_p(lin)}}{K_{s\_n(sat)}} \frac{1}{\alpha_{n(sat)}}\right)^{\frac{1}{\alpha_{n(sat)}-1}} \left(\left|-V_{DD} - V_{TH\_p(lin)}\right|\right)^{\frac{\alpha_{p(lin)}}{2\left(\alpha_{n(sat)}-1\right)}} + \\ \\ + V_{TH\_n(sat)} - \frac{K_{s\_p(sat)}\left(\left|-V_{DD} - V_{TH\_p(sat)}\right|\right)^{\alpha\_p(sat)}}{K_{l\_n(lin)}\left(V_{IH} - V_{TH\_n(lin)}\right)^{\alpha\_n(lin)/2}}. \end{array} \right)^2 \end{array}} \tag{3.31}$$

Simulation Results vs. the Analytical Expression

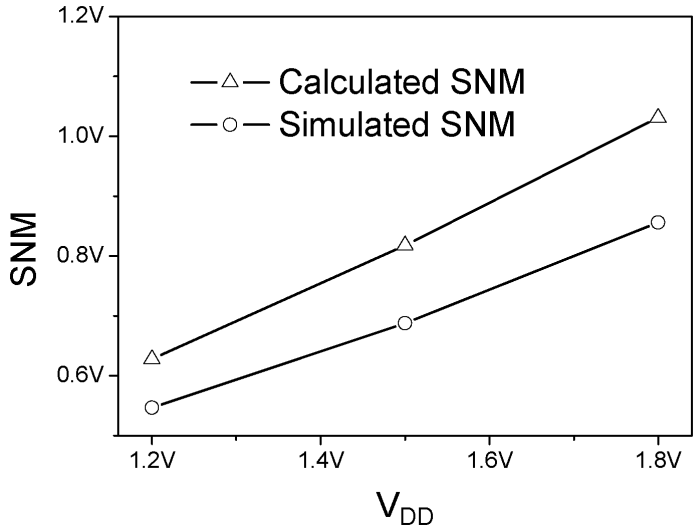


Figure 3.24 **SNM of the 4T loadless SRAM cell vs V_{DD} ($CMOS$ $0.18\mu m$ technology, $(W/L)_{driver} = 1\mu m/0.18\mu m$, $(W/L)_{access} = 0.5\mu m/0.18\mu m$). Comparison of the results using SPICE simulation and calculation using Equation 3.31.**

Table 3.1 represents the extracted parameters of the α-power law model from SPICE simulations of 0.18-μm MOSFETs used in our four-transistor loadless SRAM cell. The values of α for the NMOS and PMOS transistors were obtained from fitting the calculated V_{GS} vs. I_D ($V_{DS} = V_{DD} = 1.8V$) α-power model curves to the SPICE simulated ones.

The SNM definition depicted in Figure 3.23 is convenient for the analytical SNM calculation. The SNM, defined as the diagonal or as a side of the maximum square embedded between the two inverter voltage transfer characteristics of an SRAM cell, while can be easily applied for graphical extraction of SNM, it cannot be derived mathematically.

Note that the proposed analytical expression (Equation 3.31), unlike the previous art [34], allows one to calculate the SNM of the loadless 4T SRAM cell accounting for

Table 3.1 α**-power law MOSFET model parameters for 0.18**μ**m technology and** $V_{DD} = 1.8V$**.**

Model parameters	NMOS, $\frac{W}{L} = \frac{1\mu m}{0.18\mu m}$	PMOS, $\frac{W}{L} = \frac{0.5\mu m}{0.18\mu m}$
$\alpha\_(lin)$	1.2	1.4
$\alpha\_(sat)$	1.15	1.4
$V_{D0},\ V$	0.98	0.98
$V_{TH(lin)},\ V$ $(V_{DS} = 0.1V)$	0.49	0.522
$V_{TH(sat)},\ V$ $(V_{DS} = 1.8V)$	0.489	0.488
$I_{D0},\ mA$	0.646	0.123

the transistor parameter differences shown in Table 3.1.

Figure 3.24 presents the SNM dependence of the 4T loadless SRAM cell using "-1" slope SNM definition on V_{DD} calculated using the proposed analytical model and simulated with HSPICE. For the simulated cell with $(W/L)_{driver} = 1\mu m/0.18\mu m$, $(W/L)_{access} = 0.5\mu m/0.18\mu m$) in CMOS $0.18\mu m$ technology, the average error of the proposed analytical model with respect to the simulated data is $\simeq$14-15% across the V_{DD} values from 1.2V to 1.8V.

3.6 Summary

In this chapter we presented the concept of Static Noise Margin and the various definitions found in the literature. SNM can be used as a figure of merit when estimating the stability of an SRAM cell. Most SRAM manufacturers define the SNM as the side of the

smaller of the two largest squares drawn between the voltage transfer characteristics of the equivalent inverters comprising an SRAM cell. This SNM definition is more suitable for use in computer simulations. Other definitions may be more suitable for analytical SNM calculation.

The comprehensive SRAM cell stability sensitivity investigation presented in Section 3.4 demonstrated the various factors affecting SRAM cell stability. The growing process uncertainties give rise to inter- and intra-cell mismatches causing stability deterioration. The increasing SRAM packing density presents large critical area for the possible bridging defects, while the resistive contacts and vias can introduce data retention faults. Combined with ground bounce and coupling effects, these factors can severely compromise cell's stability and cause inadvertent bit flips. Soft errors become a growing concern in SRAM data integrity. Special measures, such as error correction or repair of unstable cells with the redundant cells, are needed.

We used the SNM definition presented in Section 3.3.3 for derivation of an analytical expression for a four-transistor loadless SRAM cell using the α-power law model presented in Section 3.5.2. The SNM values obtained using the proposed analytical expression demonstrate a good match with the simulation results.

Chapter 4

SRAM Cell Stability Detection

This chapter is dedicated to stability fault modelling, detection concepts and the existing techniques for SRAM stability testing. Section 4.1 presents Data Retention Faults (DRFs) and the regular Data Retention Test (DRT) and introduces a link between the DRF and Stability Faults (SF) in terms of the SNM. In Section 4.2 we propose an SRAM cell stability detection concept that explains the conditions required to be met in order to detect an SRAM cell with compromised stability (reduced SNM). The existing industrial strategies for SF detection in SRAM cells are categorized and presented in Section 4.3. Finally, contributions presented in this Chapter are summarized in Section 4.4.

4.1 Stability Fault Modelling

4.1.1 Data Retention Faults and Data Retention Test

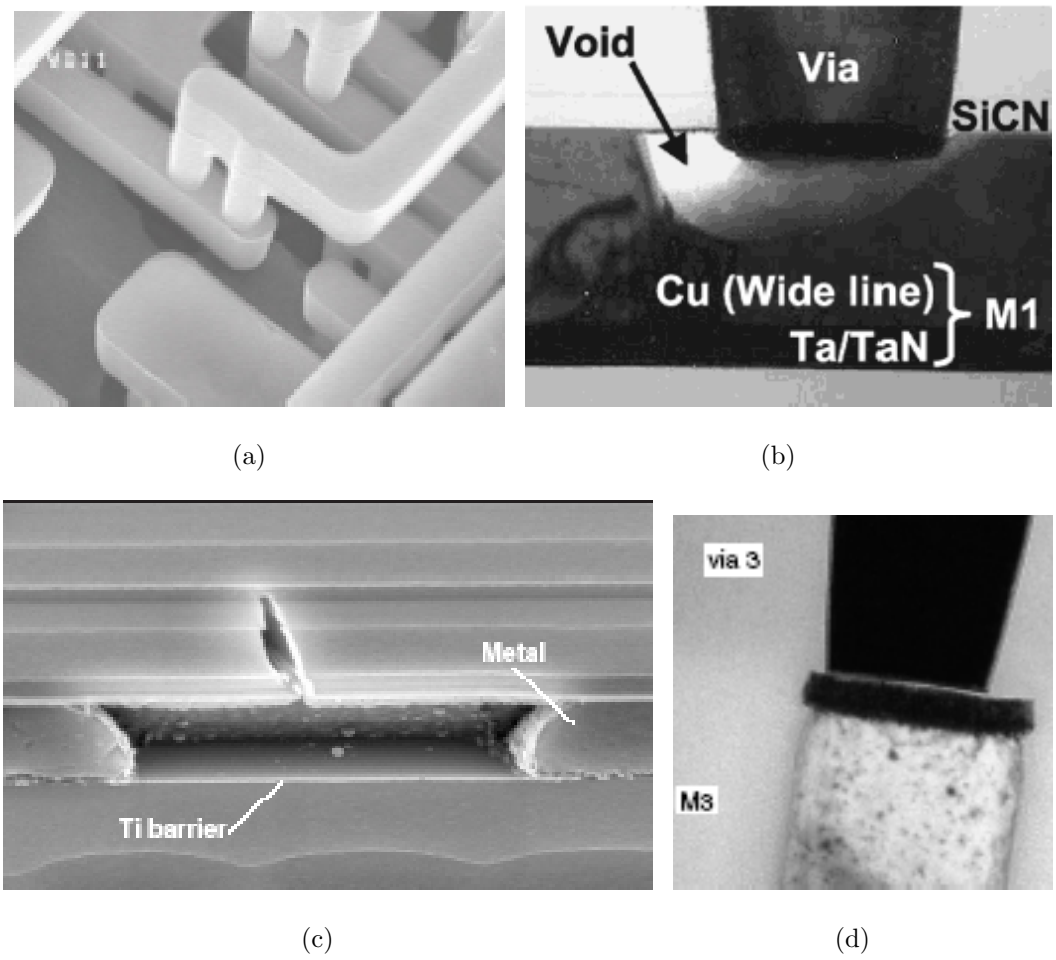
Defects causing DRFs

The operation speed of the scaled-down technologies tends to be limited by the propagation delay of their interconnects [4]. This limitation can be overcome by using Cu interconnect technology in combination with low-k dielectrics and a low-k barrier dielectric cap layer. Most of the semiconductor foundries have switched from the traditional aluminum metallization with tungsten via plugs to copper interconnects. The lower electrical resistance of copper leads to improved power distribution and device performance throughout the chip. Copper also improves the electromigration resistance, a major concern in an IC's long term reliability, by as much as 50 times.

The resistance of copper interconnects is less than two-thirds that of aluminum-tungsten interconnects. This results in RC delay reductions of 15% or more. Via series resistance runs as low as 20% of that of tungsten plugs. These benefits become even more pronounced in $0.13\mu m$ technologies and smaller [4]. The dual-damascene process is necessary to eliminate the need for copper etch and for dielectric gap fill, which becomes very challenging as dimensions continue to shrink. In the dual damascene process, the trench and via patterns are defined by etching through the dielectric material in two separate lithography and etch steps. Metal filling of the trenches and vias is accomplished in a single copper electroplating step, which is followed by the CMP to remove excess copper and obtain the desired metal pattern.

However, electromigration and stressmigration effects limit the further scaling of copper interconnects. Increasing frequency of unreliable contacts and vias have been reported due to stress-induced voiding under the vias in CMOS $0.13\mu m$ technology node and beyond [60]

An example of a failed interconnect via is presented in Figure 4.1(b) [5].



(a) (b)

(c) (d)

Figure 4.1 (a) Dual damascene copper interconnects [4]; (b) Cross-sectional TEM image of a failed copper interconnect via [5]; (c) Weak open defects: detailed cross-section of a metal open line, showing the metal cavity and formation of a weak open defect due to the Ti barrier; (d) A resistive via [6].

A typical CMOS process involves numerous contacts from metal layers to diffusion areas and vias between metal layers. Providing reliable contacts and vias is a growing

challenge [6, 5], especially in high-density SRAM arrays. Poorly formed overly resistive contacts and vias can cause delay faults if located in the timing or signal propagation paths. However, our research showed that if resistive contacts are located in the load PMOS transistors of an SRAM cell, such defects can cause a DRF or a stability fault depending on their severity and can escape traditional tests.

Figure 4.1 on the preceding page shows dual damascene copper interconnects and illustrates weak open defects in metal and via formation. Difficult-to-detect weak opens with $R_{open} < 10M\Omega$ constitute a significant part of the total number of opens, as suggested by Figure 4.2. Note that weak opens are almost equally distributed across the entire range from $10k\Omega$ to $10M\Omega$ showing relatively high and flat probability of a weak open with any resistance value.

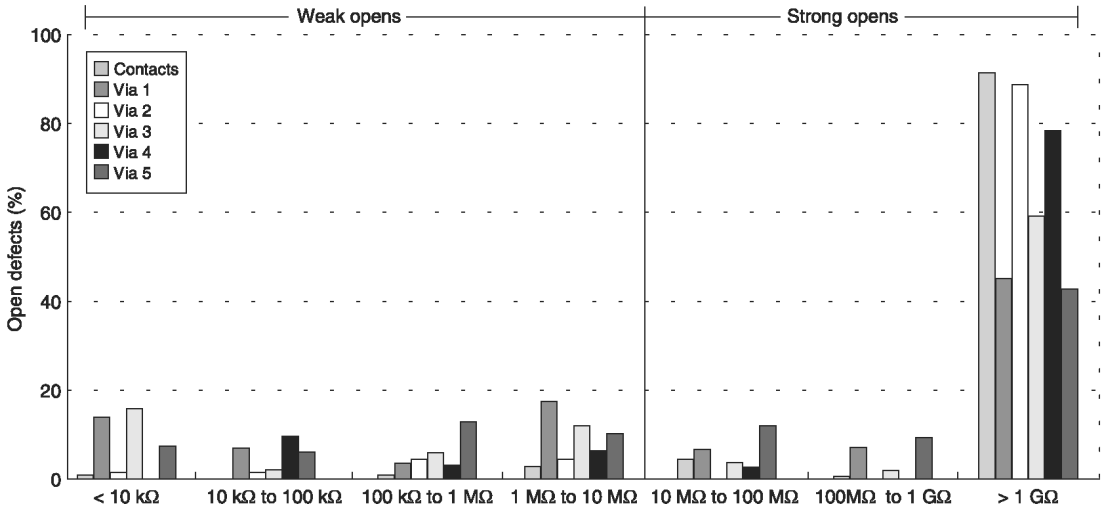


Figure 4.2 Resistance distribution for contact and via opens [6].

SRAM arrays are the densest form of circuitry and can occupy a significant percentage of silicon area. Each cell contains from ten contacts, as in the cell used in this work (Figure 4.3), to fourteen in the latest technologies [61]. These contacts are potential locations

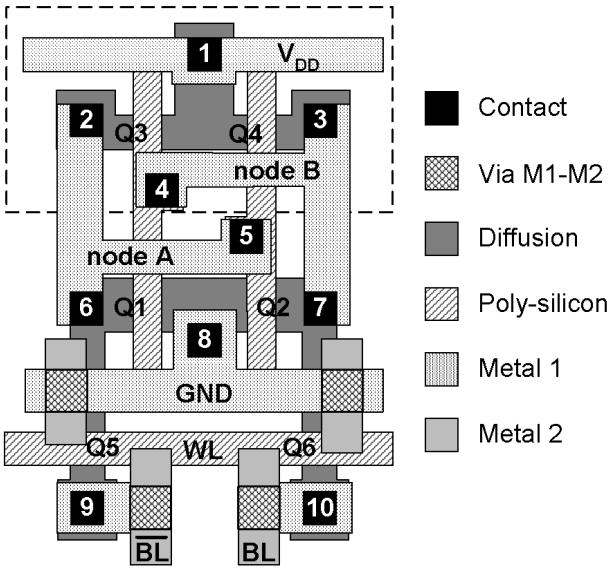


Figure 4.3 **Layout of a 6T SRAM cell, where contact resistances 1, 2 and 3 correspond to resistors R1, R2 and R3 respectively, which are shown in Figure 4.4.**

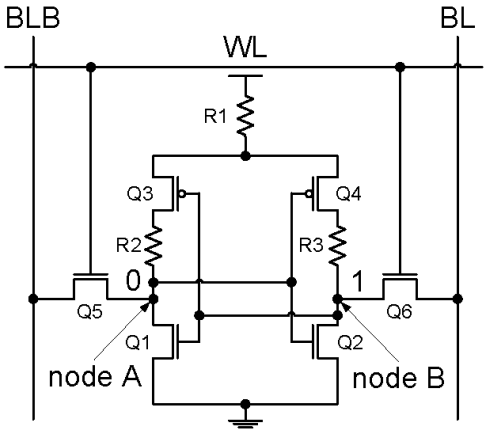


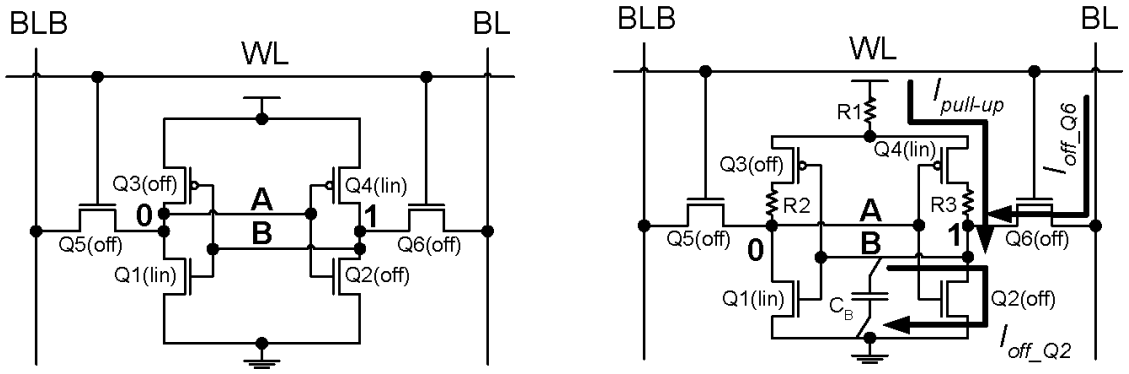
Figure 4.4 **SRAM cell schematic with resistors in place of potential weak opens that can cause stability faults as per layout in Figure 4.3.**

of weak opens. A break or a weak open in contact "1" in Figure 4.3 represents a symmetric defect when both drains of load PMOS transistors $Q3$ and $Q4$ are connected to the power

supply through a shared resistor $R1$ (Figure 4.4). An infinite value of $R1$ corresponds to an open in the cell's supply or to the situation when both PMOS transistors are missing. Opens in contacts "2" and "3" represent asymmetric defects in the left-hand and right-hand sides of the cell respectively and correspond to resistive connection of $Q3$ and $Q4$ sources to nodes A and B respectively (Figure 4.4). Infinite resistance in contacts "2" or "3" corresponds to an asymmetric defect [7]. As technology moves towards smaller feature sizes, the "split word line" cell layouts have been adopted by the foundries [62, 61]. They have separate drain contacts for each of the load PMOS devices increasing to four the total number of possible open contacts that can cause data retention or stability faults.

Data Retention Fault and Data Retention Test

Figure 4.5(a) shows a 6T SRAM cell in retention (quiescent) mode when $V_{BL} = V_{BLB} = V_{DD}$, $V_{WL} = 0$, where $Q1$ and $Q2$ are driver transistors, $Q3$ and $Q4$ are load transistors, and $Q5$ and $Q6$ are access transistors, respectively.



(a) (b)

Figure 4.5 (a) Defect-free SRAM cell 6T SRAM cell in retention (quiescent) mode when $V_{BL} = V_{BLB} = V_{DD}$, $V_{WL} = 0$; (b) SRAM cell with a symmetric (R1) and asymmetric (R2, R3) defects in data retention mode.

Poorly formed contacts in the cell shown as $R1$, $R2$ and $R3$ in Figure 4.5(b) can cause SFs in an SRAM cell. A break or a weak open represented by $R1$ creates a symmetric defect when both drains of load PMOS transistors $Q3$ and $Q4$ have a resistive connection to the power supply. Infinite value of $R1$ corresponds to an open in the cell's supply or to the situation when both PMOS transistors are missing. Opens represented by $R2$ and $R3$ create an asymmetric defect in the left-hand and right-hand sides of the cell, respectively, and correspond to the resistive connection in the pull-up path with either transistor $Q3$ or $Q4$.

A *Data Retention Fault* (DRF) is defined as the failure of an SRAM cell to retain the written data for as long as the power is supplied. Figure 4.5(b) helps to explain the conditions causing a DRF. If the off-state current I_{off_Q2} of transistor Q2 in Figure 4.5(b) is such that

$$I_{off\_Q2} > I_{pull-up} + I_{off\_Q6}, \tag{4.1}$$

where $I_{pull-up}$ is the current in the pull-up path of a cell and I_{off_Q6} is the off-state current of Q6, then after a delay proportional to

$$C_B V_B / (I_{off\_Q2} - (I_{pull-up} + I_{off\_Q6})) \tag{4.2}$$

the capacitance of node B (C_B) will discharge sufficiently for the cell to flip states. Reading the cell data after a delay on the order of 100ms and comparing it with the previously written data can detect resistive defects R1-R3 in the range of several GΩ. This algorithm is employed by the traditional passive Data Retention Test (DRT, a.k.a. Delay or Pause Test).

Conditions and the defect resistance range of a DRF detection by means of the Pause (a.k.a. DRT) test are illustrated in Figure 4.6. The DRT test reads the SRAM array after a pause on the order of $100ms$ to determine whether any cell has changed from its

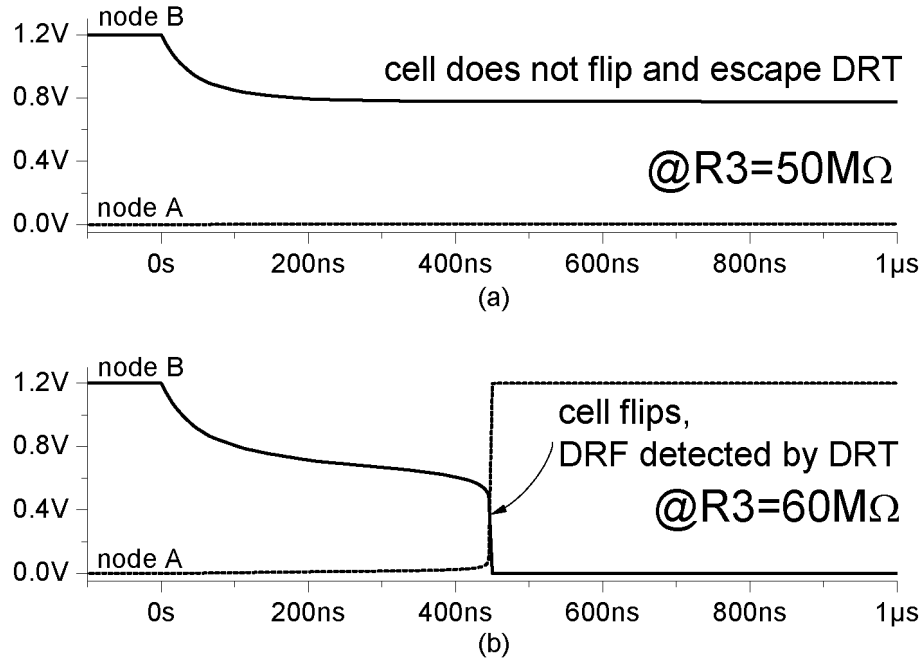


Figure 4.6 **Data Retention Fault due to the discharge of node B through the off-state current of Q2 (simulation results for CMOS 0.13μm technology, V_{DD}=1.2V, T=150^{o}C). A resistive open R3=50MΩ is insufficient to flip the cell and thus is not detected by Data Retention Test (DRT) (a), whereas R3=60MΩ causes a DRF and is detected by DRT (b).**

previously written state [7]. The DRF in the cell is modelled by an asymmetric defect resistance $R3$ (see Figure 4.4). If $R3$=50MΩ, then such a highly resistive defect is not detected even at the elevated temperature of 150^{o}C and a pause of more than 100ms. When $R3$=60MΩ, the off-state current of $Q2$ is sufficient to gradually discharge the node B capacitance and at 450ns the cell flips, destroying the stored data. If the same test is conducted at room temperature, the lowest detected value of $R3$ will be 2.75GΩ and the test time necessary to detect it will be over 60 times longer. Obviously, the detection range of the DRT is insufficient to reliably identify many manufacturing defects that cause poor cell stability. Elevated temperatures help to improve the detection range by about

45 times at the cost of the increased test time. However, resistive opens of about 50MΩ are still considered to be strong opens [6]. Unless special tests are applied, such defects will pass the standard test and an SoC with highly unstable and unreliable SRAM cells will be shipped to the customer. More subtle defects caused by the process disturbances can also reduce the stability of the cell. The likelihood of resistive bridges grows as the critical area shrinks with scaling, and resistive break defects are likely to appear in place of poor or absent contacts, vias or silicide [7, 8]. While these defects can be non-catastrophic, they can have a serious impact on the cell stability as shown in Figures 3.15 and 3.16 on page 68, respectively.

Data Retention Fault vs. Stability Fault

A DRF represents a severe case of cell instability and is typically attributed to a serious fabrication defect. Let us consider a case where the cell stability is compromised just marginally less severe than in case of a DRF.

Depending on the severity of the SNM degradation, the stability problems in SRAM cells can be classified into the Data Retention Faults and the Stability Faults (SF) with the former being a subset of the latter, as shown in Figure 4.7 on the following page. For extremely low values of the SNM the cell is likely to flip its state if not rewritten with the same data again shortly after, i.e. it fails to retain its data demonstrating a Data Retention Fault. If the SNM is sufficient to handle the NMOS off-state leakage current that discharges the storage node, under normal conditions it can retain its data as long as the power is supplied to the cell. However, under the adverse conditions such as the reduced supply voltage, elevated temperature, increased coupling and supply noise etc., i.e. the conditions contributing to further SNM degradation, this cell may become so unstable as to flip its state. And finally, cells with SNM high enough to withstand the

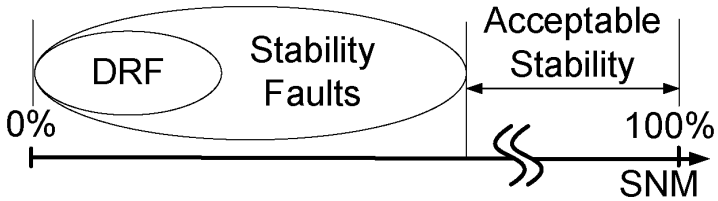


Figure 4.7 Relationship between the SNM, Data Retention Faults (DFRs) and Stability Faults (SFs) in an SRAM cell.

worst possible case scenario, are outside of the oval representing the stability faults. We will refer to the cells inside and outside of the oval representing stability faults as *weak* and *good* respectively.

A cell weakness that makes a cell vulnerable to SFs can be caused by various factors, including resistive defects (resistive breaks and bridges), excessive process shifts, mask misalignment, transistor mismatch etc. [10]. A stability fault, which is a possible consequence of cell weakness, may occur due to any electrical disturbance such as power supply noise, read/write cell disturbs, etc. during normal operation of the SRAM. These adverse conditions, especially combined, can cause a weak cell to flip its state easily and corrupt its contents.

Due to undetected resistive defects, inadequate SNM may indicate intermittent stability and possible long term reliability issues in SRAM cells, while such cells will successfully pass regular memory tests. For certain applications requiring extreme reliability (e.g. banking and enterprize servers; automotive, where SoC chips control such systems as ABS, stability control etc.) detecting all weak defects and possibly unstable cells is crucial. Thus, to achieve high product reliability and high quality tests, weak cell detection should be included in SRAM test suites.

4.1.2 Proposed SRAM Cell Stability Fault Model

With the advent of VLSI circuits, exhaustive functional testing has become unfeasible and has led to the appearance of structural tests aimed at detecting possible faulty conditions [26]. Such conditions have to be modelled by fault models. A fault model is a systematic and precise representation of physical faults in a form suitable for simulation and test generation [63]. A fault represents the electrical impact of a physical defect with a certain degree of accuracy and can be used to mimic such defect, in a simulation environment.

In a logic- or a transistor-level fault modeling all faults are assumed to be equally probable. However, in reality fault probability is a function of the probability of the defect that causes such a fault. Thus, layout-level fault modelling is essential for generation efficient and effective tests that are targeting the defects which are more likely to appear in the layout of the circuit under test in a given technology. Inductive Fault Analysis (IFA) [23], used in Section 3.4.2 to obtain the SNM sensitivity to the defect resistance showed that the most probable bridge defect in the layout of the SRAM cell under investigation was the bridge between its two data nodes.

Since the SNM is a measure of SRAM cell stability, its degraded value results in a cell stability fault that is parametric in nature. This fault can manifest itself under certain conditions by compromising the stored data integrity. We believe that development of a parametric stability fault model for SRAMs is crucial for the investigation and comparison of the effectiveness of various test algorithms and DFT techniques as well as for the stability characterization of SRAM designs. Therefore, we developed such a model. The proposed fault model mimics SNM degradation due to transistor mismatches, non-catastrophic defects and variation in operating conditions.

Let us consider the dependence of the SNM on the resistance between node A and node

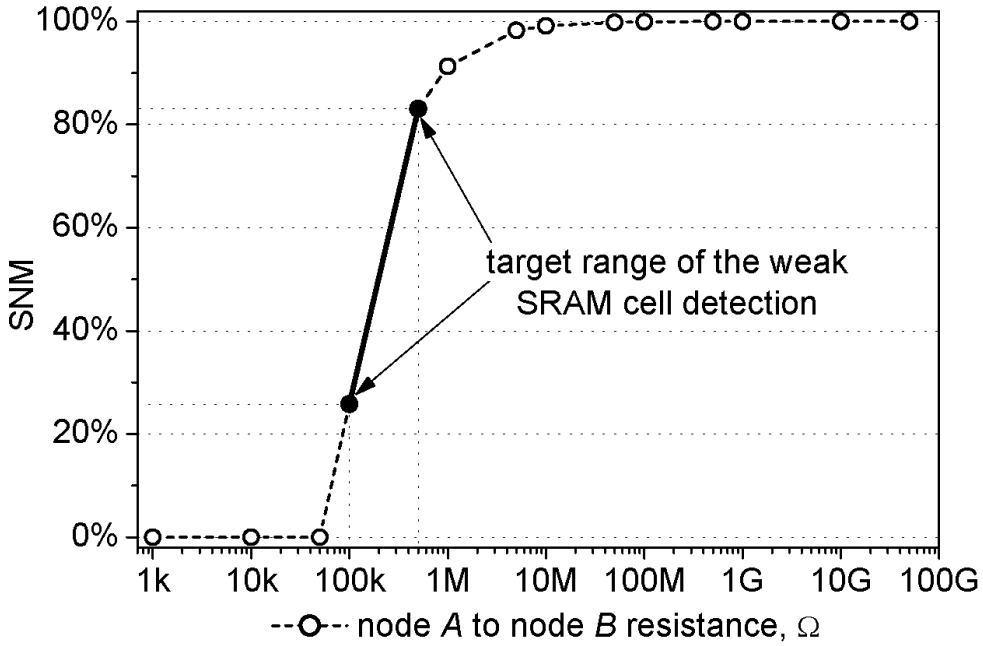
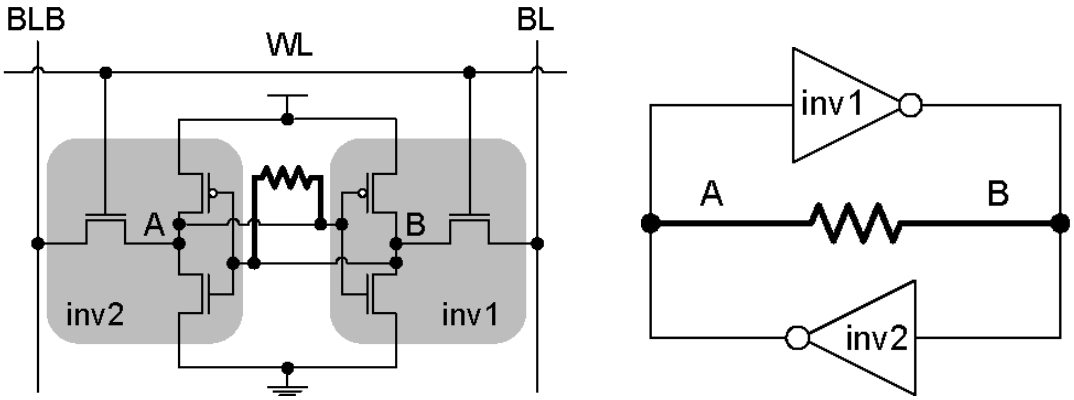


Figure 4.8 A possible target range of the SNM modelled by a given range of a resistor between node A and node B.

B as illustrated in Figure 4.8, and also in a thicker solid line in Figure 3.16 on page 68. For very large resistance values ($> 10\ M\Omega$), the cell SNM is not affected. For the resistance range between 50 $k\Omega$ and 1 $M\Omega$, the SNM is reduced linearly. The SNM becomes zero and causes catastrophic failure for the resistance values below 50 $k\Omega$. Depending on the parameters of a given SRAM cell, one can choose a particular resistance value in order to realize a weak cell with a pre-determined SNM. For instance, to obtain a cell with a half of the typical SNM, a resistor of 200 $k\Omega$ has to be used. A bold line in Figure 4.8 represents a possible target range for weak SRAM cell detection.

The resistor between node A and node B represents the proposed *weak cell fault model*, which is illustrated in Figure 4.9(a) on the next page. This SRAM cell has the worst-case SNM in the read-access mode when both the bit lines are precharged and the word line is



(a) (b)

Figure 4.9 Proposed Weak Cell Fault Model (a) and its equivalent circuit (b).

activated [34]. Each half of a read-accessed SRAM cell can be represented as an equivalent inverter, as shown in Figure 4.9(b). As we can see from Figure 4.9, a node-to-node resistive defect represents a negative feedback for the equivalent inverters comprising the SRAM cell. The corresponding reduction of the inverter gains and hence, the amount of the negative feedback in the cross-coupled inverters, is symmetrical and can be used to control the SNM. In a simulation environment, a cell with a resistor of a specified value between node A and node B can imitate a weak cell with a specified SNM value. The degree of the "weakness" is controlled by the value of the resistor. Provided other conditions are equal, the "weakened" cell has equal SNMs and symmetrical response for both high-to-low and low-to-high internal node voltage transitions. Thus, it represents a simple, symmetric and realistic weak cell fault model for simulation of parametric stability faults in SRAMs.

Intentionally inserting weak cells with the desired target SNMs into an SRAM array allows us to verify and fine-tune test techniques for parametric stability fault (weak cell) detection in the simulation environment. Having a simulation setup with a set of weakened cells with varying degrees of weakness (SNM) allowed us to evaluate various cell stability

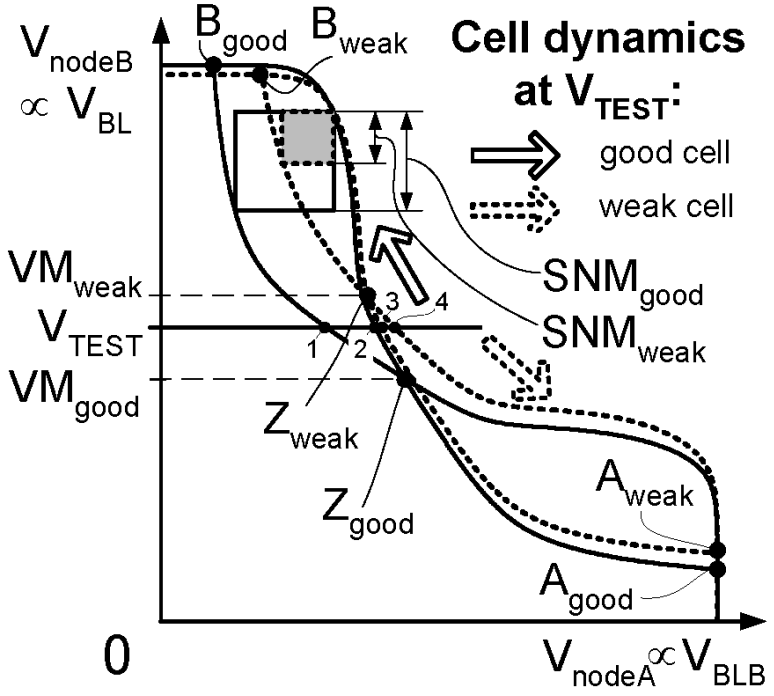


Figure 4.10 **Choice of V_{TEST} with respect to the VTCs of a typical and a weak SRAM cell.**

DFT techniques and algorithms. Moreover, such a setup can also be instrumental in the stability characterization and debugging stages of SRAM development.

4.2 SRAM Cell Stability Detection Concept

To provide better insight into the principles behind the detection of weak SRAM cells, we developed an SRAM cell stability detection concept. To illustrate the concept consider the voltage transfer characteristics of a good SRAM cell (solid lines) and a weak SRAM cell (dashed lines) presented in Figure 4.10. The axes in Figure 4.10 represent node A and node B voltages, which in turn, are proportional to the bit line voltages V_{BL} and V_{BLB}.

Normally, the real VTC of an SRAM cell is asymmetrical due to inevitable mismatches or defects. The SNM of the cell is proportional to the degree of asymmetry of the VTCs ($(SNM_{weak} < SNM_{good})$). Fluctuations in V_{TH} and L_{EFF}, the presence of defects, and poorly formed contacts and vias can make the driving strength of one of the inverters in a cell weaker. This results in the shift of the meta-stability point of the cell. Without loss of generality, let us suppose that a given cell's VTC is skewed so that its metastable point VM_{weak} is closer to node B. Since the metastable point is not equidistant from the the node potentials, the affected data node will be more vulnerable to disturbances than the other. Any noise disturbance exceeding the metastable point of the cell will cause such a cell to flip states. In other words, if a data node of an SRAM cell is driven to the level of VM, then a small voltage increment will flip the cell towards the direction of this increment.

VM_{good} and VM_{weak} represent node B voltages corresponding to the metastable points Z_{good} (Z_{weak}) of the good (weak) cell, respectively. Note that the metastable point of a weak cell is different from the metastable point of a good cell. The difference is proportional to the degree of asymmetry introduced into the weak cell VTC by a defect or a mismatch. This property of cell's VTCs can be exploited in the cell stability test techniques.

Let us assume that node B of an SRAM cell stores state "1" and that the bit lines are pre-charged to a known value (e.g. V_{DD}). Now assume that by a certain manipulation, V_{node_B} is reduced from a stable state B_{good} (or B_{weak} for a weak cell) to a certain test voltage V_{TEST}. Voltage level V_{TEST} intersects the good cell's and the weak cell's transfer characteristics at points "1" and "2" and at points "3" and "4" respectively, as shown in Figure 4.10. It is apparent from Figure 4.10 that the weak cell will flip its state if $(V_{DD} - V_{TEST}) < (V_{DD} - VM_{weak})$ in the direction of the dotted arrow. In other words, the weak cell will flip if its node B is driven below V_{TEST}.

If $V_{TEST} > VM_{weak}$, the regenerative property of a weak cell will restore the stable

state and the weak cell will not flip. This situation is similar to a non-destructive read operation of the cell under test with incompletely precharged bit lines. If $V_{TEST} < VM_{good}$, even the good cells will flip. This situation is similar to a normal write operation on the cell under test. V_{TEST} is in the target range if $VM_{good} < V_{TEST} < VM_{weak}$. This is the selectivity condition of weak cell detection during the stability test.

Upon removal of the test stimulus V_{TEST} node B of the good cell will retain its state "1", while node B of the weak cell will flip to state "0". The arrows in Figure 4.10 show the direction of the cell dynamics at V_{TEST}. After carrying out a cell stability test, all the cells, which flip after application of the node voltage above V_{TEST} are deemed "weak". The rest of the cells is assumed to have acceptable stability.

Test voltage V_{TEST} can have a fixed or a variable value. A fixed V_{TEST} allows for a single pass/fail threshold, whereas being able to vary the V_{TEST} value, one can test for a given degree of cell weakness and shift the pass/fail threshold according to the target quality requirements. Various cell stability test techniques presented next can be classified based on this and other principles.

4.3 Existing Weak SRAM Cell Detection Strategies

This section presents an overview of the existing DFT techniques for SRAM cell stability test. Based on the proposed classification of SRAM cell stability test methods, we will discuss the existing approaches seeking to replace the DRT.

4.3.1 Classification of SRAM Cell Stability Test Techniques

Practical SRAM cell stability test techniques employed in the industry can be classified based on whether special DFT circuitry is required to apply the test stress to a Cell Under

Test (CUT). Classification of SRAM cell stability test techniques is shown in Figure 4.11.

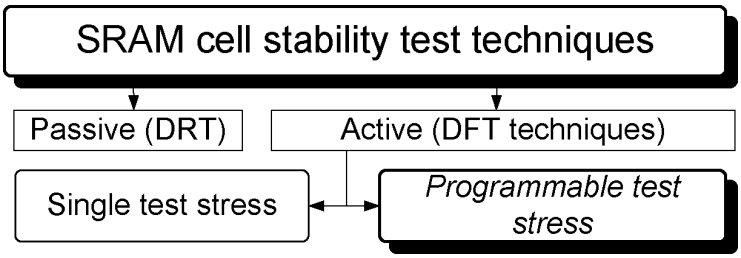


Figure 4.11 Classification of SRAM cell stability test techniques.

4.3.2 Passive Stability Test Techniques

Based on this criterion, cell stability test techniques can be classified into passive and active, where the passive techniques do not use DFT circuitry and the active techniques employ additional circuitry to actively stress the CUT. Passive test techniques include the Data Retention Test (a.k.a. Pause or Delay test) described in Section 4.1.1 and the Read Disturb test. The Read Disturb test includes writing a data background, disturbing the data by reading the array at a reduced V_{DD} and, finally, reading the array at full V_{DD} to determine if any cell has changed state [7]. Active or DFT techniques in turn can be classified into the techniques with single test stress, such as the WWTM [24] by Intel and the techniques with programmable test stress.

Due to limited defect coverage and significant test cost associated with lower tester throughput of the passive stability test techniques such as the DRT (see Section 4.1.1), the next sections will concentrate on active (DFT) techniques.

4.3.3 Active (DFT) Stability Test Techniques

To obtain acceptable noise margins in the deep sub-micron technologies with reduced supply voltages, bit lines in a vast majority of SRAM designs are precharged to the full supply voltage. However, reading a 6T SRAM cell with bit lines precharged and equalized at full V_{DD} may not detect several types of defects causing DRF or stability faults, e.g. a missing P-channel in the pull-up transistors, poor or absent vias to the pull-up transistors (an SRAM cell in this case will act as a "good" 4T DRAM cell). Detection of such cells in SRAM arrays may require a Data Retention Test. However, for large memory instances, DRT can take significant time leading to a more expensive test. Moreover, for stricter PPM levels, some cells may require excessively long DRT pause times, reduced supply voltage and high temperature and may still miss a wide range of resistive defects. Detection of such defects as poorly formed vias and contacts, shorts with non-zero resistances, and gross mismatches in cell transistors require special test conditions or stresses to be applied to make certain that parametric faults (i.e., stability faults) are reliably detected.

The standard suite of test methods often lacks sensitivity to detect parametric failures [64]. To provide better stability fault coverage, several DFT techniques for weak cell detection have been proposed in the literature [10, 7, 8, 9, 13, 16, 18, 12]. These techniques exploit the fact that the state-restoring feedback of a weak cell is weaker or absent and thus they are more susceptible to write or read disturbances. Initially, most of the weak detection techniques targeted the detection of the DRF. However, with scaling a large number of more subtle defects than completely open connections in the load transistors can remain undetected and lead to field failures, which often are intermittent and hard to diagnose. Detection of parametric stability faults needs to strike a balance between the yield loss of over-testing the cells, on one hand, and the test escapees due to under-testing the cells on the other.

Testing SRAM cells with a fixed weak write stress can lead to under- or over-testing of the targeted defects in SRAM cells due to poor process tracking characteristics. Single-threshold techniques are tuned based on the best available pre-silicon simulation data. To achieve an acceptable test quality versus test yield tradeoff, such techniques may require multiple post-silicon design iterations to account for the process modifications/changes following the initial design. However, if the weak write stress is programmable, the test quality versus test yield tradeoff can be adjusted without the design iterations and can be based on only on the results of the post-silicon testing such as using a separate test calibration block [19]. Obviously, programmable detection threshold techniques are superior in terms of time to market and test yield loss minimization and ensuring high quality.

Single Test Stress Methods

Initially, the proposed active methods offered a single stress setting, which was defined by the best available estimates of the process conditions. One of the well-known techniques, the Weak Write Test Mode (WWTM) [7], applies a weak overwrite stress to detect weak cells. The weak write circuit can be a stand-alone as in [7] or integrated into a write driver [13]. While the WWTM makes use of weaker write driver transistors, weaker access transistors [16] can also be used to apply weak write stress by underdriving the access transistors using lower word line voltage during a write operation. Conversely, an elevated word line voltage can be used to apply the test stress and detect a weak cell [18]. Detection of weak cells by reading at $V_{DD}/2$ while normal read operation is conducted after precharge to full V_{DD} is proposed in [17]. Kuo et al. [9] suggested a Soft Defect Detection (SDD) technique based on the fact that the defect-free inverters of an SRAM cell will provide certain read current upon access. If there is a break in a cell's connections, the read current is insufficient or absent. Another approach proposed by Kwai et al. is to separate the power

supply of the memory array from that of the periphery [8]. With the corresponding isolated terminal, the memory array can be operated at a lower voltage, making it susceptible to read or write disturb. However, this test alone cannot guarantee the detection of all DRFs and is mostly used for process development. Moreover, having a separate pad for each of tens of memory arrays in modern SoCs may be impractical.

Many of the single threshold methods can be altered to enable analog control of the weak overwrite stress. However, analog levels are more difficult to control on the global chip level if implemented internally, or are more pad- and tester-demanding if controlled at the ATE level.

Programmable Test Stress Methods

The increasing process spreads of modern deep sub-micron technologies have necessitated the arrival of digitally programmable techniques for weak cell detection, which better track process variation and/or allow to adjust the pass/fail threshold *during the test*. The importance of the ability to program the pass/fail threshold can be understood by inspecting Figure 4.12.

Figure 4.12 (a) shows the case when V_{TEST} is outside of the target range such that $V_{TEST} > VM_{weak}$. Applying V_{TEST} which is higher than the target range will not exsert sufficient stress to flip the weak cells and the weak cells will escape the test undetected. The case in Figure 4.12 (b) shows V_{TEST} positioned correctly between VM_{good} and VM_{weak}. And finally, Figure 4.12 (c) shows the case when V_{TEST} is outside of the target range such that $V_{TEST} < VM_{weak}$. Applying V_{TEST} which is lower than the target range will exsert too much stress causing even the good cells to flip leading to the yield loss. Cases (a) and (c) are likely to happen in case of severe process variations or using process splits during the product yield bring-up. Ability to program the pass/fail threshold in fine steps

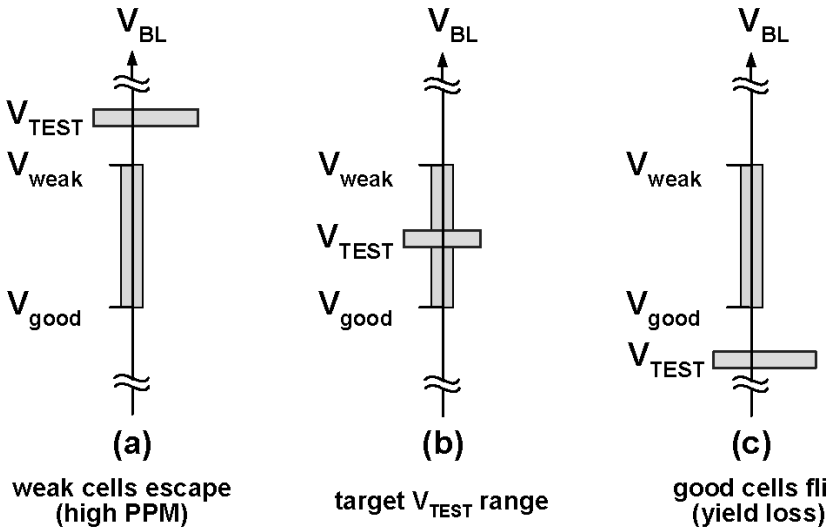


Figure 4.12 **Programming V_{TEST} to set a correct pass/fail test threshold: (a) V_{TEST} is set too high, which causes the weak cells to escape (high PPM), (b) V_{TEST} is set correctly, (c) V_{TEST} is set too low, which causes the good cells to flip (yield loss).**

provides for adjusting the pass/fail threshold on the fly without the need to redesign the DFT.

Programmable stability test techniques enhancing manufacturability by the ability to identify and repair defects are being adopted by the industry in the latest designs. For instance, the new Intel's Itanium-2 processor with the largest reported to date 6Mb on-die L3 cache [65] uses a programmable stability test technique proposed by Selvin et al [12]. It represents one of the possible extensions of the WWTM by using a decoder that switches the bias-setting transistors to vary the overwrite stress applied to the CUT [12]. This solution requires a dedicated bias voltage generator and the number of possible stress settings is limited by the number of the predefined bias settings.

We have developed three new digitally programmable techniques for the stability test in SRAM cells. Details on their concepts, design and detection capabilities are presented

in Chapter 6.

4.4 Summary

In this chapter, I described the main causes for Data Retention Faults and a presented a DRF fault model utilizing a symmetrical and asymmetrical defects in the pull-up path of an SRAM cell.

I generalized the DRF model in the proposed SRAM Cell Stability Fault Model. The proposed model allows to mimic an SRAM cell with an arbitrary value of the SNM. The proposed SRAM cell stability fault model will be used in the following chapters to verify the new digitally programmable techniques for stability tests in SRAM cells.

The introduced a concept of stability detection in SRAM cells. This concept helps to illustrate the principle behind cell stability test. Finally, I introduced a classification of the existing industrial SRAM cell stability test techniques. I showed that the single pass/fail threshold techniques are not adequate anymore for reliable and economical SRAM cell stability test. I justified the introduction of the programmable pass/fail threshold as the means to track the process or quality level changes without the need for redesigning the DFT circuitry.

Chapter 5

March Tests for Stability and Dynamic Fault Detection in SRAMs

In an effort to establish the detection capabilities of March tests I investigated March 11N and Hammer tests. Using the test bench presented in Section 5.2, I established the ability of March 11N (Section 5.3) and Hammer (Section 5.4) to detect a weak SRAM cell represented by the proposed fault model presented in Chapter 4. The detection capabilities and test potential of March tests in the presence of resistive and capacitive coupling between the aggressor and the victim cell nodes as well as the bit lines of the neighboring columns are investigated in Section 5.5. The findings are summarized in Section 5.6.

5.1 Introduction

Functional march tests, reviewed in Section 1.3.1, remain the main SRAM test method [22]. By optimizing a march test set to a particular SRAM architecture and technology, the fault

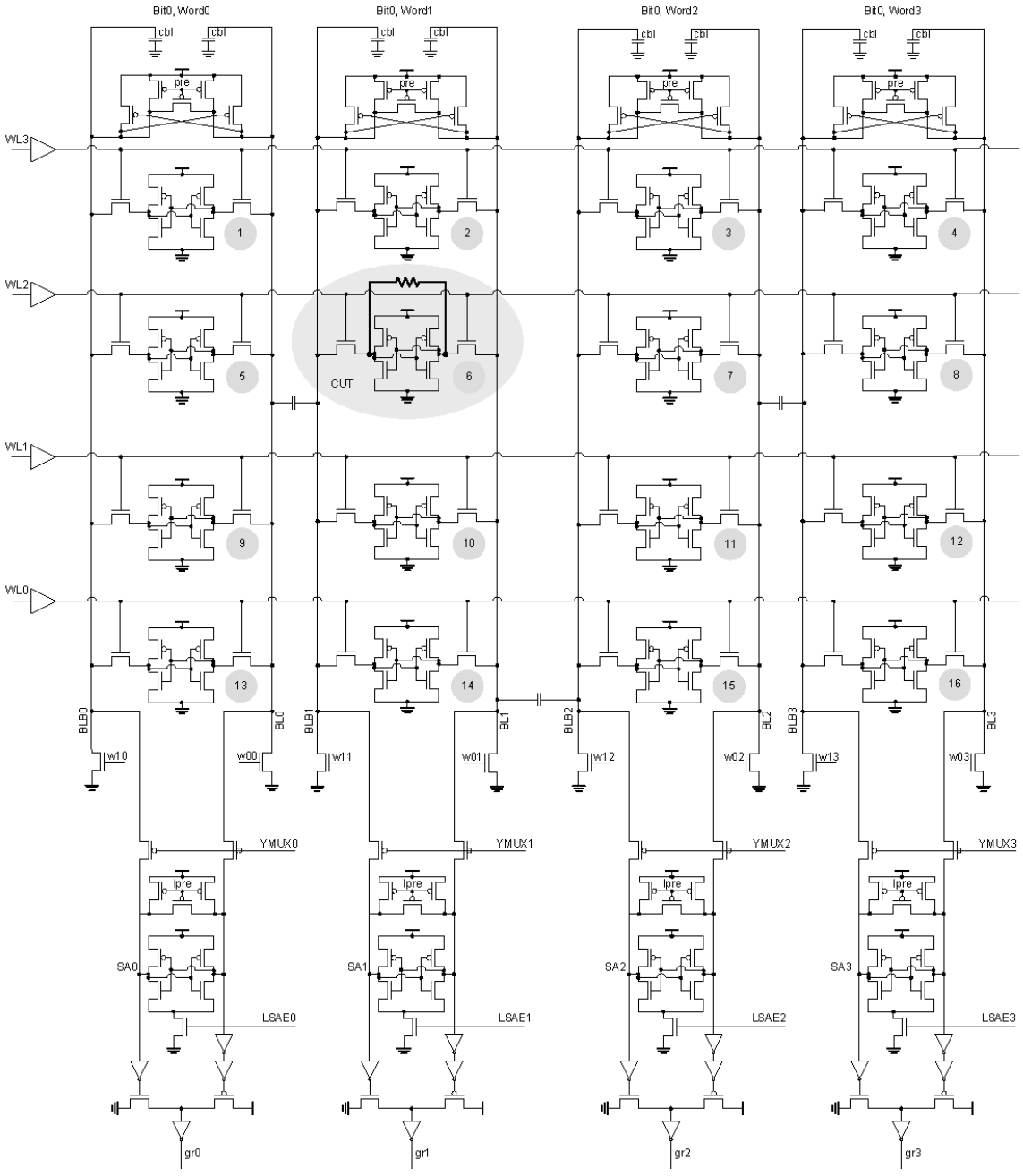


Figure 5.1 Test bench used for the march test experiments.

coverage of the "hard" faults used to be sufficient in many cases to ensure acceptable ppm levels which approach single-digit numbers [66]. However, with technology scaling, the growing probability of stability faults [10] and dynamic faults [67],[68] can cause a higher ppm, which is often unacceptable.

In this Chapter we investigate the capabilities of March 11N and Hammer tests to detect Stability Faults and Dynamic Faults in SRAMs.

5.2 Test Bench

The March 11N and Hammer tests have been simulated on an SRAM model containing an array of four columns and four rows of SRAM cells with the corresponding global (column) and local (SA) precharge/equalization, write and word line drivers, SAs with MUX pass transistors and output buffers, as shown in Figure 5.1 on the previous page. We generated timing signals on word lines, write driver inputs, column MUX, precharge and SAE for two extreme timing cases. The fastest access cycle time was assumed to be 2.4ns ($f = 417MHz$) to imitate a 4kb SRAM instance, whereas the slowest access cycle time of 6.5ns ($f = 154MHz$) was used to imitate a 1Mb SRAM. The bit line capacitance and the inter-bit line coupling capacitance were varied according to the simulated SRAM instance size.

According to the calculated word line and bit line parasitics, shown in Figure 5.2 on the following page, we distinguished three delay cases: slow, typical and fast. The choice of the delay case depends on the relative location of a cell under test to the word line driver or to the column MUX. Parasitic R and C values were varied depending on the chosen delay case.

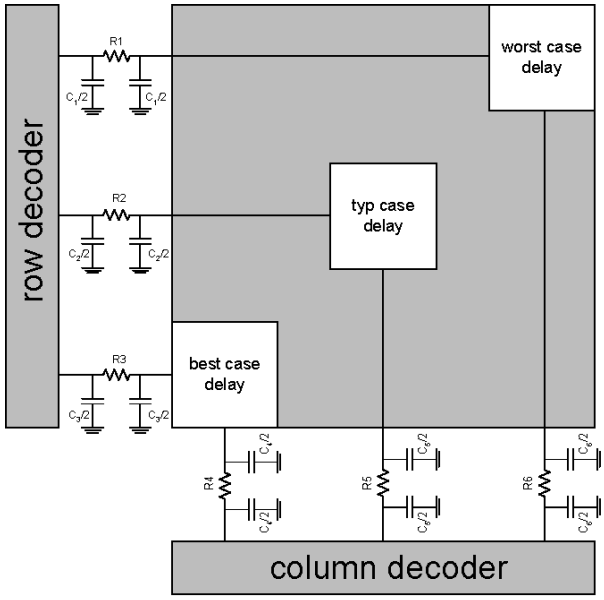


Figure 5.2 Delay cases depending on the accessed memory location in the array.

5.3 March 11N

Equation 5.1 describes the March 11N test that we applied. The first march element (w0) writes a background of all "0"s into the SRAM array.

$$\Updownarrow (w0) \Uparrow (r0, w1) \Downarrow (r1, w0, r0) \Uparrow (r0, w1, r1) \Downarrow (r1, w0) \tag{5.1}$$

Then, the value written in the first march element is read back during the first march operation (r0) with the expected value of "0". It is followed by a w1 operation that overwrites the same memory location with a "1". Next, the row address is incremented and the same march element (r0,w1) is repeated on the remaining rows in the same column. Address incrementing in this fashion is commonly called "fast column". Once the first march element has stepped through the first column, it is repeated for the next column in the same order until all columns are stepped through. Then, the third march element

(r1,w0,r0) is executed in the opposite row address order. The fourth (r0,w1,r1) and the fifth (r1,w0) march elements are incremented along the same row in the opposite order by changing the column address. Such an increment ordering is commonly called "fast row". To simplify the simulations, separate SAs were used for each column. In our experiment this is equivalent to a column decoder and a single SA corresponding to a 4 byte of X bits architecture.

Table 5.2 explains the March 11N progression through the test bench shown in Figure 5.1 on page 110. The notation used in Table 5.2 is as follows: rN denotes a read operation; wN0 denotes write a "0" and wN1 denotes write a "1", where #N is the cell number according to the numbering in Figure 5.1.

Figure 5.3 **Word line and output waveforms of an 11N March test run on the test bench shown in Figure 5.1. For a cycle time of** $2.4ns$ **and a slow process corner,** $R_{node\ A-node\ B} = 100k\Omega$ **- correct output,** $R_{node\ A-node\ B} = 90k\Omega$ **faulty output highlighted by the circles.**

An example showing the word line and output waveforms for the 11N March test with

Table 5.1 March 11N element operation sequence and addressing of the test bench in Figure 5.1 on page 110.

	fast column scan		fast row scan	
⇕(w0)	**⇑(r0,w1)**	**⇓(r1,w0,r0)**	**⇑(r0,w1,r1)**	**⇓(r1,w0)**
background	↑↘↑	↓↗↓	→↖→	→↙→
(w130)	(r13,w131)	(r1,w10,r1)	(r13,w131,r13)	(r1,w10)
(w90)	(r9,w91)	(r5,w0a,r5)	(r14,w141,r14)	(r2,w20)
(w0a)	(r5,w1a)	(r9,w90,r9)	(r15,w151,r15)	(r3,w30)
(w10)	(r1,w11)	(r13,w130,r13)	(r16,w161,r16)	(r4,w40)
(w140)	(r14,w141)	(r2,w20,r2)	(r9,w91,r9)	(r5,w0a)
(w100)	(r10,w101)	(rv,w0v,rv)	(r10,w101,r10)	(rv,w0v)
(w0v)	(rv,w1v)	(r10,w100,r10)	(r11,w111,r11)	(r7,w70)
(w20)	(r2,w21)	(r14,w140,r14)	(r12,w121,r12)	(r8,w80)
(w150)	(r15,w151)	(r3,w30,r3)	(r5,w1a,r5)	(r9,r90)
(w110)	(r11,w111)	(r7,w70,r7)	(rv,w1v,rv)	(r10,w100)
(w70)	(r7,w71)	(r11,w110,r11)	(r7,w71,r7)	(r11,w110)
(w30)	(r3,w31)	(r15,w150,r15)	(r8,w81,r8)	(r12,w120)
(w160)	(r16,w161)	(r4,w40,r4)	(r1,w11,r1)	(r13,w130)
(w120)	(r12,w121)	(r8,w80,r8)	(r2,w21,r2)	(r14,w140)
(w80)	(r8,w81)	(r12,w120,r12)	(r3,w31,r3)	(r14,w150)
(w40)	(r4,w41)	(r16,w160,r16)	(r4,w41,r4)	(r16,w160)

the addressing order shown in Table 5.2 is presented in Figure 5.3 on the preceding page. Using the proposed SRAM cell stability fault model, presented in Section 4.1.2, we reduced the SNM of the weak cell to the desired values. Application of the March 11N test for

Table 5.2 Summary of March 11N effectiveness in detecting weak SRAM cells (cycle time 2.4ns).

	slow corner	typical corner		fast corner
		best case delay	worst case delay	
$R_{not\_detected}$	$100k\Omega$	$50k\Omega$	$50k\Omega$	$30k\Omega$
$SNM_{not\_detected}$	$75mV$	$10mV$	$10mV$	$5mV$
$R_{detected}$	$90k\Omega$	$45k\Omega$	$47k\Omega$	$25k\Omega$
$SNM_{detected}$	$60mV$	$8mV$	$9mV$	$2mV$

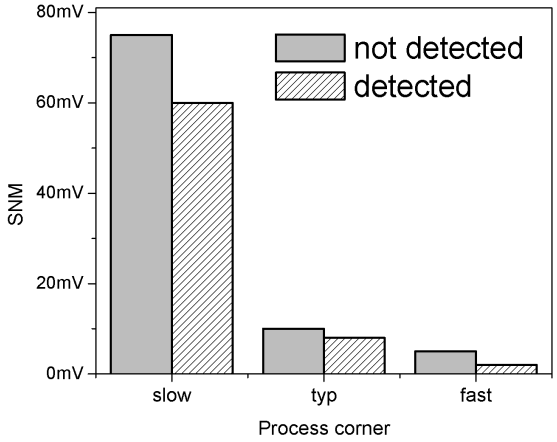


Figure 5.4 Summary of March 11N effectiveness in detecting weak SRAM cells.

the slow process corner is shown to be ineffective for the detection of a weak cell with a $R_{node\ A-node\ B} = 100k\Omega$, which corresponds to the SNM of 75mV. However, a cell with the SNM of 60mV obtained by a $R_{node\ A-node\ B}$ of $90k\Omega$ can cause a destructive read, which is detected by the March 11N. The discrepancy locations between the fault-free and the faulty behaviors are highlighted in Figure 5.3 by circles. These locations correspond to the weak cell #6 in Figure 5.1. They indicate a destructive read during the third and the fourth march elements (r1,w0,r0) and (r0,w1,r1) respectively. The first read operation of

the third march element produces a "0" instead of the expected "1", whereas the first read operation of the fourth march element produces a "1" instead of a "0".

However, for the typical and fast process corners, the detected values of the SNM drop to unacceptably low levels. Table 5.2 summarizes the detection capability of the 11N March test for various process corners. The detected SNM value in the typical process corner is around 10mV, whereas in the fast process corner it drops to 2mV, which is extremely low. An SRAM cell with such a small SNM is highly vulnerable to disturbances and cannot be tolerated. Figure 5.4 on the previous page graphically represents the decline in weak cell detection capability with the process corner variation from slow to fast.

5.4 Hammer Test

Most of the SRAM test methods have been targeting functional fault models limited to static faults. Static faults are faults that can be sensitized by performing at most one operation. However, dynamic faults have been reported recently based on defect injection and fault simulation of the industrial SRAMs [68]. Dynamic faults can be sensitized by more than one operation sequentially.

Equation 5.2 presents a Hammer test with test length of 49N.

$$\Updownarrow (w0) \Uparrow (r0, 10 * w1, r1) \Downarrow (r1, 10 * w0, r0) \Uparrow (r0, 10 * w1, r1) \Downarrow (r1, 10 * w0, r0) \quad (5.2)$$

The main feature of this Hammer test is that the write operation is performed on the same cell ten times successively. It is denoted as $10 * w1$ or $10 * w0$ for a "1" and a "0", respectively. The address increment order for the Hammer test was assumed "fast column" as opposed to the interlaced "fast column" and "fast row" of the 11N March test.

The Hammer test is classified as a Repetitive Test [66], i.e., a test that repetitively applies multiple write or read operation to a single cell. Repeating the test makes partial,

hard-to-detect fault effects become full fault effects.

Dynamic faults can be caused by the same reasons as the stability faults, i.e. mostly by mismatches, defects and environmental stresses. Since only a defective cell will exhibit a dynamic fault, in the context of this work Dynamic Faults and Stability Faults can be treated as equivalent.

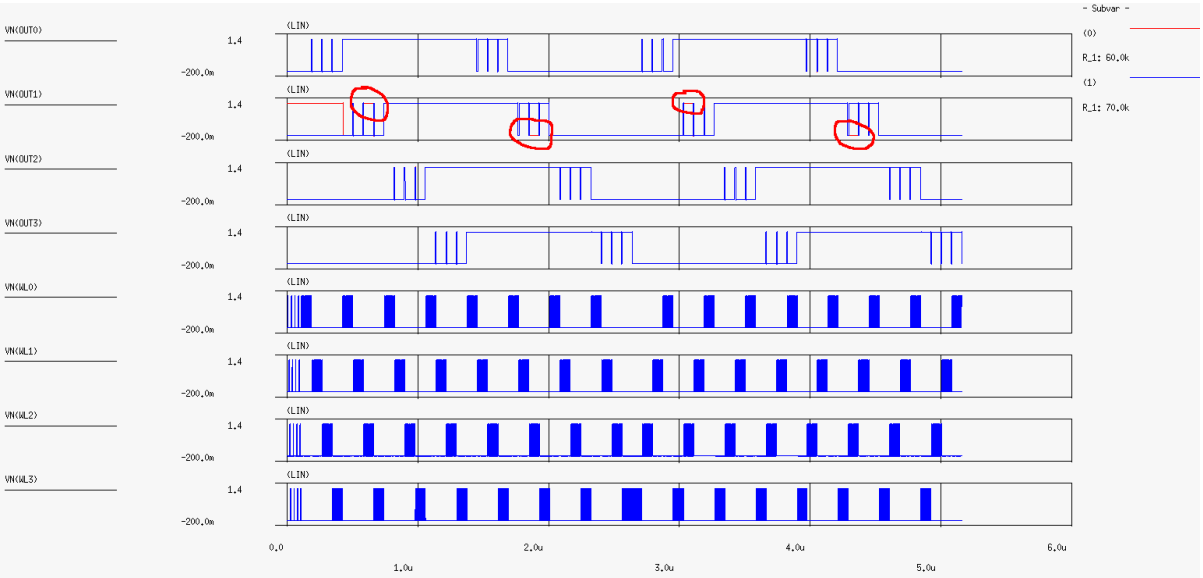


Figure 5.5 Word line and output waveforms of the Hammer test running on the test bench shown in Figure 5.1. For cycle time $6.5ns$ and typ process corner, $R_{node\ A-node\ B} = 70k\Omega$ - correct output, $R_{node\ A-node\ B} = 60k\Omega$ faulty output highlighted by the circles.

Figure 5.5 illustrates the detection capabilities of the Hammer test with cycle time of $6.5ns$. Similarly to 11N March, Hammer test can detect only unacceptably low SNM values. For instance, for the case presented in Figure 5.5, the largest detected $R_{node\ A-node\ B}$ is $60k\Omega$, which translates into the SNM value of about 12mV.

Successful read after multiple writes requires strong precharge transistors, precise timing and a cell with larger cell ratio, which translates into a larger SNM value. Experiments

with a Hammer test at 2.4ns access time with $10 * w$ operations has been unsuccessful even with $R_{node\ A-node\ B} = \infty$, i.e., with all typical cells. Investigating the reason for that could be a research topic for the future work.

Many variations of the Hammer test are possible depending on the number of times each read or write operations occur in each element of a Hammer test [66]. I experimented with modified versions of the Hammer test having $3 * w$ and $2 * r$ operations in all but the first test elements. The modified Hammer test is presented in Equation 5.3:

$$\Updownarrow (w0) \Uparrow (2{*}r0, 3{*}w1, 2{*}r1) \Uparrow (2{*}r1, 3{*}w0, 2{*}r0) \Downarrow (2{*}r0, 3{*}w1, 2{*}r1) \Downarrow (2{*}r0, 3{*}w1, 2{*}r1) \tag{5.3}$$

The waveforms showing the detection capabilities of an example of the modified Hammer test are shown in Figure 5.6.

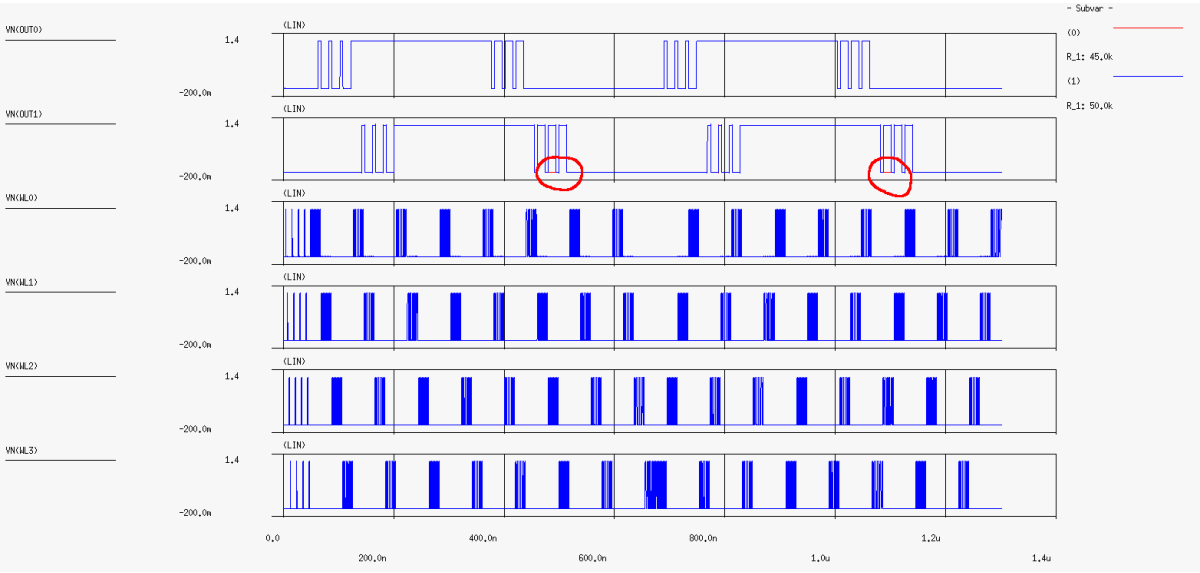


Figure 5.6 **Word line and output waveforms of Hammer+ test run on the test bench shown in Figure 5.1. For cycle time of $2.4ns$, typ process corner and the best delay case, $R_{node\ A-node\ B} = 50k\Omega$ - correct output, $R_{node\ A-node\ B} = 40k\Omega$ faulty output highlighted by the circles.**

From the results depicted in Figure 5.6 it is apparent that modification of the Hammer test does not provide improved fault detection capability. The detected SNM of a Hammer test in this case lies below 10mV, which is extremely low and unacceptable for the reliable detection of stability faults.

5.5 Coupling Fault Detection

In this work we established the detection capabilities and test potential of March tests in the presence of resistive and capacitive coupling between the aggressor and the victim cell nodes as well as the bit lines of the neighboring columns. Coupling between the neighboring rows was not considered due to power busses running between the rows on the same metal layer. A resistive bridge across such busses will likely cause a stuck-at fault and would therefore be easily detected by the standard test methods.

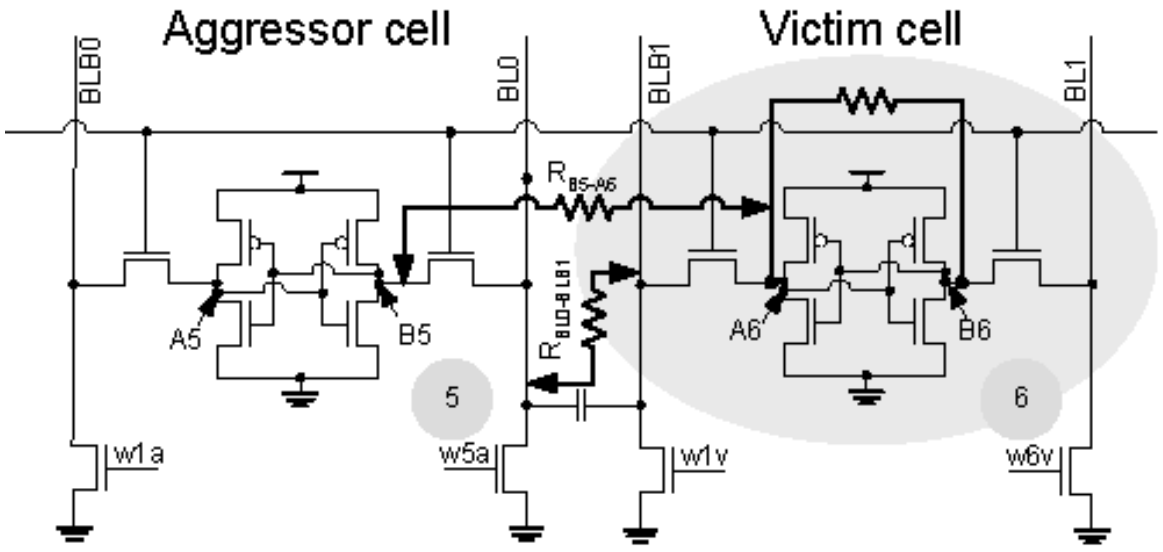


Figure 5.7 **An excerpt of Figure 5.1 on page 110 showing the coupling resistive bridge defects (R_{B5-A6} and $R_{BL0-BLB1}$) and inter-bit-line capacitance between the bit lines of the aggressor and the victim cells.**

Applying the principles of IFA, we analyzed the layout of the cell under test and defined the two most probable bridging resistive faults between two neighboring columns that are shown in Figure 5.7. Resistors R_{B5-A6} and $R_{BL0-BLB1}$ represent a resistive bridge between node B of the aggressor cell #5 (B5) and node A of the victim cell #6 (A6). The resistor between nodes A6 and B6 of the victim cell represents the Cell Stability Fault Model described in Section 4.1.2.

For the test bench in Figure 5.7 I developed a short march test consisting of two march elements:

$$(w1v, rv, w1a, rv); (w0v, rv, w0a, rv). \tag{5.4}$$

where $w1v$ and $w0v$ denote write "1" and "0" operations on the victim cell; $w1a$ and $w0a$ denote write "1" and "0" operations on the aggressor cell; and rv and ra denote read operations on the victim and the aggressor cells respectively.

Each of the test elements starts out by writing the background value to the weak victim cell #6 followed by a read operation on the same cell to establish a reference value. Next, the same data background is written to the aggressor cell #5. Since the aggressor and the victim cells are in adjacent columns, the data backgrounds written in the cells effectively result in opposite values stored on node A5 and node B6. If a bridging defect with sufficiently low resistance R_{B5-A6} or $R_{BL0-BLB1}$ is present between the columns, the victim cell can be inadvertently overwritten by the aggressor cell. The following read operation will determine whether the victim cell has changed states. The second march element is the inverse of the first and is needed for testing the victim cell's susceptibility to the aggressor cell with the inverse data background.

The second read operation in each of the march elements is redundant and was introduced for observation convenience. It can be removed without sacrificing the detection

capability. The optimized test sequence is shown below:

$$(w1v, w1a, rv); (w0v, w0a, rv) \tag{5.5}$$

In addition to the conditions specified in 5.5, the successful detection of a weak cell depends on the resistance values of R_{B5-A6} or $R_{BL0-BLB1}$ and the severity of the SNM degradation in the victim cell. All other conditions being equal, aggression on the node of the victim cell that stores a "1" overwrites the victim cell more easily. This is understood keeping in mind the design principles of 6T SRAM cells. The PMOS pull-up transistor is normally designed to be 1.5–2 times smaller than that of the NMOS driver transistor. Multiplying the size ratio by μ_n/μ_p shows that the current ratio $I_{pull-down}/I_{pull-up}$ can be significant. Therefore, if node "1" of the cell with compromised stability is coupled to the aggressor via a sufficiently low resistance of a bridging defect, it is more likely to be overwritten. Node "0" of the victim cell is vulnerable to the inadvertent overwriting when coupled by an aggressor with a smaller bridging defect resistance.

Experimenting with resistance values of R_{B5-A6}, $R_{BL0-BLB1}$ and R_{A6-B6} showed that the developed test sequence is capable of detecting resistive coupling faults with resistance below $10k\Omega$s. Bridging defects of this resistance range are usually targeted by other memory tests as well.

The test is more sensitive to R_{B5-A6} than to $R_{BL0-BLB1}$, i.e. it can detect somewhat higher values of R_{B5-A6}. The aggressor and the victim cells share the same word line. If a $R_{BL0-BLB1}$ bridge exists, then during the writing a "0" to the aggressor cell, the precharge in the aggressor column is turned off and BL0 is discharged to the ground. The ground potential is then coupled to the victim cell via the path consisting of the access transistor of the aggressor cell, $R_{BL0-BLB1}$, and the access transistor of the victim cell. Moreover, since the precharge in the victim column is on, BLB1 is being held at V_{DD} by a strong precharge transistor which is counteracting the disturbance. Once the access cycle is over,

the aggressor and the victim cells are disconnected from their bit lines and $R_{BL0-BLB1}$ has no effect on their contents. R_{B5-A6} on the other hand, is coupling the victim and the aggressor cells all the time regardless of whether this particular row is active or not. A write "0" operation to node B5 in this case can destroy a "1" stored on node A6 more easily.

Detection of a coupling defect R_{B5-A6} in the case when the aggressor node B5 is written a "1" and the victim node A6 stores a "0" is very unlikely unless the victim cell's pull-down NMOS is seriously damaged. In fact, it can even cause the situation, when a write access to the victim cell destroys the data in the aggressor cell, i.e., the aggressor and the victim cells would change places. A damaged NMOS pull-down transistor reduces the I_{read} of the cell and is easier to detect as it may fall out of the read timing specification. Such defects can be easily detected by simple functional memory tests.

Capacitive coupling between the bit lines of the neighboring columns did not play a notable role. In the considered layout the bit lines are divided by a ground wire on the same metal layer. Even in a layout without a ground wire between the columns, the capacitance between the neighboring columns is unlikely to cause a coupling fault in many SRAM architectures. Typically, bits of the same significance belonging to different bytes/words are laid out in adjacent columns and share a common SA through the column MUX. When write accessed, only one of the neighboring bit lines can be driven at a time from the precharge potential (V_{DD}) to the ground. At the same time, the neighboring bit line is held at V_{DD} by a strong precharge transistor. The capacitive coupling charge current of the neighboring bit lines is then compensated for by the current of the precharge transistor.

The principle used in this test sequence can be incorporated into a more comprehensive march test to extend their capabilities in coupling fault detection.

5.6 Summary

Detection of a weak cell is possible if it has changed states as a result of a disturbance. Some of the possible conditions for a weak cell to change states during a March test are repetitive write and/or repetitive read operations or resistive and/or capacitive coupling to its neighbors. A subsequent read operation can then detect a flipped cell. Our simulations showed that the detection capabilities of the March 11N and Hammer repetitive test are insufficient for the reliable detection of stability and dynamic faults, which are caused by the same reasons as the stability faults. More reliable test methods have to be applied to ensure a high quality stability test.

Small memory instances with a shorter access cycle are generally more demanding to the precise internal timing generation for the stable operation. Moreover, due to the smaller ratio of t_{access}/t_{flip}, stability fault detection with repetitive tests is more likely to be successful for high-speed memories with shorter access times.

We established the conditions for the successful detection of resistive coupling faults in the neighboring SRAM columns. However, detection capability has proved to be limited to low bridging defect resistance values.

In the following chapter we will introduce several novel digitally programmable DFT techniques, which exceed the capability of the march and repetitive tests and the DRT. The reduced test time and superior defect resistance coverage offered by the proposed DFTs facilitate a more economical and reliable way to ensure high quality and low PPM levels in shipped parts.

Chapter 6

Programmable DFT Techniques for Stability Fault Detection in SRAM Cells

This chapter presents the main contributions of this work. Three novel DFT techniques for SF and DRF detection in SRAM cells are presented. The Read Current Ratio with a Pass Transistor (RCRPT), the Read Current Ratio Technique with floating bit lines (RCRT) and the Word Line Pulsing Technique (WLPT) are introduced in Section 6.2, Section 6.3 and Section 6.4, respectively. All of the proposed techniques offer a flexible, digitally programmable pass/fail threshold and exceed the capabilities of the known SRAM cell stability test techniques. The advantages of the proposed DFT techniques over existing stability test methods are summarized in Section 6.6.

6.1 Introduction

As shown in Chapter 5, the regular functional tests are often ineffective and inefficient in detecting stability faults in SRAM cells. The SNM values that can be detected with the functional test are extremely small. Such small SNM values are unacceptable for the quality screening of cell stability problems in SRAMs. As an industrial guideline, $\mu - 6\sigma$ of SNM is required to exceed $\simeq 4\%$ of V_{DD} to reach 90% yield on 1Mb SRAM [50]. Typically, that translates into a requirement that $SNM_{min} \geq 20\% SNM_{typ}$. The $\mu - 6\sigma \geq 4\% V_{DD}$ rule was established to account for SNM degradation only due to V_{TH} mismatch among cell transistors.

Process variations and parameter mismatches impose fabrication specifications for parametric yield and appropriate number of σ over which an SRAM of a certain bit count should work correctly. For instance, a 4MB (32Mb) cache SRAM with ECC contains over 38 million cells. Limiting a design to one unstable cell in 38 million requires operation over a greater than 5σ parametric variation tolerance [69]. Hence, even with the ability to repair a few cells with the worst mismatch, variations must be acceptable to 5σ or beyond to achieve reasonable yield for such arrays. Cells with marginal parameter matching have to be identified and preferably repaired by replacing them with redundant cells.

However, as was shown in Chapter 3, V_{TH} offset/mismatch is only one of the many factors that can severely deteriorate the SNM and thus the stability of the cell. If more such factors are affecting a certain cell or a group of marginal cell simultaneously, the resulting SNM can become unacceptable for stable SRAM operation in real operating conditions.

Moreover, traditional functional tests, such as the march tests, may fail to provide accurate fault diagnostic and debug capabilities. The information on the nature and severity of cell defects instead of pass/fail information only may be instrumental during the product

yield ramp-up stage and redundancy calculation and allocation.

We will introduce several novel defect-oriented DFT techniques for SRAM cell stability testing that address the limitations of functional tests. The pass/fail threshold of the proposed techniques is digitally programmable. This means that the proposed techniques enable adjustments of the pass/fail SNM values "on-line", i.e. without the need for redesign and consequent fabrication a.k.a. post-silicon iterations. The variable pass/fail threshold is achieved by being able to program the overwrite stress to the cell under test (the weak cell).

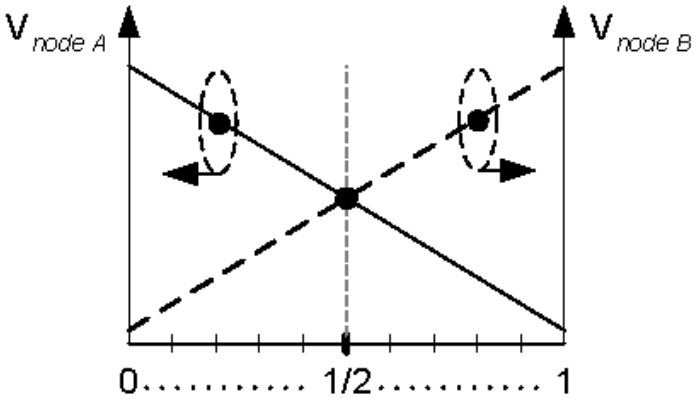
6.2 Read Current Ratio with a Pass transistor Technique (RCRPT)

6.2.1 RCRPT Concept

The first proposed DFT [10] implements the concept of programmable pass/fail threshold by using a set of n SRAM cells in a given column. Existing cells in the column or external cells can be utilized for this purpose. Let R be the ratio of cells in state "0" to the total number of cells in a set n (Figure 6.1).

We assume that the rest of the cells in a set n are in state "1". Initially, BL and BLB are precharged to V_{DD}. By manipulating the value of R, and simultaneously accessing n cells, we can manipulate the bit line voltage. As can be seen from Figure 6.1, Figure 6.4, Figure 6.5, by varying the value of R, we can control which bit line will have a higher potential.

Now, if we write a ratio R, simultaneously enable n word lines, and short the bit lines together, we can reduce V_{node_A} or V_{node_B} to a given V_{TEST} value. A large V_{TEST} will not flip



$$R = \frac{number\ of\ cells\ with\ state\ \text{“0”}\ in\ a\ set\ of\ n\ cells}{set\ of\ n\ cells}$$

Figure 6.1 V_{BL} **and** V_{BLB} **as a function of the programmable ratio R.**

any of the cells as it will be similar to a read operation. A smaller V_{TEST} (around VM_{weak}) will flip the weak cells. And finally, when V_{TEST} approaches VM_{good}, it can overwrite even the good cells. Therefore, by varying the ratio R, we can program a detection threshold for detection of SRAM cells with varying degrees of weakness.

Since R defines how many cell read currents I_{read} will be discharging each of the bit lines BL and BLB, we coined this testing technique the Read Current Ratio with a Pass transistor Technique (RCRPT). The purpose of the pass transistor (transistor $Q3$ in Figure 6.3 on page 129) will be explained below.

The flow diagram shown in Figure 6.2 on the following page depicts the sequence of steps necessary for digitally programmable weak cell detection. It is assumed that we can program the trip point of a weak cell by selecting an appropriate 0/1 ratio (R) of cells. Since the defects in an SRAM cell can reduce the write margins asymmetrically, applying the stability test to only one of the two cell storage nodes is insufficient. An inverse 0/1 ratio is necessary to detect the weak cells that may flip in the opposite direction.

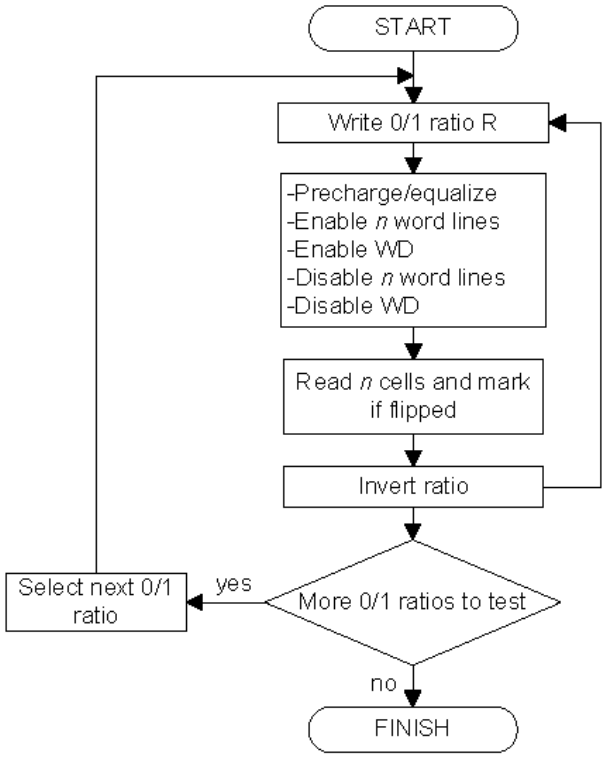


Figure 6.2 Flow diagram of RCRPT for programmable stability fault detection in SRAM cells.

6.2.2 RCRPT Implementation

Figure 6.3 on the next page shows the hardware required for one of the implementations of the proposed technique. Ellipses surround additional and/or modified circuitry. Figure 6.3 represents one of the SRAM cells in a column, two cross-coupled PMOS transistors ($Q1$, $Q2$) to pull up the bit lines, three other PMOS transistors ($Q4 - Q6$) to precharge the bit lines to V_{DD}, one NMOS pass-transistor ($Q3$) to provide a conductive path between the bit lines. It also includes special logic to issue the Weak Detect (WD) signal, and a modified word line decoder to allow the simultaneous enabling of n word lines.

The weak-cell detection phase starts by programming the trip point that is necessary to

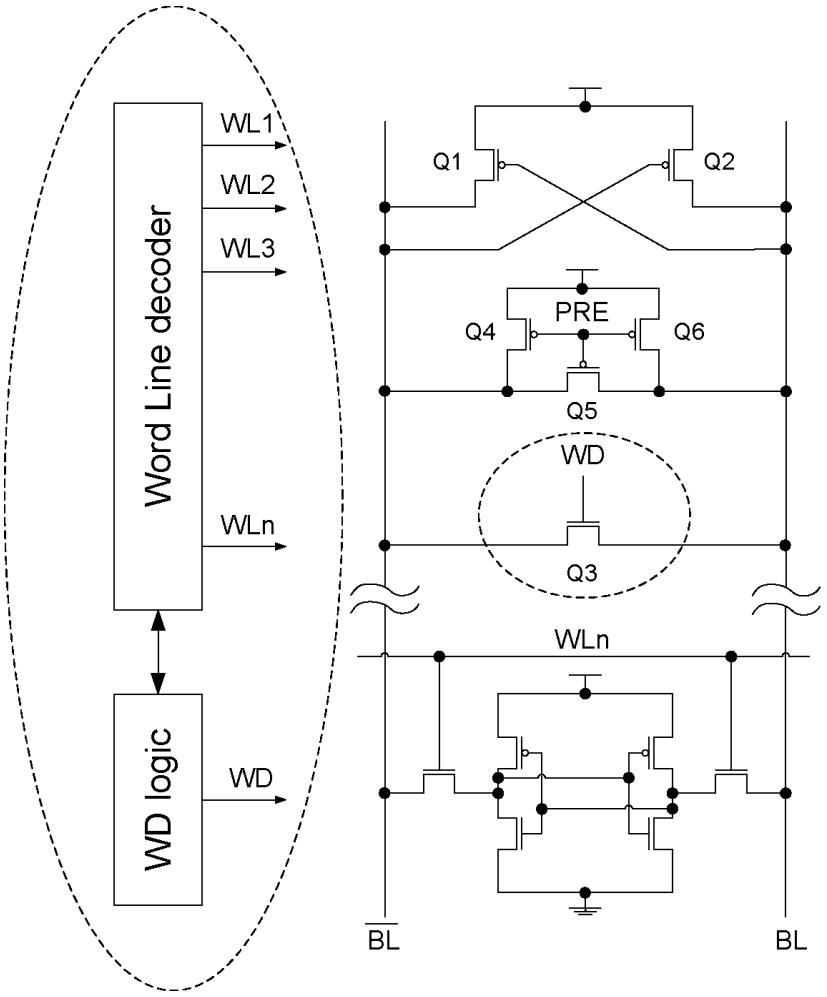


Figure 6.3 RCRPT hardware implementation

detect cells with the SNM below the target value. This is done by writing a predetermined number of cells with either a "1" or a "0" state. After normal bit line precharging finishes, n word lines are simultaneously enabled connecting in parallel n cells of the same column. Under this configuration, access transistors of each side of an SRAM column share a common gate and a common bit line nodes. The other terminal of each of the access transistors is connected either to the ground or to V_{DD} through the corresponding driver

or load transistors of an SRAM cell. The access transistors work as resistors dividing the power supply voltage on each bit line between V_{DD} and the ground depending upon the equivalent dc path resistance. For instance, bit line potentials will be around $V_{DD}/2$ when 50% of cells are in state "0" and 50% of cells are in state "1" because the path resistance to the ground and V_{DD} is the same, i.e. $R = 0.5$ (Figure 6.1).

When n word lines are enabled, the capacitance of each bit line discharges according to the time constant created by the corresponding equivalent path. If the bit lines discharge below the metastable point VM_{good}, even the good cells will flip pulling one of the bit lines even further to the ground and restoring the other one to V_{DD}. To prevent the cells from reaching the metastable point VM_{good}, the bit lines are shorted together through an NMOS pass transistor by applying a WD (weak detection) pulse. This causes the voltages at the bit lines to remain at around $V_{DD}/2$ while the cell dynamics finds a new equilibrium. In other words, the bit lines are not pulled to complementary logical values. However, a bit line voltage around $V_{DD}/2$ is already sufficient to flip the weaker cells with insufficient SNM. For a ratio $R \neq 0.5$, the corresponding path resistances to V_{DD} and the ground will be different and thus the bit line voltage is pulled earlier above or below $V_{DD}/2$.

6.2.3 RCRPT Detection Capability

To prove the effectiveness of this method we designed a setup with eight six-transistor SRAM cells in CMOS 0.13 μm technology with $V_{DD} = 1.2\ V$. The degree of weakness of one of the cells was manipulated by varying the resistance value of the resistor between node A and node B of this cell as per the proposed weak cell fault model. To verify the data retention fault detection capabilities, we also simulated the proposed DFT implementation with inserted resistive breaks in the load transistors.

Figure 6.4 and Figure 6.5 illustrate the voltage dynamic of node B and other signals

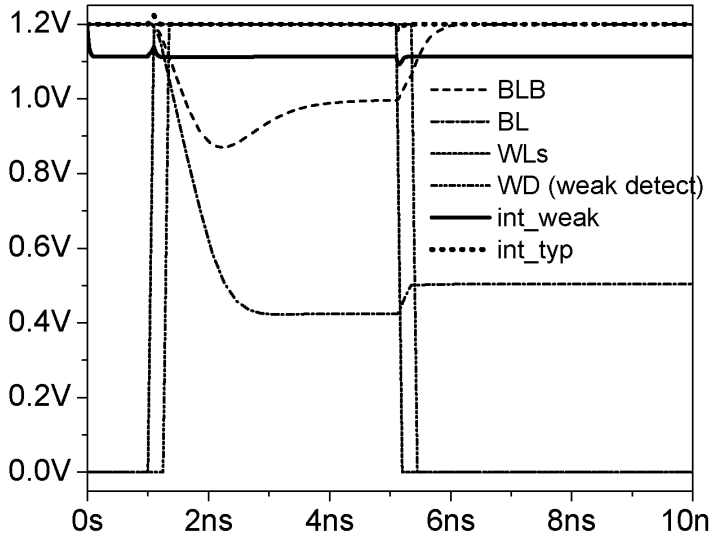


Figure 6.4 Voltage dynamics of node B and other signals for a weak cell (int_weak) and a reference normal cell (int_typ) for ratio R of 3/8

for the weak cell (*int_weak*) and a reference typical cell (*int_typ*) for ratio R of 3/8 and 5/8, respectively.

One cell was forced into a weak state by connecting nodes A and B with a resistor of 200 $k\Omega$. Evaluation of this cell with the inserted resistor of 200 $k\Omega$ gives an SNM of around 50% of the typical SNM.

A logical "1" state was stored in node B of the weak cell as well as in node B of a reference typical cell. To have a more realistic situation, the bit line precharge was simulated as well. After precharging both bit lines to V_{DD} and equalizing them, we enabled n word lines and shortly after that enabled the weak detect (WD) pulse to enter the weak detection mode. The cells state can be inspected at around 6 ns point.

When the ratio R is 4/8, the bit line-bar voltage drops to around 0.6 V ($V_{DD}/2$) but the weak cell's state does not flip. If the ratio R is 3/8, V_{BLB} rises up to about 1 V (see Figure 6.4). This voltage strengthens the weak cell and helps it to remain in its logical

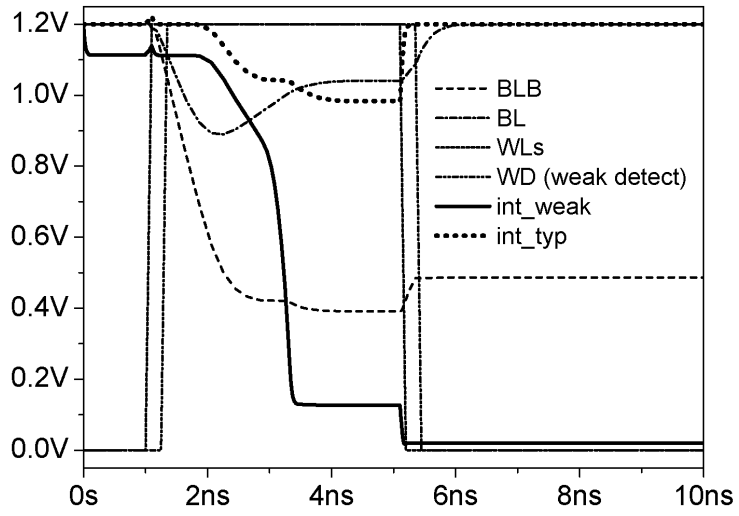


Figure 6.5 Voltage dynamic of node B and other signals for weak cell (int_weak) and a reference normal cell (int_typ) for ratio R of 5/8

"1" state (bold solid line *int_weak* in Figure 6.4). With ratio R of 5/8 (Figure 6.5), V_{BLB} drops down to about 400 mV forcing the weak cell to flip states (bold solid line *int_weak* in Figure 6.5).

Figure 6.6 demonstrates the detection capability of the proposed method. The resistance value of the node A to node B resistor for imitating a weak cell was swept from 100 $k\Omega$ to 500 $k\Omega$ and we used 0/1 ratio of 5/3.

Signal *int_weak* in Figure 6.6 represents node B of the weak cell. We can see that after applying the test sequence, the weak cell flips for resistance between node A and node B of 100 $k\Omega$ and 200 $k\Omega$. It this case, the SNM of the weak cell is too small to resist the overwriting disturbance and the cell is overwritten. Note that the cell does not flip with the node-to-node resistor values of 300 $k\Omega$, 400 $k\Omega$, and 500 $k\Omega$. In this case the SNM is large enough to resist the flipping. However, it is still possible to force such a cell to flip by choosing a different ratio R of n cells, i.e., by digitally programming the pass/fail

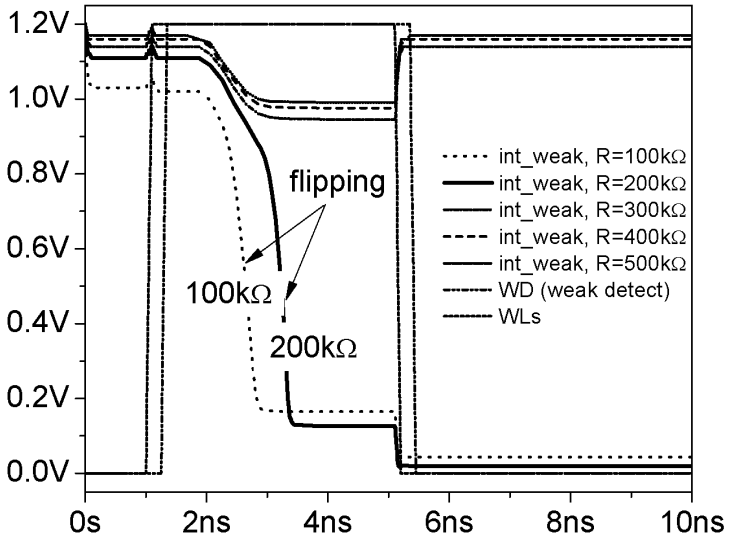


Figure 6.6 Detection capability of RCRPT for ratio R of 5/8

threshold of the test.

Similar waveforms were obtained for the detection of resistive opens in one or both load transistors, which confirms that the proposed DFT can be effectively used to detect both Data Retention and Stability Faults in SRAM cells.

6.3 Read Current Ratio Technique with Floating Bit Lines

6.3.1 RCRT Concept

This section introduces another SRAM cell stability test technique [70] based on the concept of Read Current Ratio. Unlike the RCRPT, the Cell Under Test (CUT) is not among the n cells whose word lines are simultaneously enabled. Moreover, the n cells forming the ratio R are used only to precondition the bit line potentials. The stability test stage is separated from the bit line preconditioning stage and occurs after the bit lines have been partially discharged.

Now, if we write a certain ratio of "0"s and "1"s to a set of n cells and then simultaneously enable the n word lines, the bit line potentials will be reduced according to the chosen ratio R. Subsequent application of the reduced bit line potentials to the CUT is used to apply the test stress. If the application of the reduced bit line potentials results in such a V_{TEST} value that $VM_{good} < V_{TEST} < VM_{weak}$, then only a weak CUT will flip its state, whereas the good cell will withstand such test stress. The subsequent read operation with the regularly precharged bit lines will detect such a weak CUT.

The VM_{good} and VM_{weak} values may change depending on the process conditions. Therefore, by varying the ratio R, we can program a detection (pass/fail) threshold for detecting SRAM cells with varying degrees of stability degradation. The programmability can also be used to set the V_{TEST} value to track the process conditions such that $VM_{good} < V_{TEST} < VM_{weak}$. This flexibility allows us to compensate for the influence of the process spread to avoid test escapees on one hand and marking the cells with acceptable stability as defective on the other.

The flow diagram shown in Figure 6.7 on the following page depicts the sequence of

steps necessary to implement the RCRT. Note that enabling of the n word lines of the cells forming the ratio R is followed by disabling the precharge. Then, reading the CUT with partially discharged bit lines applies the test stress to the CUT. A weak CUT will flip, whereas a good CUT will withstand the stress. An inverse of the current 0/1 ratio is necessary to detect the weak cells that may flip in the opposite direction due to the possible asymmetry in the VTCs of the CUT.

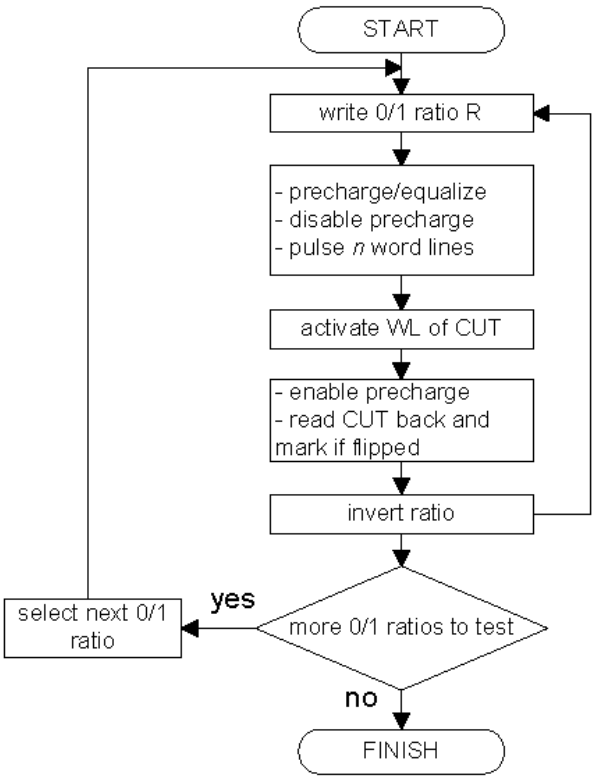


Figure 6.7 Flow diagram of RCRT for programmable stability fault detection in SRAM cells.

One of the possible implementations of the proposed technique is shown in Figure 6.8. The weak cell test starts by determining the minimal acceptable cell stability. The cells with SNM below that minimum must flip during the read operation with the partially

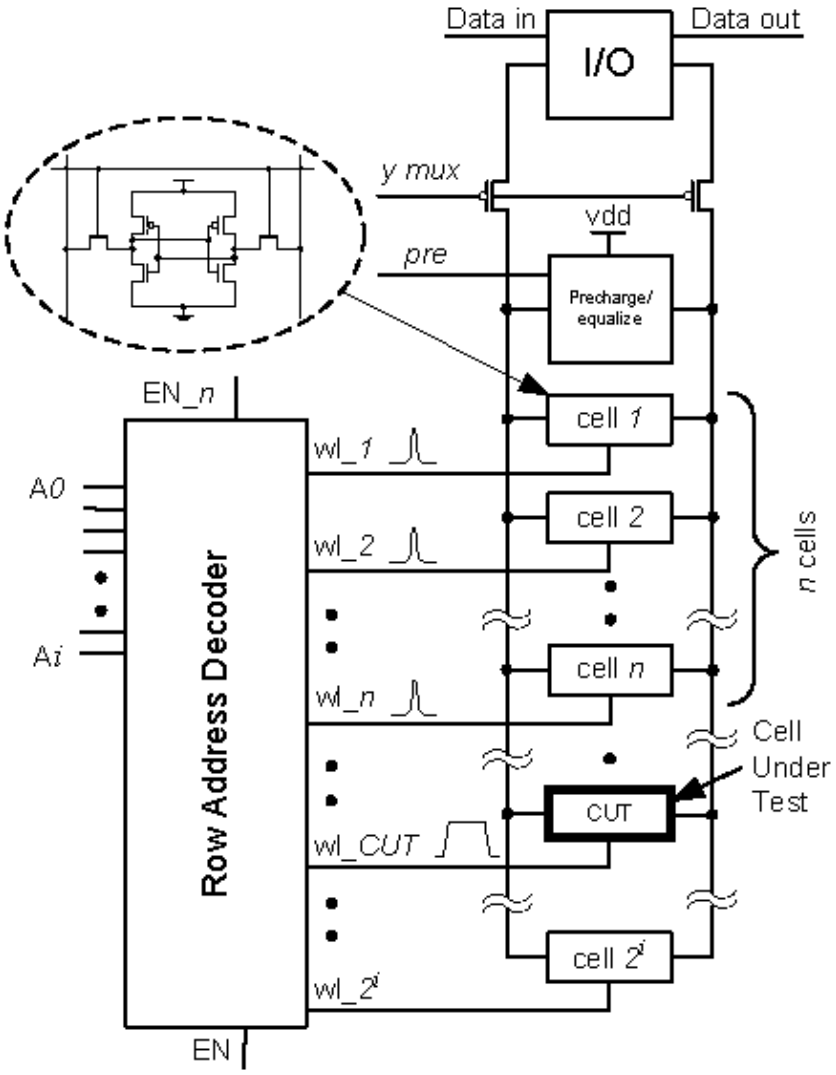


Figure 6.8 Block-level diagram showing the principle of RCRT.

discharged bit lines upon the application of the test stress.

The required bit line potentials are formed by writing n cells in the column with a ratio R that provides a desired V_{TEST} after applying a short pulse to the n word lines simultaneously. When the n word lines are enabled, the capacitance of each bit line discharges

according to the time constant created by the corresponding equivalent path. If the bit lines discharge too much, then upon the enabling of WL_{CUT} pulse, V_{TEST} can drop below the meta-stable point VM_{good}. In this case, even the good cells with acceptable SNM will flip. This situation can be corrected by shortening the pulse width of the pulse enabling n word lines or by reducing ratio R.

For $R = 0.5$, the bit line voltage will be approximately equal. If $R \neq 0.5$, the corresponding path resistances to V_{DD} and the ground will be different and thus the bit lines will be discharged to different levels. After the bit lines have been preconditioned, the precharge of the column must be disabled to allow the reading of the CUT with the bit lines precharged below the standard value of V_{DD}. If the CUT is weak, i.e. has inadequate SNM, then reading it with reduced bit line voltages will cause it to flip. By controlling the degree of the bit line discharge we can shift the pass/fail threshold of the test. Both the ratio R and the pulse width $wl_1 - wl_n$ pulse can be digitally reprogrammed to set a new weak cell detection threshold.

Detection of the CUT status after application of the test stress requires a consequent normal read operation after enabling the precharge in the column.

In practice, one is free to use various arrangements to form the ratio R. Ratio R can be formed either by the regular cells from the same column, or by external dedicated cells, or by a combination of the above. To improve the resolution of the RCRT, the number of cells n forming the ratio R can be increased. A larger n will also help to mitigate the effect of the possible read current mismatch among the n cells. To further improve the reliability of the proposed technique, two groups of n cells can be used in each column. In this case, either group of n cells can be used to test the cells in the column. Moreover, one group of n cells can be used to ensure the stability of the cells comprising the other group of n cells.

Higher capacitance of the bit lines will provide higher detection accuracy. Therefore,

the proposed DFT technique may be more attractive for larger SRAM instances with taller columns, which are common in large SRAM caches [65],[19].

6.3.2 RCRT Test Chip Design

For verification and for proof of concept we designed an asynchronous full-custom SRAM test chip with embedded circuitry for conducting the RCRT steps. The test chip has been designed, laid out and fabricated in CMOC 0.18μm TSMC technology. Technology access was provided by the Canadian Microelectronics Corporation (CMC).

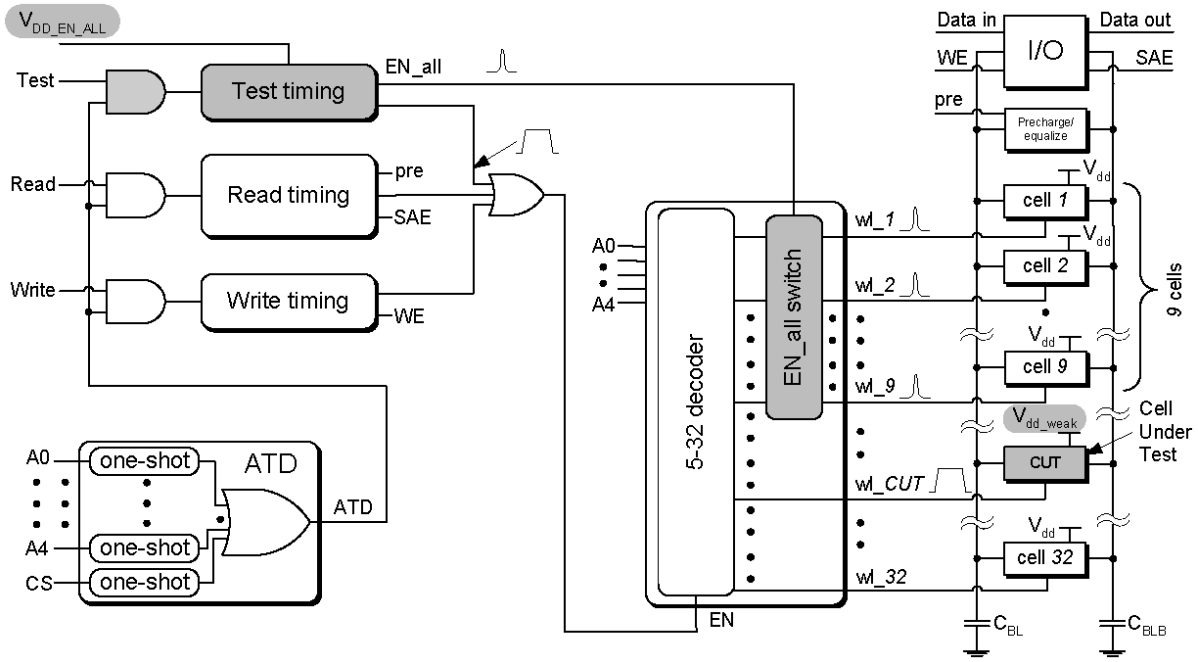


Figure 6.9 Block-level diagram of the test chip containing asynchronous SRAM and RCRT circuitry.

The main blocks of the test chip are presented in Figure 6.9. A write, a read or a test operation is chosen by providing the corresponding signal and the ATD pulse. The ATD block produces a short pulse every time an address (A0-A4) or the chip select (CS) makes

a transition. The timing blocks are designed using a delay timing scheme similar to that presented in Figure 2.15(a) on page 43.

The write timing block provides the write enable (WE) and word line enable (EN) signals such that the input data can be successfully written into the addressed cell. The read timing block stops precharge and issues the EN pulse to the word line of the addressed cell to be read. Once the cell has developed around 300mV on the bit lines the SAE pulse enables the sense amplifier and the bit line differential voltage is amplified. The test timing block provides a pulse controlled by the external voltage $V_{DD\_EN\_ALL}$. In the test mode, this pulse is supplied to the EN_all switch incorporated into the address decoder. The EN_all switch enables the word lines of the nine cells (cell 1–cell 9) that have been used to form the ratio R.

The power supply of the cell under test (CUT) can be varied from the external pin V_{DD_weak}. Reducing V_{DD_weak} enables control the SNM of the CUT. To imitate the bit line capacitance of a column with 256 cells, capacitors $C_{BL} = C_{BLB} = 220fF$ are connected to the corresponding bit lines. The I/O block features separate input and output parts. A latch-type SA is controlled by SAE pulse provided by the read timing block. The write driver is similar to the one shown in Figure 2.11(a).

A microphotograph of the test chip is shown in Figure 6.10. The test chip comprises an asynchronous SRAM with two columns of 32 cells each (1) with extra capacitors (2) connected to each of the bit lines; self-timed blocks to provide read (3), write (4) and test (5) timing; address decoder (6); address transition detector (7); sensing and writing circuitry (8); word line switches (9) and a set of weak cells (10).

To demonstrate the detection capabilities of the proposed technique, we used nine regular ($n = 9$) SRAM cells to form the ratio R. To enable nine word lines simultaneously, our test chip row address decoder was modified to include the EN_all switch (9) on the

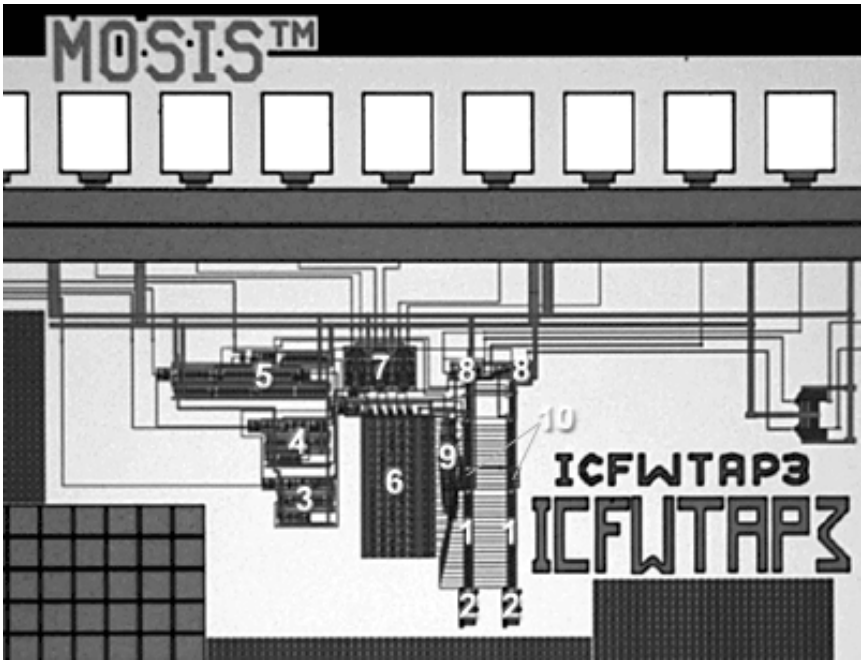


Figure 6.10 Test chip microphotograph.

first nine word lines. In practice, the switching function can be performed by two-input OR gates between the post-decoder and the word line buffers. When activated by a pulse coming from the test timing block, all nine word lines are pulled up simultaneously. The pulse with the required pulse width can be formed locally by a simple one-shot circuit. The width of the pulse activating all nine word lines of the n cells in the ratio R must be short enough, so that the word lines are deactivated before the cells forming ratio R have flipped.

6.3.3 RCRT Detection Capability

Figure 6.11 shows that depending on the number of cells carrying "0" among the nine cells comprising ratio R, the same pulse width of the pulse enabling all nine word lines will

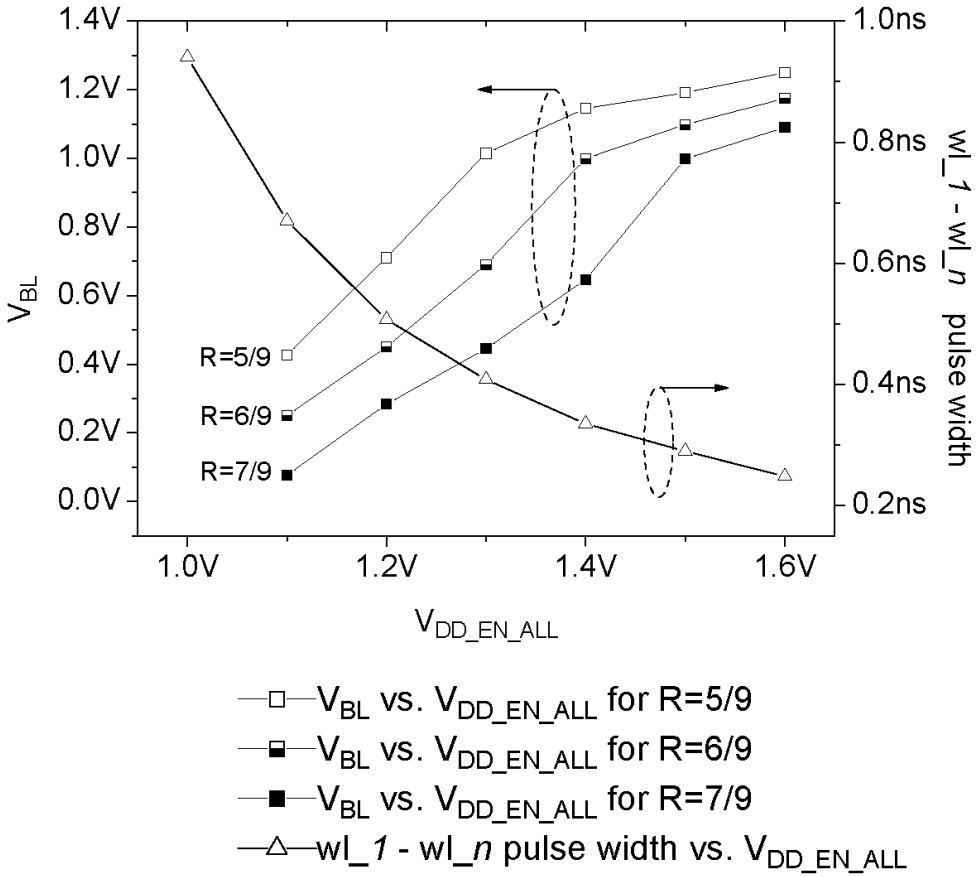
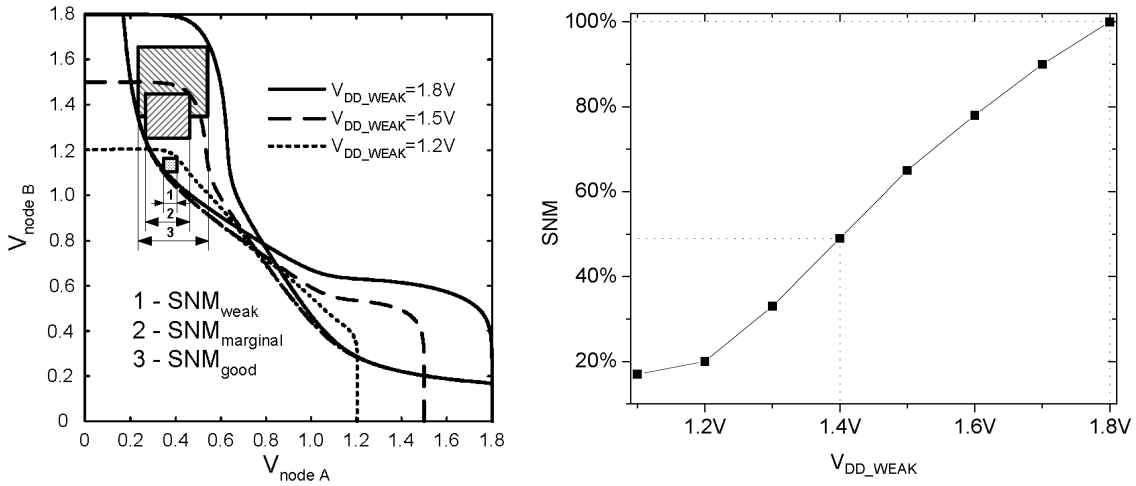


Figure 6.11 **Bit line voltage and the pulse width of the $wl\_1 - wl\_n$ pulse as a function of $V_{DD\_EN\_ALL}$ (post-layout simulation results).**

discharge the corresponding bit line to a greater or a lesser extent. To be able to control the pulse width of this pulse in the test chip, we utilized an external voltage $V_{DD\_EN\_ALL}$, which supplies the delay line in a one-shot circuit and thus modulates the width of the produced pulse. The required pulse width can also be specified and fixed by proper sizing of the delay chain inverters. In this case, adjusting of the detection threshold is done only by changing ratio R.

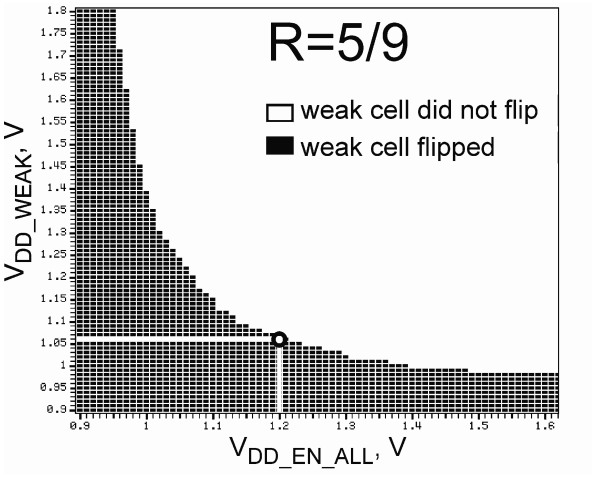
After the bit lines have been discharged to a certain extent, defined by the chosen



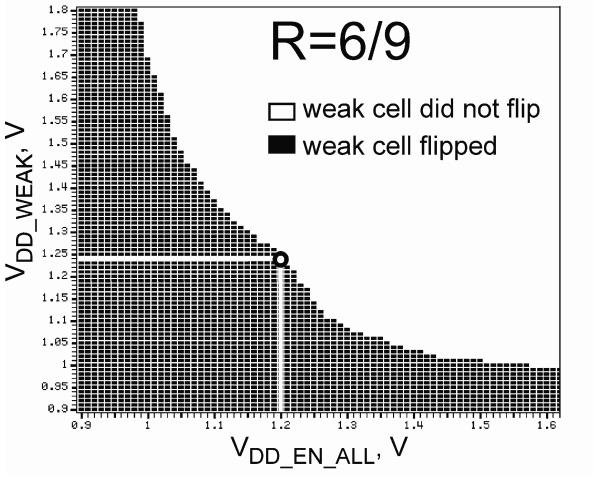
(a) (b)

Figure 6.12 Dependence of the VTC shape (a) and SNM (b) on the cell supply voltage V_{DD_WEAK} in the RCRT test chip.

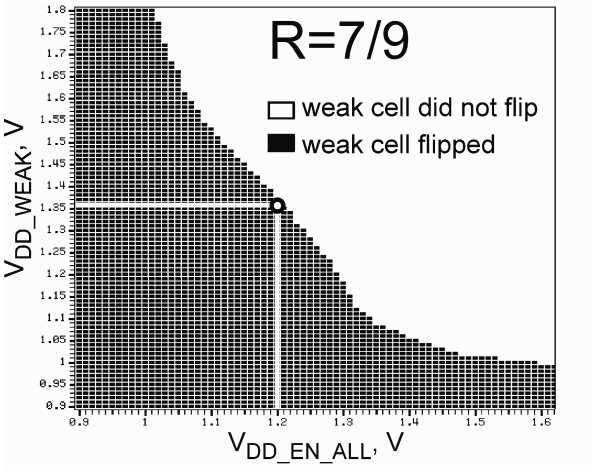
ratio R and the EN_all pulse width, the word line of the cell under test is activated and the reduced voltages of the floating bit lines are applied. From Figure 6.12(a) we can see that a good (defect-free) cell has significantly larger SNM than a weak (defective) cell. Therefore, a weak cell will flip when read-accessed with a lower V_{BL} applied to the node storing a "1" and be detected, whereas a good cell will withstand this stress.



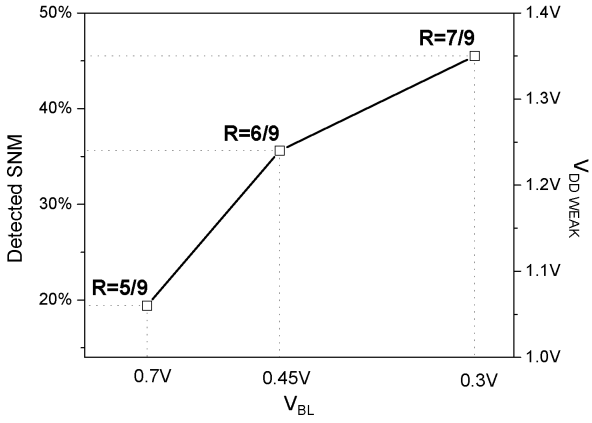
(a) Shmoo plot for $R = 5/9$; a circled dot represents the detection threshold for $V_{DD\_EN\_ALL} = 1.2V$.



(b) Shmoo plot for $R = 6/9$; a circled dot represents the detection threshold for $V_{DD\_EN\_ALL} = 1.2V$.



(c) Shmoo plot for $R = 7/9$; a circled dot represents the detection threshold for $V_{DD\_EN\_ALL} = 1.2V$.



(d) Detection capability of the proposed DFT technique for $V_{DD\_EN\_ALL} = 1.2V$. Programming R from 5/9 to 7/9 changes the weak cells detection threshold from 18% to 46% of the nominal SNM.

Figure 6.13 **Measured Shmoo plots for ratios $R = 5/9$, $R = 6/9$ and $R = 7/9$ (Figure 6.13(a)-6.13(c) respectively) and a summary for $V_{DD\_EN\_ALL}$ fixed at $1.2V$ (Figure 6.13(d)), which corresponds to 500ps pulse width of $wl_1 - wl_n$ pulse (see Figure 6.11).**

To imitate weak cells in this work we used several cells with a separate supply voltage V_{DD_WEAK} (Figure 6.10(10)), which can be adjusted independently from the V_{DD} of the rest of the chip. Figure 6.12(b) on page 142 shows how reducing V_{DD} of an SRAM cell reduces the SNM of the cell and, thus, the cell stability. For instance, to imitate a CUT with 50% of SNM, we need to reduce the power supply voltage V_{DD_weak} of that CUT to approximately 1.4V.

Test stimuli were provided by an Agilent 93000 SOC series tester. For each of the combinations of V_{DD_WEAK} and $V_{DD\_EN\_ALL}$ a predetermined ratio R of "0"s and "1"s was written into the n cells, sweeping V_{DD_WEAK} and $V_{DD\_EN\_ALL}$ from 0.9V to 1.8V. After applying the test sequence shown in Figure 6.7 on page 135 we determined whether the weak cell had flipped.

Figure 6.13 on the preceding page presents the measured Shmoo plots for ratio R of five "0"s and four "1"s (a), six "0"s and three "1"s (b), and for seven "0"s and two "1"s (c) among the nine cells. The black rectangles represent the combinations of V_{DD_WEAK} and $V_{DD\_EN\_ALL}$ at which the CUT flipped and a stability fault has been detected. The white rectangles present the combinations of V_{DD_WEAK} and $V_{DD\_EN\_ALL}$ at which the CUT maintained its data. V_{DD_WEAK} and $V_{DD\_EN\_ALL}$ were swept with a step of 10mV, i.e., each rectangle in the Shmoo plots in Figure 6.13 represents the test response of the RCRT with the resolution of 10mV. Each of the Shmoo plots is combined of 8100 test results, which illustrates the high repeatability of the proposed RCRT.

From analyzing the Shmoo plots it can be seen that the detected degree of cell weakness for every fixed value of $V_{DD\_EN\_ALL}$ will be different depending on the set ratio R. For example, if $V_{DD\_EN\_ALL}$ is fixed at $1.2V$, the pulse width of $wl\_1 - wl\_n$ (EN_all) pulse is set to $\simeq 500ps$. Then, depending on the chosen ratio $R = 5/9$, $6/9$ and $7/9$, the CUT will flip its state after the application of the proposed test sequence at $V_{DD\_WEAK} = 1.06V$,

$1.24V$ and $1.36V$ respectively (white lines and circled dots in Figure 6.13(a)–6.13(c)). The lower values of V_{DD_WEAK} correspond to a detected CUT with smaller noise margin.

Table 6.1 Detection capabilities of the proposed technique

Ratio R	5/9	6/9	7/9
V_{BL} (at $V_{DD\_EN\_ALL}$=1.2V)	0.71V	0.45V	0.28V
Detected V_{DD_WEAK}	1.06V	1.24V	1.36V
Detected SNM corresponding to the detected V_{DD_WEAK} (% of nominal SNM)	18%	36%	46%

The measurement results for this case are summarized in Table 6.1 and in Figure 6.13(d). They show that by programming the ratio R to be 5/9, 6/9 and 7/9 and applying the RCRT test sequence, the pass/fail threshold of the RCRT stability test is programmed from 18% to 46% of the nominal SNM. In other words, by setting EN_all pulse to 500ps and R to 5/9, the RCRT can screen out SRAM cells having less than $\sim$ 20% stability of the typical cell.

The pass/fail threshold can be adjusted by several means. First, by adjusting the pulse width of the EN_all pulse. Second, by choosing a different ratio R. The precision of the pass/fail threshold setting can be modified by changing the ratio R and/or applying the RCRT to an SRAM with more capacitive bit lines.

6.4 Word Line Pulsing Technique for Stability Fault Detection

Another novel DFT technique for stability fault detection in SRAM cells developed in the course of this work uses step-wise reduction of the bit line potential produced by multiple word line pulses of the reference cell [71]. We coined it Word Line Pulsing Technique (WLPT). The following sections will explain the WLPT in detail.

6.4.1 WLPT Concept

Let us consider the circuit in Figure 6.14 representing a section of a column in an SRAM array with two identical cells. We will call the top cell the reference cell and the bottom cell – the Cell Under Test (CUT).

Suppose node A of the CUT carries a "0", node B carries a "1" and resistors $R1$ or $R3$ represent opens in the pull-up path. It is well known that the hard opens in the pull-up path of an SRAM cell can cause Data Retention Faults if the pull-up current through $R1 - Q4 - R3$ in Figure 6.14 fails to compensate for the off-state leakage of the driver transistor $Q2$. Given sufficient time, the leakage current of $Q2$ will discharge the capacitance of node B. Once V_{node_B} has crossed the switching threshold of the cell, the cell will flip and can be detected by a consecutive read operation. This situation is similar to a very slow weak overwriting of the cell. For higher values of the $I_{pull-up\_path}/I_{leakage\_driver}$ ratio, the time required to discharge node B and flip the CUT can be longer than is economical for the Data Retention Test (DRT) to detect such a defect. Moreover, if the pull-up current is even marginally greater than the off-state leakage current of $Q2$, the cell may escape the DRT even with extended test delay periods and elevated temperature. Memories with such highly unstable cells may be shipped out and cause customer returns.

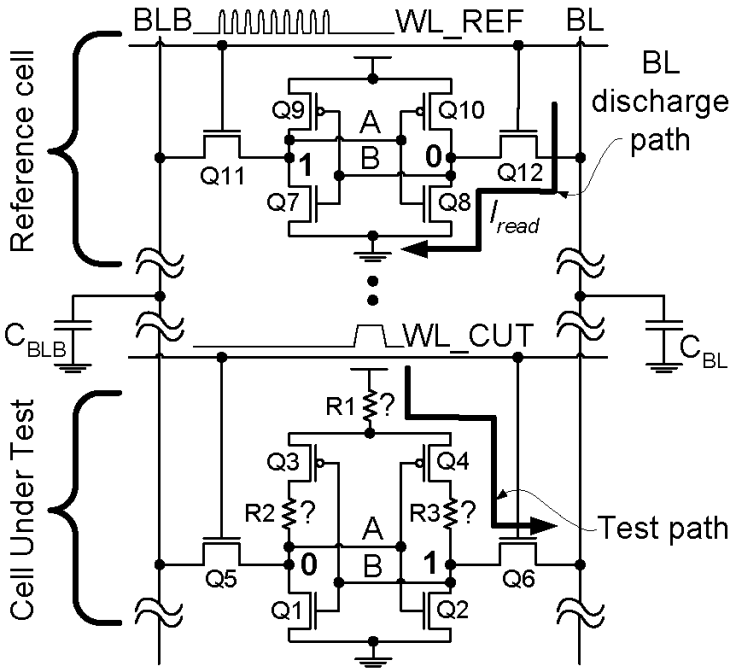


Figure 6.14 Reference cell and Cell Under Test (CUT). R1 represents a symmetric and R2, R3 represent asymmetric defects.

The gate voltage of $V_{G\_Q1} = V_{node\_B}$ will not change significantly unless $I_{leak\_Q2}(R1+R3)$ is significant, i.e. at least one of the weak opens $R1$ or $R3$ qualifies for a hard open. So, node A and the gate of $Q4$ will remain at "0" and the effective pull-up current for node B of the CUT will be proportional to the equivalent resistance of the path $(R1 - Q4 - R3)$.

Suppose we have the means to freely change the potential on the bit line (BL) and set it to be $0.6V$ while the complementary bit line (BLB) remains fully precharged to the supply voltage. After enabling WL_CUT, $Q6$ will pass the reduced V_{BL} onto node B. The node B potential is proportional to the ratio I_{Q4}/I_{Q6}. The overwrite condition for node B is ensured if we can pull node B below the switching threshold of the inverter formed by transistors $Q1$ and $Q3$. Since the effective pull-up drive of node B is weakened by the defect resistance $(R1 + R3)$, the overwrite condition is met earlier and the weak cell is

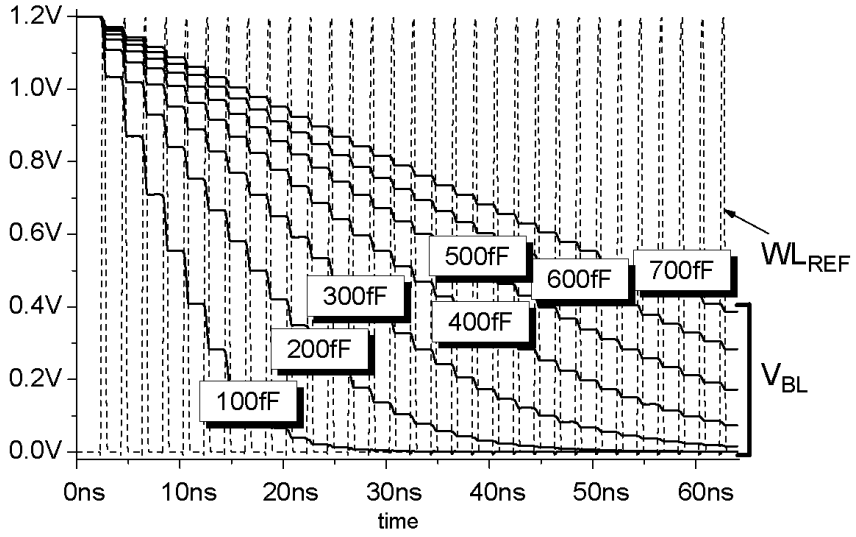


Figure 6.15 Word line pulses of the reference cell discharge the bit line for various values of the bit line capacitance. Results obtained for CMOS 0.13μm technology and the pulse width of the reference cell word line of 410ps.

overwritten, whereas a good cell with $(R1 + R3) \rightarrow 0$ can withstand the same stress.

The WLPT is based on the realization that the precharged bit line BL coupled through the access transistor $Q12$ to the node B of the reference cell carrying a "0" (Figure 6.14) is gradually discharged by the I_{read} of the reference cell. The discharge rate can be expressed by Eq.6.1 and is a function of the *total duration* that the word line of the reference cell has been enabled.

$$\Delta V_{BL} = \frac{I_{read} * t_{WL\_REF\_pw}}{C_{BL}} \tag{6.1}$$

where: ΔV_{BL} is the discharge of the bit line after each enabling of the WL_REF; I_{read} is the cell read current of the reference cell; $t_{WL\_REF\_pw}$ is the pulse width of the reference cell word line pulse; C_{BL} is the bit line capacitance.

Figure 6.15 shows the waveforms of the bit line voltage discharge after the application of each of the 32 WL_REF pulses. Since ΔV_{BL} is inversely proportional to C_{BL}, i.e. the bit

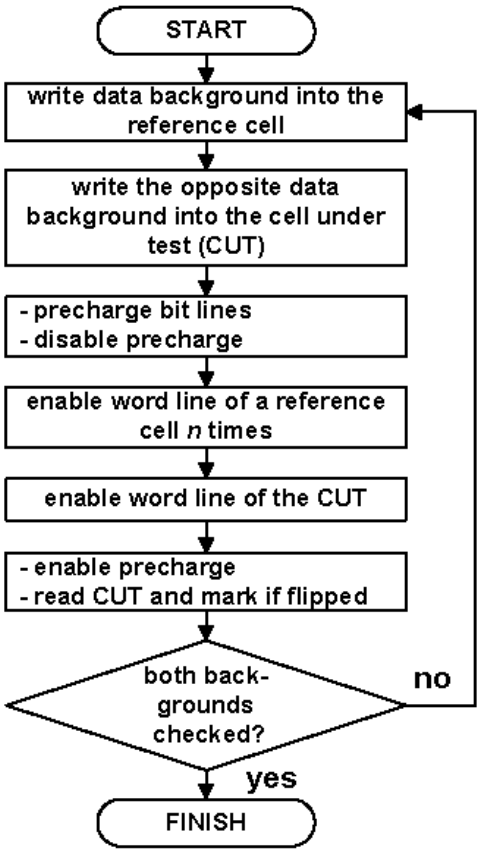


Figure 6.16 Flow diagram of the proposed Word Line Pulsing technique for stability fault detection in SRAM cells.

line discharge rate is slower and the precision of the bit line voltage setting is higher for the higher values of the bit line capacitance. The bit line capacitance in SRAM architectures with local and global bit lines can be increased in the test mode by enabling all of the bit line MUX transistors that connect the local bit lines to the global in a given column.

The flow diagram of the proposed weak cell detection technique is shown in Figure 6.16. The test procedure consists of writing the opposite data backgrounds into the reference cell and the CUT. Next, after the normal precharge we enable the word line of the reference cell N times gradually discharging the bit line as shown in Figure 6.15. Then, we enable

the word line of the CUT and after that perform a normal read operation to determine whether the CUT has flipped. To ensure the coverage of asymmetric faults as well, we invert the data background stored in the reference cell and the CUT and repeat the test sequence. To further reduce the test time, this test can be conducted in parallel on a word line per word line basis with one reference cell and one CUT per column.

Implementation Variations

Since the rate of the bit line discharge by the reference cell is determined by the total duration that the word line of the reference cell has been enabled, the required degree of the bit line discharge can be achieved by several methods.

The first method is illustrated in Figure 6.15. The total enabled duration of WL_REF can be composed of N pulses. The desired resolution of the bit line discharge can be provided by changing the number of the WL_REF pulses.

The second possible approach is to fix the number of WL_REF pulses N while changing their pulse width.

The third approach is to hold WL_REF for a predefined time period ($N = 1$) to discharge the bit line to the required level. The reference cell can be interchanged with the CUT or another cell sharing the same bit lines with the CUT. Thus, the SNM of all the cells in a given column can be tested. Inter- and intra-cell read current spread of the reference cells caused by the process variations can be alleviated by using a larger dedicated reference cell. Another degree of freedom can be provided by employing several dedicated reference cells with different read currents. In other words, the WLPT offers extended flexibility in setting the weak overwrite stress superior to the prior art.

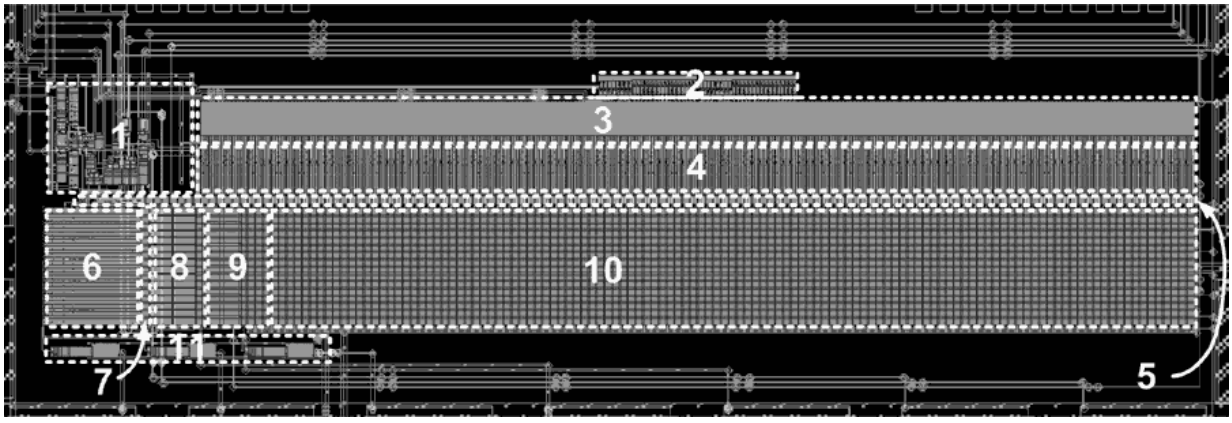


Figure 6.17 **8Kb synchronous SRAM test chip with WLPT (IBM CMOS 0.13μm 8 metal process): (1) control block, (2) decoder, (3) post-decoder, (4) word line drivers, (5) dummy column and dummy SA, (6) SAs, column MUXs, write drivers and precharge/equalization, (7) dummy row, (8) reference SRAM cells for WLPT, (9) weak SRAM CUTs for WLPT, (10) regular SRAM cells and (11) pad drivers.**

6.4.2 WLPT Test Chip Design

A full-custom synchronous self-timed 8Kb 300MHz SRAM test chip in IBM CMOS 0.13μm process was designed and taped out to prove the concept of the WLPT. SRAM circuit occupies $150\mu m * 672\mu m$. The top-level layout and and the block-level diagram of the test chip are presented in Figure 6.17 and Figure 6.18, respectively.

The SRAM array is composed of 32 columns and 256 rows in a 2 words by 16 bits architecture. The test chip features a fully self-timed control block. The read access time and power is minimized by using the replica (dummy) loop based timing technique. The Control block contains a reset-dominated Finite State Machine (FSM). When it is set by the rising edge of the clock (CLK), it disables precharge and enables the Dummy Word Line (DWL) and a regular word line. Once the single-ended Dummy Sense Amplifier (DSA) flips, it resets the FSM stopping further discharge of the bit line. Next, the Sense Amplifier

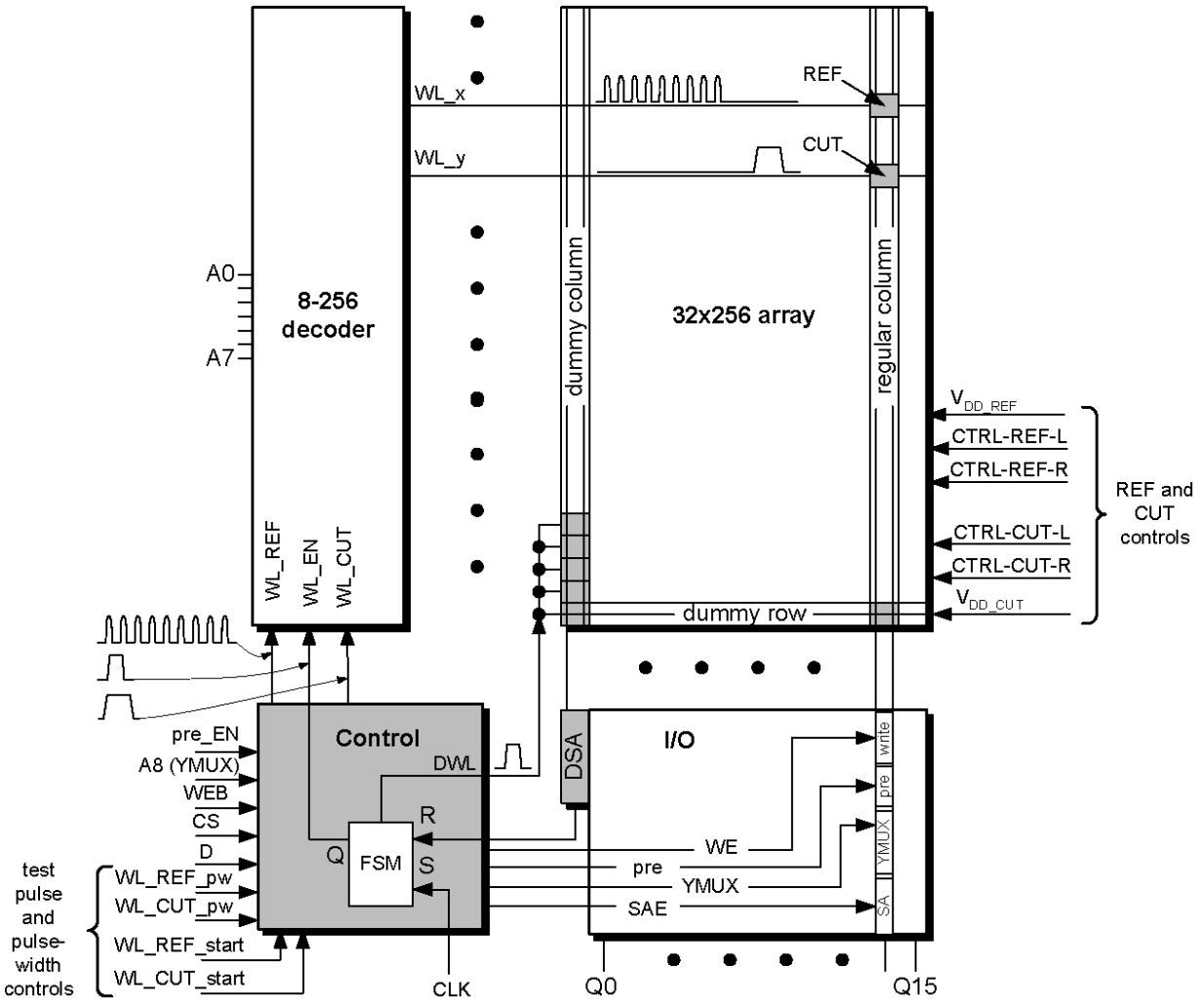


Figure 6.18 Block-level diagram of the WLPT test chip.

Enable (SAE) is issued to latch the V_{diff} on the bit lines. The YMUX is disabled shortly after to prevent the bit lines from being discharged by the SA. More details on the principle of the replica-loop based timing is presented in Section 6.4.2.

The Chip Select (CS) input is latched on the rising clock edge ensuring that the subsequent read or write operation is applied to this chip. A high level on Write Enable Bar

(WEB) enables activation of the SA and YMUX and disables the write drivers providing for a successful read operation. Conversely, a low level of the WEB switches SRAM into the write mode by blocking the SAE and YMUX and enables the write drivers.

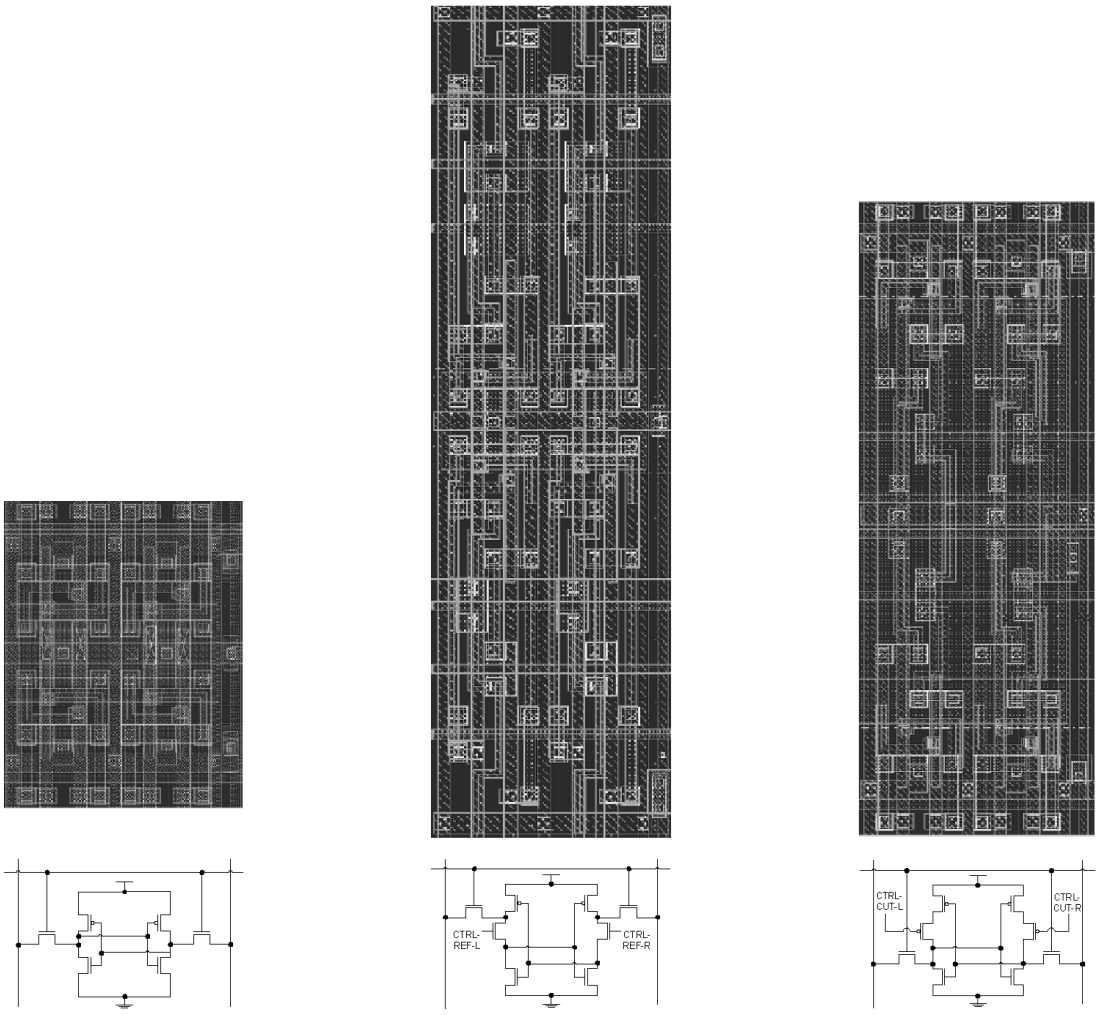
Address bit A8 (YMUX) provides an address for selection of the column to be connected to the SA by YMUX. Combined with the row address bits A0–A7, that makes for nine address bits overall.

Signals WL_REF_start and WL_CUT_start can be connected to the system CLK to initiate each of the WL_REF or WL_CUT pulses for the reference cell or the CUT, respectively. For added flexibility, the WL_REF or WL_CUT pulse width is adjustable via WL_REF_pw or WL_CUT_pw levels. These levels control the PMOS bias in the current-starved inverters used in WL_REF or WL_CUT one-shot circuits. The formed WL_REF or WL_CUT pulses gate the accessed word line. An example of a REF cell and the CUT is shown in Figure 6.18.

Cells

Figure 6.19 on the next page shows the cells used in the test chip in blocks of 2x2 cells with their corresponding schematics. The cell area of a regular six-transistor SRAM cell laid out in the general logic $0.13\mu m$ process using the minimal design rules is $3.15\mu m^2$. Adjacent cell in the column are mirrored to share a V_{DD} and the bit line contacts achieving higher packing density.

Besides regular cells, shown in Figure 6.19(a), two kinds of special cells were introduced. The first is the reference cells shown in Figure 6.19(b). The reference cells include extra NMOS transistors in series with the driver transistors in the pull-down path controlled by CTRL-REF-L and CTRL-REF-R for the left and right part of the cell, respectively. The function of these extra transistors is to control I_{read} of the reference cell. Controlling the



(a) (b) (c)

Figure 6.19 Blocks of 2x2 SRAM cells and the corresponding cell schematics used in WLPT test chip: (a) regular cells; (b) reference cells; (c) cells under test.

I_{read} allows finer discharge steps of the bit line voltage and facilitates test chip characterization. The I_{read} of another type of a reference cell (not shown in Figure 6.19) used in the test chip is controlled by a variable supply voltage V_{DD_REF}. These reference cells are

otherwise identical to the ones shown in Figure 6.19(a).

The second special cell is the Cell Under Test (CUT) shown in Figure 6.19(c). The pull-up path of the CUT can be weakened by controlling the gate voltages of the extra PMOS transistors in series with the regular PMOS transistors (CTRL-CUT-L and CTRL-CUT-R for the left and right part of the cell). Similarly to the reference cells, another kind of CUT with a variable supply voltage V_{DD_CUT} (not shown in Figure 6.19) is used to extend flexibility during the test.

Note that the larger reference cells and the CUT described above will not be present in a real embedded SRAM instance. Their only purpose is to facilitate the testing of the test chip and provide proof of concept.

Figure 6.20 presents a zoomed-in view on the central part of the chip. It shows the relative location of the reference cells and the CUTs as well as the dummy column and the dummy row discussed in Section 6.4.2. Reference cells-1 and reference cells-2 correspond to the cells with an extra NMOS transistors and with a variable supply respectively. Similarly, CUT-1 and CUT-2 correspond to the cells with extra PMOS transistors in the pull-up path and with a variable supply respectively.

The WLPT introduces one minimal-size NOR gate into the word line drivers shown as WLPT_EN in Figure 6.20. The area overhead is minimal and is estimated to be 1.3% of the total SRAM area in Figure 6.17. Generally, area overhead of a DFT or a BIST is often decreasing with the area increase of the tested SRAM array.

Replica-loop based timing

Timing control circuitry used in the test chip comprises a replica (a.k.a. dummy) column, replica (a.k.a. dummy) row and an inverter acting as a dummy SA. Employing replica timing provides the fastest and the most power-efficient operation. It can be achieved if

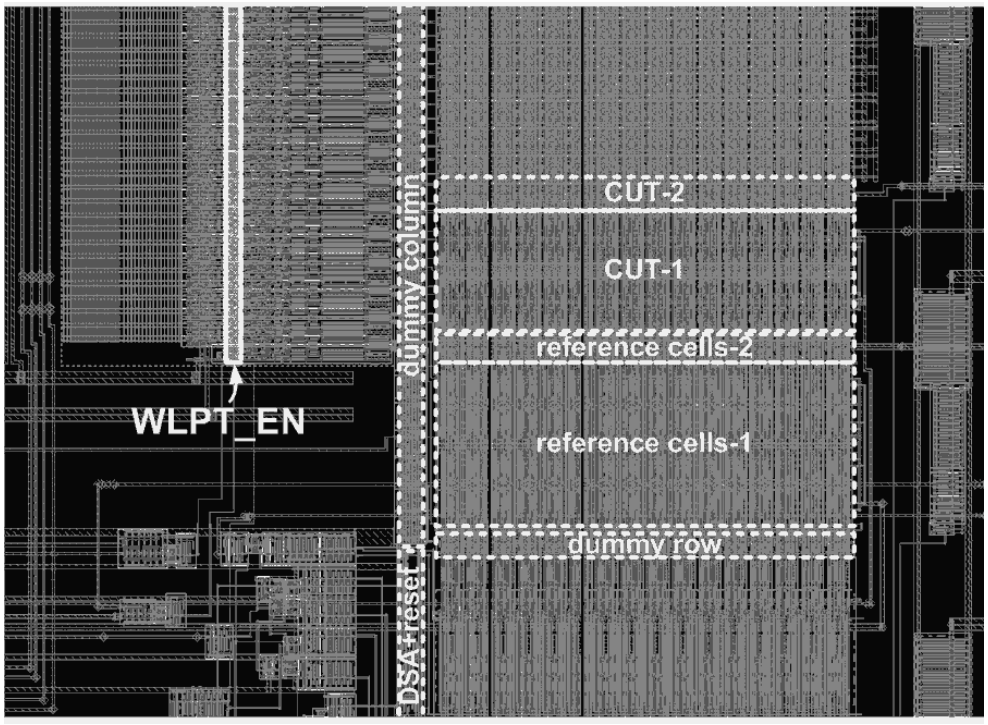


Figure 6.20 **WLPT offers a low area overhead of $\sim 1.3\%$ adding just one minimal-sized NOR gate to each word line driver (WLPT_EN shown shaded in a solid-line rectangle).**

the time T_{dummy} to discharge the replica bit line to the switching threshold V_{dummy} of the dummy SA equals to the time T_{diff} a regular SRAM cell needs to discharge the active bit line to achieve sufficient differential voltage V_{diff} for reliable sensing by the bit SA. In other words,

$$T_{diff} = \frac{C_{BL\_active} V_{diff}}{I_{read}} \tag{6.2}$$

should be equal to

$$T_{dummy} = \frac{C_{BL\_dummy} V_{dummy}}{I_{dummy}}. \tag{6.3}$$

The bit line capacitance of the dummy column is a replica of a regular column. Combining Equations 6.2 and 6.3 and assuming that $T_{dummy} = T_{diff}$ and $C_{BL\_active} = C_{BL\_dummy} = C_{BL}$

gives:

$$\frac{C_{BL}V_{diff}}{I_{read}} = \frac{C_{BL}V_{dummy}}{I_{dummy}} \tag{6.4}$$

I_{dummy} can be expressed as:

$$I_{dummy} = \left(\frac{V_{dummy}}{V_{diff}}\right) I_{read} \tag{6.5}$$

Since the dummy SA is single-ended (inverter), its switching threshold differs from the one of the active latch-based differential SA. The switching threshold V_{dummy} of the dummy SA is chosen to be $V_{DD}/2$, whereas V_{diff} for reliable sensing is assumed to be at least $V_{DD}/10$. Substituting V_{dummy} and V_{diff} in Equation 6.5 by their values gives:

$$I_{dummy} = \left(\frac{V_{DD}/2}{V_{DD}/10}\right) I_{read} = 5I_{read} \tag{6.6}$$

In other words, in order to satisfy $T_{dummy} = T_{diff}$, the dummy bit line has to be discharged with the current which is equivalent to five cell read currents. This condition is satisfied if the dummy bit line is discharged by simultaneous activation of five SRAM cells connected to it.

Once the dummy SA switches, it resets the FSM in the control block. The reset signal disables the word line of the accessed row. The regular cells in the accessed row stop discharging the corresponding bit lines. Since $C_{BL\_active} = C_{BL\_dummy}$ and the propagation paths of the dummy loop and the regular access path are matched, V_{diff} seen by the active SA should be sufficient for reliable sensing. The SAE signal is issued and the SA amplifies the applied V_{diff}, latches the data and passes it to the global read bus.

The matched delay of the dummy and active path are achieved by replicating the column and row propagating paths. The dummy column and a dummy row (Figure 6.20) replicate the bit line and the word line capacitance of an active column and an active row respectively. The dummy word line activation delay is matched to the longest access delay

of SRAM array. That ensures that the most remote memory location has sufficient time to discharge its bit line for reliable sensing. Moreover, the replica self-timing technique is shown to provide more reliable tracking of process variations [72] allowing high-speed SRAMs to be designed with tighter timing margins [73].

Decoders

Row decoding of the 256 row x 32 column array is done by multi-stage decoders similar to the one shown in Figure 2.12(b) on page 38. Two 4-16 decoders (block (2) shown in Figure 6.17 on page 151) provide 32 partially decoded address lines to the post-decoder(block (3)). The post-decoder consists of NAND gates decoding the partial products into a unique row address.

Column decoding in the 2 words by 16 bits architecture is provided by a simple 1-2 MUX switching the SA between two neighboring columns.

Sense amplifier, precharge/equalize and write driver

Latch-type sense amplifiers (SAs) and precharge/equalization circuitry used in the test chip are similar to shown in Figure 2.10 on page 35. Write drivers similar to shown in Figure 2.11(c) on page 36 writes the addressed cell by driving its complementary bit line to the ground.

6.4.3 WLPT Detection Capability

This section demonstrates the weak cell detection capabilities of the WLPT.

The WLPT has been verified for various values of bit line capacitance, number and pulse width of the WL_REF pulses. It proved to be capable of detecting a wide range of

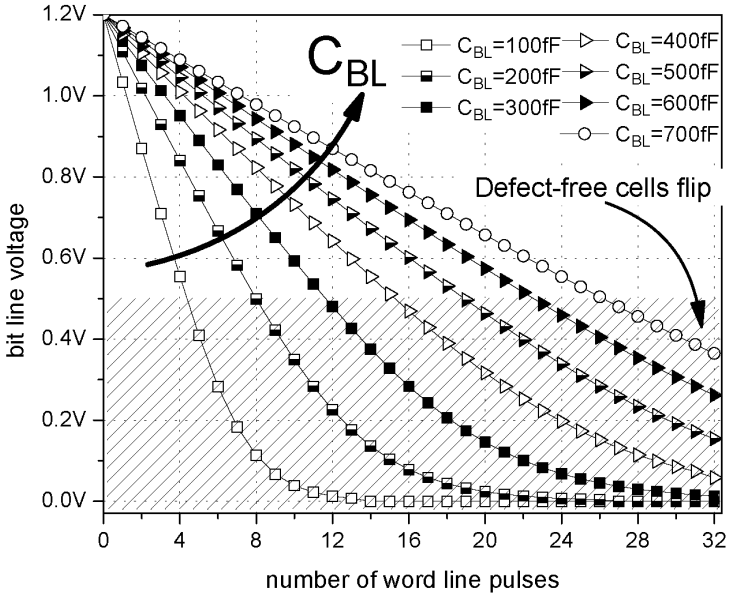


Figure 6.21 Discharge of the bit line as a function of the number of word line pulses of the reference cell and the bit line capacitance.

resistive defects. To demonstrate the detection capabilities of the WLPT we used twelve WL_REF pulses with a pulse width of $410ps$, $C_{BL} = 400fF$.

It was found that if the overwrite stress is too strong, i.e. the bit line is discharged below a certain point, which in our simulations was 0.55V, even the defect-free cells will flip. The area where the defect-free cells flip is represented in Figure 6.21 by the patterned rectangle. As was mentioned before, the rate of the bit line discharge is a function of the bit line capacitance. Depending on bit line capacitance values, a different number of WL_REF pulses is required to reach the desired degree of the bit line discharge.

Figure 6.22 shows the signal waveforms illustrating detection of a symmetric defect (resistive contact "1" in Figure 4.3 and $R1$ in Figure 6.14). After WL_REF has been enabled twelve times, the bit line has been discharged to 650mV. Figure 6.22(a) demonstrates that for $R1 = 80k\Omega$ the potentials of nodes A and B of the cell have not reached

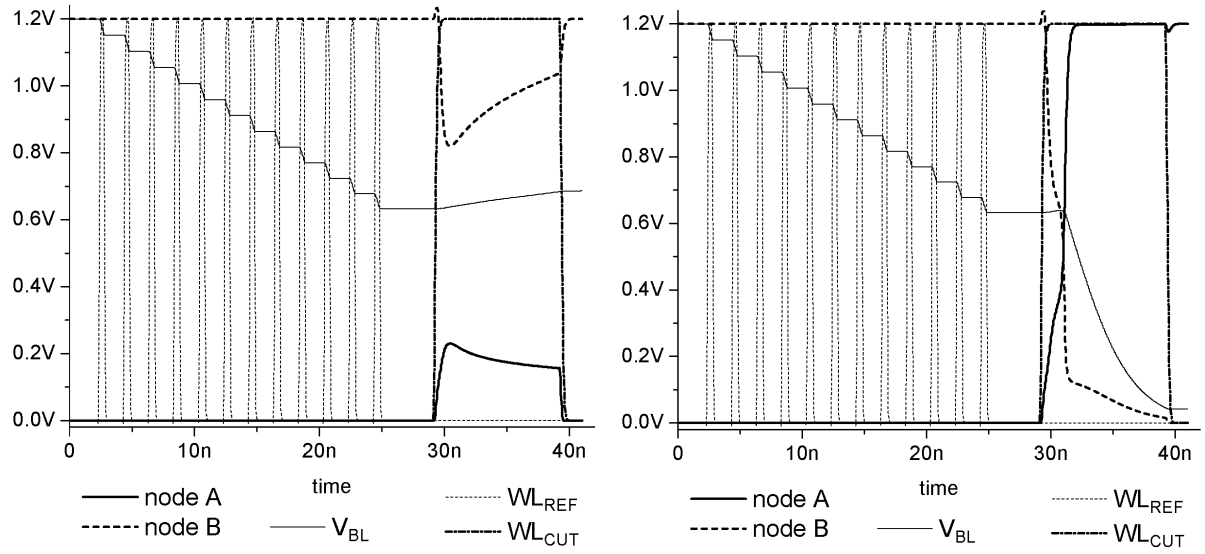


Figure 6.22 Detection of a symmetric defect in the pull-up path of an SRAM cell. A symmetric defect with $R1 = 80k\Omega$ is not detected (a), whereas if $R1 = 120k\Omega$, it is detected (b) (CMOS 0.13μm, $C_{BL} = 400fF$).

the metastable point and a defect with $80k\Omega$ resistance has not been detected. However, for the defect resistance of $120k\Omega$ (Figure 6.22(b)), node B has been driven low enough to cross the switching threshold of the inverter formed by transistors $Q1$ and $Q3$ and the CUT has flipped its state. Consequently, when the CUT is read back after the application of the test procedure, the cell with $R1 = 120k\Omega$ will be marked as defective.

Note that the slope of the cell's VTC in the metastability region dV_{out}/dV_{in}, which is proportional to the AC gain of the cell, is steeper for the cell with the higher SNM (Figure 4.10). The resolving capability (gain-bandwidth product) of the cell in the metastable region is proportional to the SNM [29]. Therefore, if the SNM of the cell is higher, the cell will have a higher immunity against metastability and quickly recover from the test disturbance. However, if the stability of the cell is weakened by a defect or a mismatch

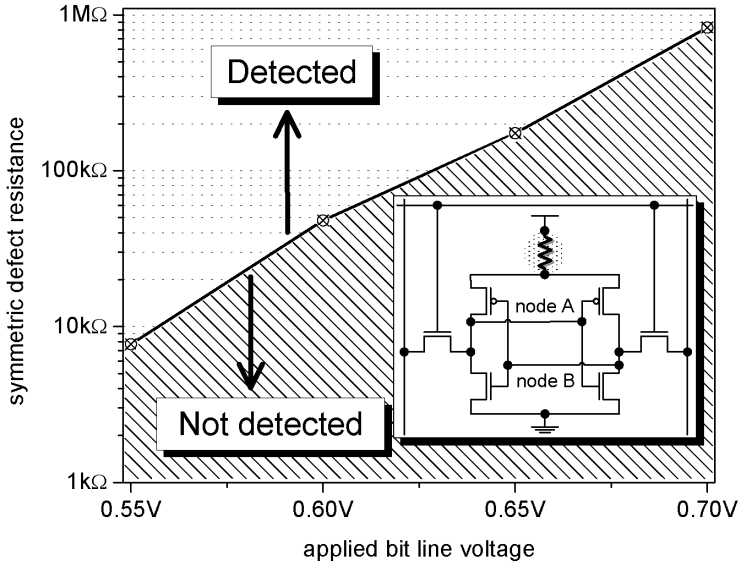


Figure 6.23 WLPT detection of a symmetric defect resistance in the pull-up path of an SRAM cell as a function of applied bit line voltage (CMOS 0.13μm, $C_{BL} = 400fF$).

and the SNM is reduced, such a cell will stay in the metastability region longer. Thus, to extend the range of the detected defect resistance, the duration of the WL_CUT pulse should be sufficient to allow for the extended metastability window of the weaker cells before a stable state is resumed.

For instance, if the value of $R1$ is between $80k\Omega$ and $120k\Omega$, e.g. $95k\Omega$, the potentials of node A and node B move closer to each other and to the metastable point of the cell. Due to the reduced AC gain in this region, the WL_REF pulse should be asserted for several nanoseconds for such a cell to reach a stable state.

Waveforms very similar to those presented in Figure 6.22 have been obtained for an asymmetric defect $R3$ with $R3_{detected} = 180k\Omega$ and $R3_{undetected} = 190k\Omega$. The WLPT has also been verified to successfully detect hard opens in the gates of PMOS transistors.

The WLPT has shown excellent detection capability. Figures 6.23,6.24 and 6.25 sum-

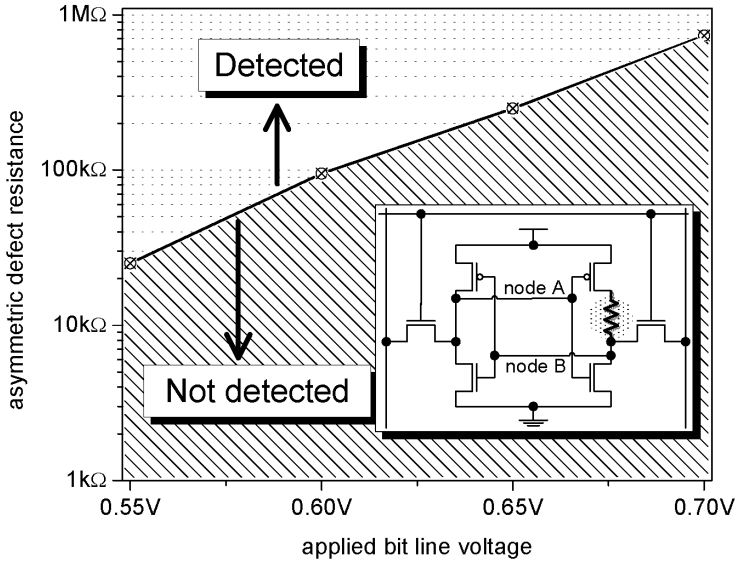


Figure 6.24 WLPT detection of a asymmetric defect resistance in the pull-up path of an SRAM cell as a function of applied bit line voltage (CMOS 0.13μm, $C_{BL} = 400fF$).

marize our findings of the WLPT detection range for $C_{BL} = 400fF$. Symmetric and asymmetric defects are the typical cause of data retention and stability faults. The WLPT demonstrated high selectivity of the detected defect resistances that can cause such faults. The solid line represents the detection boundary, where the area above is the detected values of the defect resistance and the patterned area represents the resistance values beyond the WLPT detection capability. Figures 6.23 and 6.24 demonstrate that the WLPT can detect a weak open defect with resistance as small as $10k\Omega$, which is not far from the normal path resistance of the pull-up path in an SRAM cell.

Figure 6.25 on the following page demonstrates the detection of the bridge resistance between node A and node B as per the weak cell fault model presented in Section 4.1.2. According to our estimations, a resistive bridge between node A and node B is one of the more probable bridge defects in the cell layout under consideration. Since it also represents

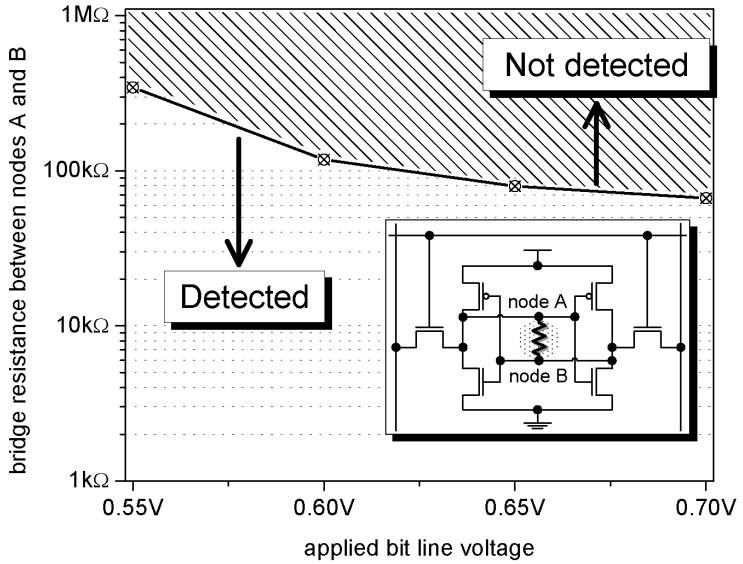


Figure 6.25 Detection of a resistive bridge between node A and node B (as per the proposed stability fault model introduced in Section 4.1.2) as a function of applied bit line voltage (CMOS 0.13μm, $C_{BL} = 400fF$).

a symmetric negative feedback branch for the two cell's inverters, its resistance directly reduces the AC gain of the cell in the metastable region, it changes the shape of the VTCs and, thus, reduces the SNM of the cell. We believe that it can mimic the behavior of a multitude of resistive bridges in an SRAM cell. Since the SNM dependence on the bridge resistance is the inverse of that of the resistive opens, the solid line in Figure 6.25 also has the inverse slope compared to Figures 6.23 and 6.24. The higher bridge resistance values correspond to a more stable cell and thus the detected resistance region in Figure 6.25 is above the solid boundary line.

A comparison between detected defect resistance in the pull-up path of an SRAM cell between the Data Retention Test (DRT) vs. the WLPT is presented in Figure 6.26. As apparent from Figure 6.26, the WLPT exhibits an open defect detection capability exceeding the one of the traditional Data Retention Test by nearly four orders of magnitude.

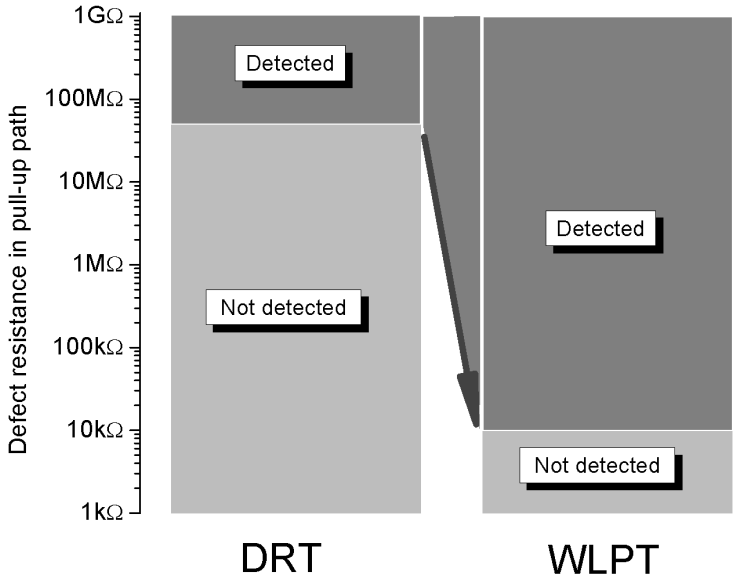


Figure 6.26 Detection range of defect resistance in the pull-up path of an SRAM cell: Data Retention Test (DRT) vs. the WLPT.

6.5 Comparison of the Proposed DFT Techniques and the DRT

Table 6.2 presents comparative analysis on the proposed DFT techniques and the DRT with respect to their fault coverage, test granularity, test time and the area overhead.

Table 6.2 Comparison of the proposed DFT techniques and the DRT.

	fault coverage	granularity	test time	area overhead
RCRPT	SF+DRF	$r/2$	$t_{cycle} * 2N(r+2)$	1 NOR gate/row, 1 NFET/column
RCRT	SF+DRF	$r/2$	$t_{cycle} * 2N(r+2)$	1 NOR gate/row
WLPT	SF+DRF	z	$t_{cycle} * 2N(z+3)$	1 NOR gate/row
DRT	DRF	1	$t_{cycle} * 4N + 2t_{delay}$	none

Table 6.2 uses the following notation: r is the number of cells in the same column which are used to form the ratio R; z is the number of pulses of WL_REF signal which is necessary to achieve the target bit line voltage; N is the number of rows in the array and t_{cycle} is the access cycle time; t_{delay} is the delay (pause) time in the DRT. The delay time is typically around 100ms/data node [7], which results in a delay of around 200ms required to test both the data nodes of an SRAM cell. This is the dominating component of the DRT test time.

All of the proposed DFT techniques cover both SFs and DRFs whereas the DRT will only cover the DRFs. DRT is based on the leakage of the driver transistor being larger than the pull-up current. All other defects that also can cause a cell to have a reduced SNM and be unstable will not be detected by the DRT.

The number of programmable steps, or the granularity of the test for the proposed DFT techniques, is to a large extent a design variable. Depending on the target degree of required precision in setting the pass/fail threshold of the test, a designer can modify the granularity of the chosen DFT technique by varying the number of cells in the same column r which are used to form the ratio R (RCRPT and RCRT) or by changing the pulse width or redesigning the reference cell for a smaller read current (WLPT). The WLPT provides a designer with more flexibility in choosing the ways to control the test granularity compared to the RCRPT or RCRT.

The test time of the proposed DFT techniques is similar for the most practical cases. In Table 6.2 the test time is presented per memory word. Table 6.3 compares the test time per word for a sample SRAM array with 128 rows and access cycle time of 3ns (333MHz). From Table 6.3 it is obvious that the proposed DFT techniques offer superior test time to that of the DRT, while providing better defect coverage and flexibility. The major test time component that provides the advantage of the DFT techniques over the DRT is the

absence of the delay time required by the DRT. The test time of the DRT is offset by $2t_{delay}$ which is typically in the range of several hundreds of milliseconds. The write/pulse operations necessary to create the reduced bit line voltage in the proposed DFTs and the test application can be executed in parallel over all the words/blocks of the SRAM under test. The read operations necessary to determine whether the CUTs have flipped have to be done sequentially on the word-by-word basis due to the limited number of I/Os. Hence, the test time of the large SRAM instances will be limited by the time required to read out the test results from every block of the SRAM under test. However, the DRT executed on large SRAM instances will also exhibit increased test time due to the time required to read out the test results. The reduction in test time can offered by the proposed DFTs can significantly reduce the total test cost thanks to the higher tester throughput.

Table 6.3 Test time comparison example, $t_{cycle} = 3ns$.

test technique	test time
RCRPT, $r = 9$	$\simeq 8.5\mu s$
RCRT, $r = 9$	$\simeq 8.5\mu s$
WLPT, $z = 12$	$\simeq 11.5\mu s$
DRT, $t_{delay} = 100ms$	$\simeq 200ms$

All of the proposed techniques introduce minimal area overhead. The RCRPT will require an extra NOR gate per row and an extra minimum-size NMOS transistor per column. The RCRT and the WLPT reduce the area overhead to one extra NOR gate. The relative impact of the area overhead decreases as the array size is growing. For example, the WLPT area overhead for a 256x32 array is estimated to be around 1.2%. Doubling of the array size is (256x64) will reduce the array area overhead by half (0.6%).

The speed impact of the proposed DFTs is approximately 40ps of extra delay introduced

by the converting an inverter into a NOR gate. For an SRAM instance running at 300MHz (3.3ns cycle time), the cycle time impact will be around 1.2%.

6.6 Summary

In this chapter, we introduced three novel Design For Testability (DFT) techniques for SRAM cell stability testing. The test stress to the CUT in Read Current Ratio with a Pass Transistor (RCRPT) and the Read Current Ratio Technique with floating bit lines (RCRT) techniques is formed by the read current ratio of a number of cells. The Word Line Pulsing Technique (WLPT) forms the test stress by pulsing the word line of the reference cell with subsequent application of modified bit line potentials to the CUT.

All of the proposed techniques offer a flexible digitally programmable pass/fail threshold and are proven to be effective in detecting Stability and Data Retention Faults. They exceed the capabilities of known SRAM cell stability test techniques with regard to flexibility in setting (programming) the test stress while also providing excellent defect resistance coverage.

Two test chips have been designed to provide a proof of concept for the proposed DFT techniques. The RCRT has been verified in a full-custom asynchronous self-timed SRAM test chip fabricated in CMOS 0.18μm technology. The measurement results presented in Figure 6.13 on page 143 show the effectiveness of the RCRT in detecting SRAM cells with compromised stability. The WLPT is implemented in a full-custom synchronous self-timed SRAM test chip designed in CMOS 0.13μm technology. The detection capability of the WLPT is shown to exceed one of the traditional Data Retention Test by nearly four orders of magnitude.

Chapter 7

Conclusion

7.1 Summary

In this thesis we discussed the challenges of the designing and testing of deep submicron embedded SRAMs focusing on cell stability issues. Based on the results of the investigation into the factors affecting cell stability, we developed insight into weak cell detection principles. This enabled us to develop three novel design for testability techniques that exceed the capabilities of the Data Retention Test in the detection of stability faults. We will briefly summarize the main findings and contributions of our work and outline some of the possible future work in this area.

7.2 Stability Characterization and Detection

The high bit count and packing density of embedded SRAMs makes them yield limiters in SoCs. Large defect-sensitive SRAM arrays exhibit a growing number of unstable cells. We established that process variations, such as V_{TH} offset and mismatch, photo-lithography

non-idealities causing L_{EFF} and W_{EFF} variations, can severely deteriorate the Static Noise Margin of an SRAM cell. The additional impact of subtle defects and extreme operating conditions can cause many cells in an SRAM array to have marginal stability and inadvertently flip their state.

Data Retention Fault modelling using extra resistors in the pull-up path of the cells has been supplemented by the proposed novel Stability Fault Model. The proposed SF model mimics the impact of various stability deteriorating factors on the SNM of an SRAM cell. It has been validated by the proposed DFT techniques.

Analytical prediction of the effect of transistor parameter changes as well as optimizing the design of SRAM cells is instrumental in SRAM cell design. We developed an analytical method of Static Noise Margin (SNM) calculation for the recently proposed loadless four-transistor SRAM cells. The proposed model is based on the alpha-power law and tracks the simulated SNM value within 15%.

7.3 March Tests for Cell Stability Test in Embedded SRAMs

Detection of a weak cell by a functional test is possible if it has changed states as a result of a certain disturbance. Some of the possible conditions for a weak cell to change states during a March test are repetitive write and/or repetitive read operations or resistive and/or capacitive coupling to its neighbors. A subsequent read operation can then detect a flipped cell. My simulations showed that the detection capabilities of the March 11N and Hammer repetitive tests are insufficient for the reliable detection of stability and dynamic faults, which are caused by the same reasons as the stability faults. More reliable test methods have to be applied to ensure a high quality stability test. Small memory

instances with shorter access cycle are generally more demanding to the precise internal timing generation for stable operation. Moreover, due to smaller ratio of t_{access}/t_{flip}, the detection of stability faults detection with repetitive tests is more likely to be successful for high-speed memories with shorter access times.

I established the conditions for the successful detection of resistive coupling faults in the neighboring SRAM columns. However, the detection capability has proved to be limited to low resistive bridging defects.

7.4 DFT Techniques for Cell Stability Test in Embedded SRAMs

Reliable and economical cell stability detection in SRAMs calls for the use of defect-oriented Design for Testability (DFT) techniques. In this work we developed three novel DFT techniques for SRAM cell stability testing. The test stress to the CUT in Read Current Ratio with a Pass Transistor (RCRPT) and the Read Current Ratio Technique with floating bit lines (RCRT) techniques is formed by the read current ratio of a number of cells. The Word Line Pulsing Technique (WLPT) forms the test stress by pulsing the word line of the reference cell with subsequently applying modified bit line potentials to the CUT.

All of the proposed techniques offer a flexible digitally programmable pass/fail threshold and were proven to be effective in detecting Stability and Data Retention Faults. They exceed the capabilities of the known SRAM cell stability test techniques with regard to the flexibility of the applied test stress while also providing excellent defect resistance coverage.

Two test chips have been designed to provide a proof of concept for the proposed DFT techniques. The RCRT has been verified in a full-custom asynchronous self-timed SRAM test chip fabricated in CMOS 0.18μm technology. The measurement results proved

the effectiveness of the RCRT in detecting SRAM cells with compromised stability. The WLPT is implemented in a full-custom synchronous self-timed SRAM test chip designed in CMOS 0.13μm technology. The detection capability of the WLPT is shown to exceed one of the traditional Data Retention Tests by nearly four orders of magnitude. The obtained test results can be used for yield diagnostic and debug, for cell stability binning or as a criterion for redundancy calculation and application.

7.5 Future Work

We now discuss some of the possible areas of future work that can be pursued based on the results presented in this thesis.

Physical layout design of an SRAM cell plays a major role in ensuring cell stability and fabrication yield maximization. Optical proximity effects on the critical diffusion, poly, contact and M1 masks can deteriorate the SNM of an SRAM cell. Recently, the soft error rate of the new SRAM generations with reduced supply voltages and various low power and sleep transistor design techniques has started to grow. Layout-dependent capacitance on the critical nodes and their collection efficiency affect the soft error immunity. Maximizing the SNM by increasing the driver transistor size makes a cell faster but the packing density is reduced. A co-optimization study of SNM, soft-error immunity, speed and packing density for various bitcell architectures presents an interesting topic for future work.

Most of the memory test literature has considered on static functional fault models. Recent work shows that defect injection and SPICE simulation reveals that dynamic faults can take place in the absence of static faults. An example of a sequence sensitizing a dynamic fault is a fast read-after-write operations. In this case, the precharge time may be insufficient for adequate recovery after a write driver completely discharges a bit line and

the subsequent read operation may render a cell unstable. Most of the currently used tests are designed for static faults and may be unsuitable for detecting these dynamic faults. Different dynamic faults are reported for Intel and ST Microelectronics SRAM chips [68] implying that dynamic faults are architecture/technology dependent. March tests directed at the detection of the dynamic faults in the cells could be explored more thoroughly with respect to different cell architectures, defect types and locations in the SRAM cell and under the influence of PVT variations.

A new version of the Itanium-2 processor has incorporated enhanced manufacturability features such ECC in the L1 and L2 caches, the Programmable Weak Write Test Mode (PWWTM) [12] and extensive redundancy. Both of these features increase the ability to identify and repair defects [65]. The PWWTM is an active programmable detection technique for stability fault detection and is similar to the detection techniques proposed in this work. Even though the focus of this research is not the BIST implementation, the functionality and design of the proposed DFT techniques can be extended so that the proposed DFTs can be integrated with BIST circuitry to offer a complete SRAM cell stability test solution. The integration would require the addition of a NOR gate in the word line decoder, additional precharge gating in the test mode and a programmable counter. The BIST should include the means to ensure its functionality, such as the test scan chains.

It is worth mentioning that the detection concept introduced in this work is applicable to all other types of SRAM cells (such as the loadless 4T cells and 4T cells with resistive loads). The proposed DFT techniques similar to those presented in this thesis, can be generally extended and modified to apply to other types of SRAM cells as well.

Publications and Patents Resulted from this Work

Publications

- **A. Pavlov**, M. Sachdev and J. Pineda de Gyvez, "Weak Cell Detection in Deep-Submicron SRAMs: A Programmable Detection Technique", *(accepted for publication in IEEE Journal of Solid State Circuits (JSSC))*, pp. 1–10.

- **A. Pavlov**, M. Azimane, J. Pineda de Gyvez and M. Sachdev, "Programmable Techniques for Cell Stability Test and Debug in Embedded SRAMs", in Proc. IEEE Custom Integrated Circuits Conference (CICC-2005), San Jose, CA, Sept. 2005, pp. 443–446.

- **A. Pavlov**, M. Azimane, J. Pineda de Gyvez and M. Sachdev, "Word Line Pulsing Technique for Stability Fault Detection in SRAM cells", in Proc. IEEE International Test Conference (ITC-2005), Austin, TX, Nov. 2005, pp. 1–10.

- **A. Pavlov**, M. Sachdev and J. Pineda de Gyvez, "An SRAM Weak Cell Fault Model and a DFT Technique with a Programmable Detection Threshold", in Proc. IEEE International Test Conference (ITC-2004), Charlotte, NC, Oct. 2004, pp. 1106–1115.

- **A. Pavlov**, M. Sachdev and J. Pineda de Gyvez, "A Parametric-Stability Fault Model for Embedded SRAMs", 8-th IEEE European Test Workshop, Maastricht, The Netherlands, 2003 (poster presentation).

- **A. Pavlov**, "Investigation of Design for Testability Issues in Embedded SRAMs", Philips Research Technical Note, Eindhoven, The Netherlands, 2003, pp. 1–61.

- O. Semenov, **A. Pavlov** and M. Sachdev, "Sub-Micron SRAM Cell Stability in Low-Voltage Operation: A Comparative Analysis", IEEE International Integrated Reliability Workshop, Oct. 2002, pp. 168–171.

Patents

- J. Pineda de Gyvez, M. Azimane and **A. Pavlov**, "DFT for Weak SRAM Cell Detection", (patent application, 2004).

- J. Pineda de Gyvez, M. Sachdev and **A. Pavlov**, "Method and Apparatus to Detect Weak SRAM Cells", (patent application, 2003).

References

[1] International Technology Roadmap for Semiconductors (ITRS-2004) update. [Online]. Available: http://www.itrs.net/Common/2004Update/2004Update.htm

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

[3] J. Rabaey, A. Chandrakasan, and B. Nicoloc, *Digital Integrated Circuits: A Design Prospective. Second Edition.* Prentice Hall, 2003.

[4] TSMC literature. [Online]. Available: http://www.tsmc.com/english/function/f07.htm

[5] T. Saito, H. Ashihara, K. Ishikawa, M. Miyauchi, Y. Yamada, and H. Nakano, "A reliability study of barrier-metal-clad copper interconnects with self-aligned metallic caps," *IEEE Transactions on Electron Devices*, vol. 51, pp. 2129–2135, Dec. 2004.

[6] R. Montanés, J. Pineda de Gyvez, and P. Volf, "Resistance characterisation of open defects," *IEEE Design and Test of Computers*, vol. 19, pp. 18–26, 2002.

[7] A. Meixner and J. Banik, "Weak write test mode: An SRAM cell stability design for test technique," in *Proc. IEEE International Test Conference (ITC)*, Nov. 1997, pp. 1043–1052.

[8] D.-M. Kwai, H.-W. Chang, H.-J. Liao, C.-H. Chiao, and Y.-F. Chou, "Detection of SRAM cell stability by lowering array supply voltage," in *Proc .of the Ninth Asian Test Symposium (ATS 2000)*, Dec. 2000, pp. 268–273.

[9] C. Kuo, T. Toms, B. Neel, J. Jelemensky, E. Carter, and P. Smith, "Soft-defect detection (SDD) technique for a high-reliability CMOS SRAM," *IEEE J. Solid-State Circuits*, vol. 25, pp. 61–67, Feb. 1990.

[10] A. Pavlov, M. Sachdev, and J. Pineda de Gyvez, "An SRAM weak cell fault model and a DFT technique with a programmable detection threshold," in *Proc. IEEE International Test Conference (ITC)*, Nov. 2004, pp. 1106–1115.

[11] International technology roadmap for semiconductors - 2003 (ITRS-2003). [Online]. Available: http://public.itrs.net/

[12] E. Selvin, A. Farhang, and D. Guddat, "Programmable weak write test mode," U.S. Patent 6 778 450 B2, Aug. 17, 2004.

[13] D. Weiss, J. Wuu, and R. Reidlinger, "Integrated weak write test mode WWTM," U.S. Patent 6 192 001 B2, Feb. 20, 2001.

[14] P. Wong and F. Towler, "Method and apparatus for identifying SRAM cells having weak pull-up PFETS," U.S. Patent 6 552 941 B2, Aug. 17, 2004.

[15] M. Mehalel, "Short write test mode for testing static memory cells," U.S. Patent 6 256 241 B1, July 3, 2001.

[16] W. Schwarz, "Data retention weak write circuit and method of using same," U.S. Patent 5 835 429, Nov. 10, 1998.

[17] T. Liston and L. Herr, "Method and apparatus for soft defect detection in a memory," U.S. Patent 6 590 818, July 8, 2003.

[18] R. H. W. Salters, "Device with integrated SRAM memory and method of testing such a device," U.S. Patent 6 757 205, June 29, 2004.

[19] J. Wuu, D. Weiss, C. Morganti, and M. Dreesen, "The asynchronous 24MB on-chip level-3 cache for a dual-core Itanium-family processor," in *IEEE International Solid-State Circuits Conference*, Apr. 2005, pp. 488–489.

[20] B. F. Cockburn, F. Lombardi, and F. J. Meyer, "DRAM architecture and testing," *IEEE Design and Test of Computers*, pp. 19–21, Jan. 1999.

[21] A. Sharma, *Andvanced Semiconductor Memories: Architectures, Designs and Applications*. Wiley Inter-Science, 2003.

[22] A. van de Goor, *Testing Semiconductor Memories: Theory and Practice*. A. van de Goor, 2001.

[23] F. Ferguson and J. Shen, "Extraction and simulation of realistic CMOS faults using inductive fault analysis," in *IEEE International Test Conference*, Nov. 1988, pp. 475–484.

[24] J. Banik, A. Meixner, G. King, and D. Guddat, "Static random access memory SRAM having weak write test circuit," U.S. Patent 5 559 745, Sept. 24, 1996.

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

[26] J. Pineda de Gyvez and D. Pradhan, *Integrated Circuit Manufacturability*. Institute of Electrical and Compuiter Engineers Inc., 1998.

[27] T. Harazsti, *CMOS Memory Circuits*. Kluwer Academic Publishers, 2000.

[28] (2005) Sizing Cu-11 dense SRAM and fuse redundancy options. [Online]. Available: http://www-306.ibm.com/chips/techlib/techlib.nsf/techdocs/82C332ECD949BD0F8725700A0050054B

[29] L.-S. Kim and P. W. Dutton, "Metastability of CMOS latch/flip-flop," *IEEE J. Solid-State Circuits*, vol. 25, pp. 942–951, Aug. 1990.

[30] K. Imai *et al.*, "A 0.13μm CMOS technology integrating high-speed and low power/high density devices with two different well/channel structures," in *IEDM Technical Digest*, Oct. 1999, pp. 667–690.

[31] NEC, in *ISSCC Digest of Technical Papers*, Feb. 2001.

[32] S. Masuoka *et al.*, "A 0.99-m^2 loadless four-transistor SRAM cell in 0.13μm generation CMOS technology," in *Proceedings of Symposium on VLSI Tech.*, June 2000, pp. 164–165.

[33] K. Takeda *et al.*, "A 16-Mb 400-MHz loadless CMOS four-transistor SRAM macro," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1631–1640, Nov. 2000.

[34] E. Seevinck, F. List, and J. Lohstroh, "Static-noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 748–754, Oct. 1987.

[35] F. List, "The static noise margin of SRAM cells," in *in Proc. of ESSCIRC*, May 1986, pp. 16–18.

[36] B. Cheng, S. Roy, and A. Asenov, "The impact of random doping effects on CMOS SRAM cell," in *Proc. IEEE Solid-State Circuits Conference ESSCIRC*, Leuven, Belgium, 2004, pp. 219–222. [Online]. Available: http://www.esscirc.org/

[37] T. Hirose *et al.*, "A 20-ns 4-Mb CMOS SRAM with hierarchical word decoding architecture," *IEEE Journal of Solid-State Circuits*, vol. 25, pp. 1068–1073, Oct. 1990.

[38] B. Amurtur and M. Horowitz, "A replica technique for wordline and sense control in low-power SRAMs," *IEEE Journal of Solid-State Circuits*, vol. 33, pp. 1208–1219, Aug. 1998.

[39] M. Eisele *et al.*, "The impact of intra-die device parameter variations on path delays on the design for yield of low voltage digital circuits," in *IEEE International Symposium Low Power Electronic Design*, Oct. 1996, pp. 237–242.

[40] S. Tachibana *et al.*, "A 2.6-ns wave-pipelined CMOS SRAM with dual-sensing-latch circuits," *IEEE Journal of Solid-State Circuits*, vol. 30, pp. 487–490, Apr. 1995.

[41] S. Schuster *et al.*, "A 15-ns CMOS 64k RAM," *IEEE Journal of Solid-State Circuits*, vol. 21, pp. 704–711, Oct. 1986.

[42] C. Hill, "Definitions of noise margin in logic systems," *Mullard Tech. Commun.*, vol. 89, pp. 239–245, Feb. 1967.

[43] J. Lohstroh, "Static and dynamic noise margins of logic circuits," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 591–598, June 1979.

[44] J. Lohstroh, E. Seevinck, and J. de Groot, "Worst-case static noise margin criteria for logic circuits and their mathematical equivalence," *IEEE J. Solid-State Circuits*, vol. SC-18, pp. 803–807, Dec. 1983.

[45] T. DeMassa and Z. Ciccone, *Digital Integrated Circuits*. Jonh Wiley & Sons, 1996.

[46] S. Mitra, *An Introduction to Digital and Analog Integrated Circuits and Applica-tions*. Harper & Row Publishers, 1980.

[47] L. Glasser and D. Dobberpuhl, *The Design ans Analysis of VLSI Circuits*. Addison-Welsley Publishing, 1985.

[48] J. Hauser, "Noise margin criteria for digital logic circuits," *IEEE Transactions on Education*, vol. 36, pp. 363–368, Nov. 1993.

[49] A. Bhavnagarwala, X. Tang, and J. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid State Circuits*, vol. 36, pp. 658–665, Apr. 2001.

[50] P. Stolk, H. Tuinhout, *et al.*, "CMOS device optimization for mixed-signal technologies," in *IEEE International Electron Devices Meeting IEDM Technical Digest*, Oct. 2001, pp. 10.2.1–10.2.4.

[51] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1837 – 1852, 2003.

[52] Carafe inductive fault analysis IFA tool. [Online]. Available: http://sctest.cse.ucsc.edu/carafe/

[53] T. Karnik, P. Hazucha, and J. Patel, "Characterization of soft errors caused by single event upsets in CMOS processes," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, pp. 128–143, Apr. 2004.

[54] R. Baumann and E. Smith, "Neutron-induced boron fission as a major source of soft errors in deep submicron SRAM devices," in *Proceedings of IEEE International Reliability Physics Symposium*, Apr. 2000, pp. 152–157.

[55] P. Roche *et al.*, "Determination of key parameters for SEU occurrence using 3-D full cell SRAM simulations," *IEEE Transactions on Nuclear Science*, vol. 46, pp. 1354–1362, Dec. 1999.

[56] R. Baumann, "Ghost in the machine: A tutorial on single-event upsets in advanced commercial silicon technology," in *a tutorial at IEEE International Test Conference (ITC)*, Nov. 2004.

[57] D. Kang and Y.-B. Kim, "A deep sub-micron SRAM cell design and analysis methodology," in *Proc. of MWSCAS-2001*, Aug. 2001, pp. 858–861.

[58] T. Sakurai and A. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, pp. 584 – 594, Apr. 1990.

[59] S. Vemuru, N. Scheinberg, and E. Smith, "Short-circuit power dissipation formulae for CMOS gates," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 1993, pp. 1333–1336.

[60] E. Ogawa *et al.*, "Stress-induced voiding under vias connected to wide Cu metal leads," in *Proc. of IRPS*, Oct. 2002, p. 312321.

[61] D. Adams, R. Abbott, X. Bai, D. Burek, and E. MacDonald, "An integrated memory self test and EDA solution," in *IEEE International Workshop on Memory Technology, Design and Testing (MTDT'04)*, Aug. 2004, pp. 92–95.

[62] F. Arnaud, F. Boeuf, F. Salvetti, D. Lenoble, F. Wacquant, *et al.*, "A functional $0.69\mu m^2$ embedded 6T-SRAM bit cell for 65nm CMOS platform," in *Symposium on VLSI Technology*, June 2003, pp. 55–56.

[63] M. Sachdev, *Defect Oriented Testing for CMOS Analog and Digital Circuits*. Kluwer Academic Publishers, 1998.

[64] J. Segura, A. Keshavarzi, J. Soden, and C. Hawkins, "Parametric failures in CMOS ICs – a defect-based analysis," in *Proc. IEEE International Test Conference (ITC)*, Oct. 2002, pp. 90–98.

[65] S. Rusu *et al.*, "A 1.5-GHz 130-nm Itanium-2 processor with 6-MB on-die L3 cache," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 1887–1895, Nov. 2003.

[66] A. van de Goor, "An industrial evaluation of DRAM tests," *IEEE Design and Test of Computers*, vol. 21, pp. 430–440, Sept. 2004.

[67] S. Hamdioui, Z. Al-ars, and A. van de Goor, "Testing static and dynamic faults in random access memories," in *IEEE VLSI Test Symposium*, Oct. 2002, pp. 395–400.

[68] S. Hamdioui, R. Wadsworth, J. Reyes, and A. van de Goor, "Importance of dynamic faults for new SRAM technologies," in *IEEE European Test Workshop*, May 2003.

[69] R. Heald and P. Wang, "Variability in sub-100nm SRAM designs," in *IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, Nov. 2004, pp. 347–352.

[70] A. Pavlov, M. Sachdev, and J. Pineda de Gyvez, "Weak cell detection in deep-submicron SRAMs: A programmable detection technique," *to appear in IEEE Journal of Solid-State Circuits*.

[71] A. Pavlov, M. Azimane, J. Pineda de Gyvez, and M. Sachdev, "Word line pulsing technique for stabilty fault detection in SRAM cells," in *IEEE International Test Conference (ITC)*, Nov. 2005.

[72] B. Amrutur and M. Horowitz, "A replica technique for wordline and sense control in low-power SRAMs," in *IEEE Journal of Solid State Circuits (JSSC)*, Aug. 1998, pp. 1208–1219.

[73] H. Nambu *et al.*, "A 1.8ns access, 550MHz 4.5Mb CMOS SRAM," in *IEEE Int. Solid-State Circuits Conference (ISSCC)*, Oct. 1998, pp. 360–361.

Glossary

A

ATD Address Transition Detector – a circuit used in asynchronous SRAMs for detecting transitions of address or Chip Select pulse and initiating a read or write operation.

B

BL, BLB Bit Line, Bit Line Bar (complementary bit line).

C

CMP Chemical-Mechanical Polishing a.k.a. Chemical-Mechanical Planarization – removing of the excessive material from the wafer by using both the chemical and mecanical action.

CUT Cell Under Test.

D

DFT Design For Test – a design approach allowing for enhanced test capabilities of a circuit.

DRAM DRAM – Dynamic Random Access Memory.

F

FSM Finite State Machine.

K

KCL Kirchhoff's Current Law – based on the conservation of charge, the total charge flowing into a node must be the same as the the total charge flowing out of the node.

M

Moore's Law Moore's Law - An observation by Gordon Moore (Intel) that the market demand (and semiconductor industry response) for functionality per chip (bits, transistors) doubles every 1.5 to 2 years. He also observed that MPU performance [f_{clk}(MHz)$*$instructions per clock = millions of instructions per second (MIPS)] also doubles every 1.5 to 2 years. Although viewed by some as a self-fulfilling prophecy, Moores Law has been a consistent macro trend and key indicator of successful leading-edge semiconductor products and companies for the past 30 years.

MPU MPU – Microprocessor Unit.

P

PVT Process, Voltage, Temperature – the typical sensitivity parameters.

R

RCRPT Read Current Ratio with a Pass transistor Technique – a proposed stability test technique that uses the ratio of read currents of a number of cells and a pass transistor to precondition the bit lines applied directly to the CUT.

RCRT Read Current Ratio Technique – a proposed stability test technique that uses the ratio of read currents of a number of cells to precondition the bit lines and then applies the partially discharged floating bit lines to the CUT.

S

SER Soft Error Rate, measured in $FIT = 1\ failure/10^9\ dev - hours$.

SEU Single Event Upset.

SNM Static Noise Margin of an SRAM cell, defined as the side of the smaller of the two squares that can be embedded between VTC curves of the two equivalent inverters of an SRAM cell.

SRAM SRAM – Static Random Access Memory.

T

technology node ITRS Technology node – minimum half-pitch of custom-layout metal interconnect for memory (typically DRAM) and physical bottom gate length

for MPU.

V

VTC Voltage Transfer Characteristic, a plot of V_{in} vs. V_{out}.

W

WL Word Line.

WLPT Word Line Pulsing Technique – a proposed stability test technique that uses the pulsing of the word line to precondition the bit lines and then applies the partially discharged floating bit lines to the CUT.

WWTM Weak Write Test Mode – a stability test technique that uses a weak write operation to determine cell's stability (© Intel).