Efficient Testing of SRAM With Optimized March Sequences and a Novel DFT Technique for Emerging Failures Due to Process Variations

Qikai Chen, *Student Member, IEEE*, Hamid Mahmoodi, *Student Member, IEEE*, Swarup Bhunia, *Student Member, IEEE*, and Kaushik Roy, *Fellow, IEEE*

***Abstract*—With increasing inter-die and intra-die parameter variations in sub-100-nm process technologies, new failure mechanisms are emerging in CMOS circuits. These failures lead to reduction in reliability of circuits, especially the area-constrained SRAM cells. In this paper, we have analyzed the emerging failure mechanisms in SRAM caches due to transistor V_t variations, which results from process variations. Also we have proposed solutions to detect those failures efficiently. In particular, in this work, SRAM failure mechanisms under transistor V_t variations are mapped to logic fault models. March test sequences have been optimized to address the emerging failure mechanisms with minimal overhead on test time. Moreover, we have proposed a design for test circuit to complement the March test sequence for at-speed testing of SRAMs. The proposed technique, referred as double sensing, can be used to test the stability of SRAM cells during read operations. Using the proposed March test sequence along with the double sensing technique, a test time reduction of 29% is achieved, compared to the existing test techniques with the same fault coverage. We have also demonstrated that double sensing can be used during SRAM normal operation for online detection and correction of any number of random read faults.**

***Index Terms*—Design for test (DFT), failure mechanism, March test, process variation, SRAM.**

I. Introduction

AS THE silicon industry moves toward the end of the technology roadmap, controlling the fabrication of scaled devices is becoming a great challenge. The device parameter variations (such as the variation in channel length, oxide thickness, and random placement of dopants in channel) are expected to be significantly large in future generations [1]. Process variations can be classified as inter-die or intra-die variations. Due to inter-die variations, the same device on a chip can have different characteristics across different dies. On the other hand, intra-die variations are the variations of transistor characteristics within a single chip. Although intra-die variations in terms of channel length, width, and oxide thickness are expected to exhibit spatial correlations to some extent, random dopant fluctuations in sub-50-nm technology make every transistor in a die independent in terms of threshold voltage [1]. These atomic level fluctuations are difficult to eliminate through external control of the manufacturing process and are most pronounced in minimum geometry transistors commonly used in area-constrained circuits such as SRAM cells.

Memory subsystems dominate the chip area of the state-of-the-art microprocessors (predicted to occupy about 94% of die area by 2014 [2]). Process variations have different impacts on different components of a memory subsystem. Delay of the memory address decoder varies from the target value due to inter-die variations [3]. With random dopant fluctuations, similar transistors on a die may have different strengths (different threshold voltages). Such variations affect the stability of six-transistor (6T) SRAM cells. Moreover, the variations also affect proper functionality of sense amplifiers (SAs).

There is considerable previous work on memory tests. In [4]–[6], authors discuss the March test, which is the most commonly used manufacturing test for memories. In [7], the failure probability due to parameter variations is analyzed. In [8], authors propose architectural techniques to reconfigure the memory in order to improve yield under process variations. In [9] and [10], authors investigate the yield of SAs, which are an integral part of memories.

In this paper, we have analyzed the impact of individual transistor V_t deviations on the SRAM physical failure mechanisms in all major components of a memory subsystem. Moreover, we have proposed efficient test solutions to cover the emerging faults introduced by process variations through optimized test algorithms and a design for test (DFT) circuit technique. In particular, the main contributions of this paper are as follows:

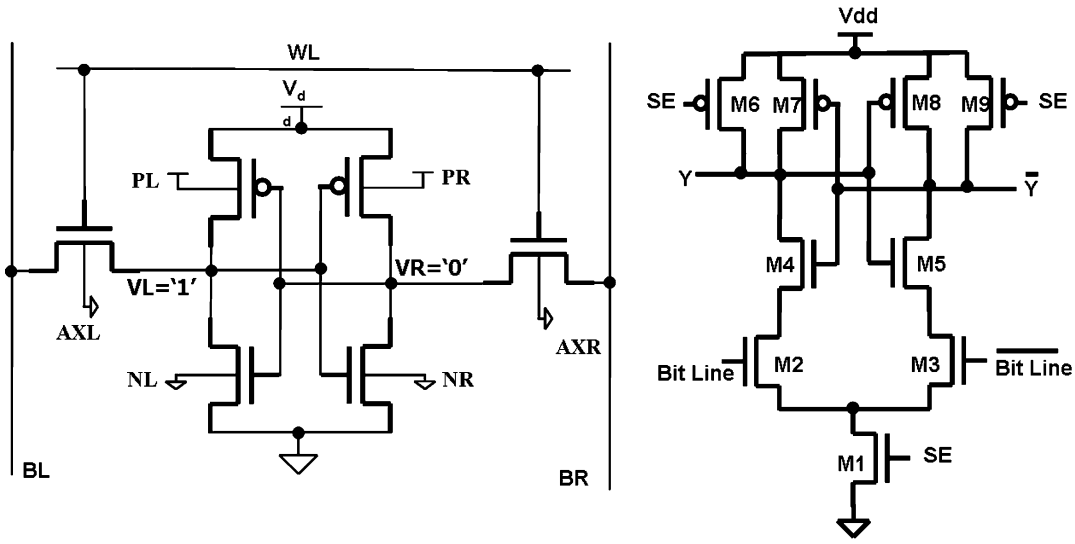
- analysis of failure mechanisms due to individual transistor V_t variations in SRAM subsystems and development of corresponding logic fault models;
- development of two March test sequences to achieve improved fault coverage (considering the failures due to transistor V_t variations) with minimum increase in test time;
- a novel DFT circuit, referred to as double sensing, to reduce test application time of the March test;
- application of double sensing in online testing so as to efficiently detect and correct specific faults for SRAM caches during normal mode of operation.

Manuscript received January 13, 2005; revised June 23, 2005 and June 27, 2005. This work was supported in part by MARCO GSRC and the National Science Foundation.

Q. Chen, S. Bhunia, and K. Roy are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: qikaichen@purdue.edu; bhunias@purdue.edu; kaushik@purdue.edu).

H. Mahmoodi is with the School of Engineering, San Francisco State University, San Francisco, CA 94132 USA and also with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: mahmoodi@purdue.edu).

Digital Object Identifier 10.1109/TVLSI.2005.859565



(a) (b)

Fig. 1. (a) 6T-SRAM cell. (b) SA.

The remainder of this paper is organized as follows. In Section II, the SRAM failure mechanisms under transistor V_t variations are analyzed and categorized into the corresponding logic fault models. Section III presents optimized March test sequences which improve the fault coverage under process variations. In Section IV, a novel DFT technique is introduced to complement the March test so as to reduce the test time. In Section V, the applications of the proposed DFT technique are discussed. Section VI draws the conclusions.

II. Failure Mechanisms and Fault Models

To compare logical behavior of faulty memories against good ones, modeling the physical failure mechanisms as logic fault models is required. This section first introduces the established logic fault models within the scope of interest. Then SRAM physical failure mechanisms under V_t deviations are investigated in the subsystem components, including SRAM cells, SAs, and address decoders. Finally, those physical failures are mapped to the logic fault models.

A. Logic Fault Models

In [4], established logic fault models in SRAM are summarized. The fault models of interest are single-cell and coupling fault models (namely, Stuck-at Fault; Transition Fault; Write Disturb Fault; Read Destructive Fault; Deceptive Read Destructive Fault; Incorrect Read Fault; Random Read Fault; Data Retention Fault; State Coupling Fault; Disturb Coupling Fault; Transition Coupling Fault; Read Destructive Coupling Fault; Incorrect Read Coupling Fault and Deceptive Read Destructive Coupling Fault). It should be noted that the following sections of the paper mainly focus on the test of Deceptive Read Destructive Fault and Data Retention Fault. Deceptive Read Destructive Fault refers to the SRAM logical behavior in which SRAM cell contents flip during read operations but the read output is the initially stored value. Data Retention Fault is the SRAM logical behavior in which an SRAM cell loses its stored value when the cell is not accessed. As observed in the following sections, these two types of faults have high probabilities of occurrence due to V_t variations. Moreover, these two types of faults are not detected efficiently by the conventional SRAM test techniques. Therefore, in Section III, the March test is optimized to cover the above-mentioned fault models with minimal increase in test time. Also, in Section IV, a DFT technique is introduced to complement the March test so as to detect Deceptive Read Destructive Fault without test time increase.

B. Failure Mechanisms in SRAM Under Process Variations

Intra-die variations, resulting from mismatches in parameters of similar transistors (threshold voltage V_t and geometry L or W), may lead to new failures in memories. These mismatches modify the strength of individual transistors resulting in various failures. The principal source of mismatch is the intrinsic fluctuation of V_t due to random dopant effect [11]. Therefore, in this work, we focus on V_t variations. In this work, V_t's of transistors are considered to be independent random variables. The V_t shift of each transistor is considered to be a zero mean Gaussian distribution, with [12]

$$\sigma_{\mathrm{Vt}} = \sigma_{\mathrm{Vt0}}\sqrt{(L_{\min}/L)(W_{\min}/W)} \tag{1}$$

where σ_{Vt0} is the standard deviation of the V_t shift of a minimum-sized transistor, σ_{vt0} depends on the doping concentration and the oxide thickness of a particular technology [12], and $L_{\min}$ and $W_{\min}$ are the minimum length and width of a transistor in a given technology.

1) SRAM Cell Failure Mechanisms: A 6T-SRAM cell is shown in Fig. 1(a). Due to its area constraint, SRAM is particularly vulnerable to process variations. V_t mismatch in the six transistors of SRAM cells may result in [7] one of the following conditions.

1) A decrease in the current that discharges the bit-lines during read operations.

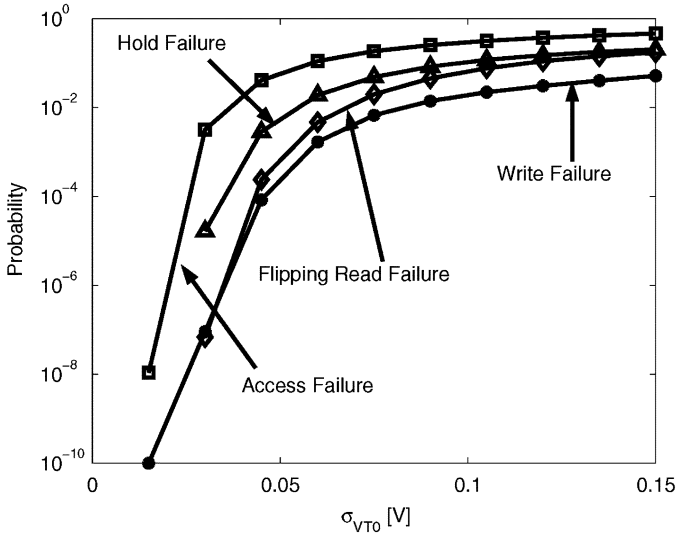


Fig. 2. Failure probability versus σ_{Vt0} [7].

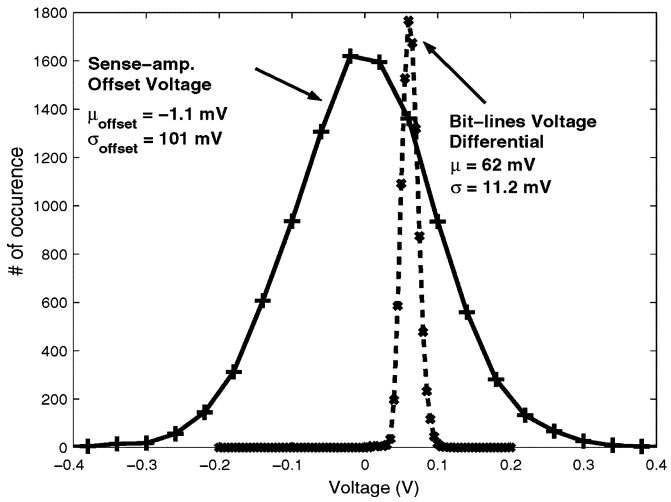


Fig. 3. Distributions of SA offset and bit-line voltage difference (σ_{vt0}=80 mv; 50-nm process [14]).

Weak access transistors (AXL or AXR) or weak pull-down transistors (NL or NR) result in a reduction of the bit-line discharging current. This mechanism leads to less voltage differential between the bit-lines when the SA fires. Therefore, it may result in a wrong evaluation in the SA. We refer to this failure as *access failure*. The reduced bit-line differential may lead to a wrong evaluation of the SA due to its own offset voltage shift (discussed in Section II–B2).

2) Increased disturbance to the content of the SRAM cell during read operations.

   During read operations, comparatively strong access transistors or weak pull-down transistors lead to an increase in the internal node voltage that stores "0" in the SRAM cell [VR in Fig. 1(a)]. If the pull-up transistor of the cross-coupled inverter [PL in the case of Fig. 1(a)] is weak, the SRAM cell may flip during the read operation. We refer to this failure as flipping read failure. It should be noted that the flipping of the internal voltage can happen in the late stage of the read cycle. Therefore, the flipping may not change the bit-line differential. Also, the output of the read operation may still be the correct value, which is the initial stored value.

3) Unsuccessful write operations.

   Comparatively strong pull-up transistors or weak pull-down transistors shift the trip point of the cross-coupled inverters to a higher voltage. This may lead to an unsuccessful write operation of the SRAM cell. We refer to this failure as write failure.

4) Instability of a SRAM cell in holding its content even when it is not accessed.

   Due to excessive mismatch in the cross-coupled inverters, a SRAM cell may lose its value without the cell being accessed by any operations, especially when the supply voltage is lowered to save leakage. We refer to this failure as hold failure.

In Fig. 2, the probabilities for the above-mentioned failures with different V_t variation values (σ_{vt0}) are plotted [7]. The X axis shows the assumed σ_{Vt0} values. Then the V_t distribution of each transistor in a SRAM cell is calculated based on (1), taking its size into consideration. The failure probabilities are estimated by Monte Carlo simulations for a 50-nm predictive technology [13]. As shown in Fig. 2, the failures are more likely to occur with an increase in V_t variations.

2) SA Failure Mechanisms: Fig. 1(b) shows an SA used in SRAM design. Functional failures in SAs result from V_t mismatch between the differential pair input transistors (M2 and M3) as well as V_t mismatches in the cross-coupled inverters (M4, M5, M7, and M8). The differential pair transistors (M2 and M3) have dominant impact on the SA functional failures. During the evaluation period, unbalanced V_t's of M2 and M3 modify the amount of current on each side discharging nodes Y and $\bar{\mathrm{Y}}$. This causes a shift in the offset voltage (the minimum voltage differential between two inputs for the SA to function correctly [9]). Monte Carlo simulation is performed to obtain the offset voltage distributions. In the simulation, we have assumed σvt0 to be 80 mv. Fig. 3 shows the the offset voltage distribution of an SA when its output is evaluated to "1." In Fig. 3, the X axis is the offset voltage. The Y axis represents the number of cases when the offset voltage falls into the corresponding voltage range. It should be noted that, due to V_t variations, the offset voltage [differential input between M2 and M3 in Fig. 1(b)] can be negative for the SA to evaluate to "1." Also, as shown in Fig. 3, the mean offset voltage (μ) is -1.1 mv, which is close to 0. Therefore, the SA offset voltage distribution can be approximated by a zero-mean Gaussian distribution with $\sigma = 101$ mv. It is observed from Fig. 3 that the offset voltage is sensitive to the transistor V_t variations. Together with the offset voltage distribution, Fig. 3 also plots the distribution of the voltage differential between the two bit-lines at the time when they are sampled. In this scenario, a "1" is stored in the accessed SRAM cell. As shown in Fig. 3, although most of the SRAM cells can develop enough voltage differential, for those SAs whose offset voltages are larger than the bit-line differential, they still evaluate to a wrong output due to the offset voltage shift.

Assuming one SA per column, the probability that the output of one particular column is incorrect corresponds to the probability that the offset voltage of the SA is larger than the min-

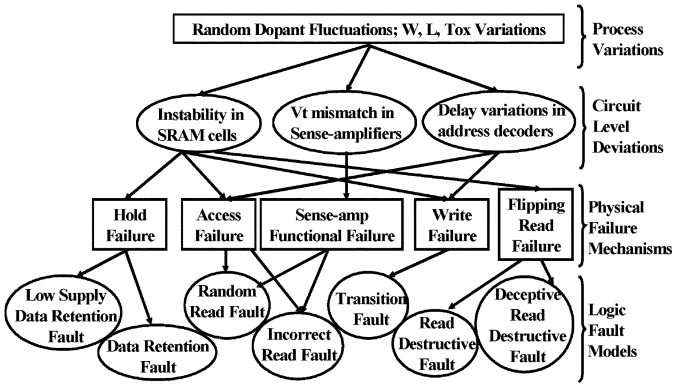


Fig. 4. Failure mechanisms and logic fault models.

imum voltage differential developed by any SRAM cell in that column. Therefore, the probability that the output of a SA is incorrect can be estimated as

$$P_{\text{Incorrect/Column}} = \int_{-\infty}^{+\infty} f_{\text{SenseAmp}}(x)\{1-[1-F_{\text{SRAMCell}}(x)]^n\}dx \quad (2)$$

where $f_{\text{SenseAmp}}(x)$ is the probability density function (pdf) of the SA offset voltage distribution, $F_{\text{SRAMCell}}(x)$ is the cummulative density function (cdf) of the bit-line differential distribution, and n is the number of SRAM cells per column. The pdf or cdf of the SA offset voltage and bit-line voltage differential can be evaluated by Monte Carlo simulations or semi-analytical methods as described in [14].

3) Address Decoder Failure Mechanisms: Under individual transistor V_t deviations, the delay of the address decoder suffers from variations. Delay variations of the address decoder can result in shorter time left for accessing the SRAM cell. This may consequently result in access failures or write failures in SRAM cells. The probability that the address decoder fails to meet the delay target can be obtained from statistical delay distribution of the decoder. For practical purposes, statistical delay distribution of a combinational circuit can be caculated as Gaussian distribution using the method proposed in [15] and [16].

C. Fault Models

Fig. 4 summarizes the failure mechanisms in SRAM under process variations and shows mapping of the failures to the logic fault models. Among the failure mechanisms: 1) SRAM access failures and SA functional failures may show themselves as either incorrect read faults or random read faults, depending on the noise level and the sense-ampifier offset voltage; 2) SRAM flipping read failure is modeled by read destructive fault or deceptive read destructive fault, based on the time the cell flips and how fast bit-lines responds to that flip; and 3) SRAM hold failure is modeled as data retention fault if the failure occurs at the nominal supply voltage. However, most of the hold failures happen in the standby mode (when the supply voltage is decreased to reduce leakage power). Extending the concept of data retention fault, we introduce a new fault model named the low supply data retention fault to describe flipping failures occuring due to application of low supply voltage in the standby mode.

By mapping the process variation related failure mechanisms to logic fault models, memory test can be designed to target failures in nano-scale SRAMs. As shown in Fig. 2, the flipping read failure in SRAM cells has a high probability of occurrence ($\sim$2% of cells at $\sigma_{\text{vt0}} = 80$ mv). Most of the flipping read failures show themselves as deceptive read destructive faults, which is overlooked in conventional memory test. Moreover, low supply data retention faults are also likely to occur (about 4% with $\sigma_{\text{vt0}} = 80$ mv). These types of faults are not emphasized by the conventional March test. Therefore, in the remainder of this paper, we propose techniques to address efficient detection of these logical faults through optimization of the March test and a novel low-cost DFT circuit.

III. March Test Optimization

The March test is prevalently applied to the SRAM test [17]. A March sequence consists of several March elements. A March element is a set of operations on the memory cell, including W0 (write 0), W1, R0 (read and expect 0 for output), and R1. All of these operations of one March element are applied to a certain address before proceeding to the next address, either in an increasing (denoted by $\Uparrow$ before a March element) or a decreasing (denoted by $\Downarrow$) order.

To cover the traditional single-cell and coupling fault models, March C- is popularly used as a base sequence [4]. However, a serious problem with March C- is that it does not detect deceptive read destructive faults, which are very likely to happen in memories due to individual transistor V_t deviations. In [5], a test sequence named March SR is proposed that covers the deceptive read destructive faults. However, the sequence has a test time of $14N$ (where N is the number of addresses in a memory), which is significantly higher compared to the test time of March C- ($10N$). Moreover, conventional March test sequences overlook data retention faults and low supply data retention faults. It is increasingly important to test for these faults because of the necessity to apply low supply voltage in the standby mode to reduce leakage power consumption in scaled technologies. Due to V_t mismatches from intra-die variations, lowering supply voltage induces hold failures in SRAM cells. Hence, modification of the March sequences to test for low supply data retention fault is an emerging requirement. We propose two March sequences that can cover the above fault models with a minimum impact on test time.

The first sequence is based on the well-known March C-. We have observed that two extra read operations are required to detect deceptive read destructive faults. In addition, for detection of low supply data retention faults, proper places are identified in the sequence (denoted by HOLD) to lower the supply voltage: keep the memory at a lower supply for a specified time and then bring it back to normal supply. The March C- sequence with the above-mentioned modifications is presented as follows:

$$\textbf{Extended March C}-: \begin{array}{l} \Updownarrow(\text{W0}) \Uparrow(\text{R0W1}) \Uparrow(\text{R1W0}) \\ \Downarrow(\text{R0W1})(\text{HOLD}) \\ \Downarrow(\text{R1R1W0})(\text{HOLD}) \Updownarrow(\text{R0R0}). \end{array}$$

TABLE I
COMPARISONS OF MARCH SEQUENCES

Logic Fault Models	Conventional Test Sequences			Proposed Sequences	
	March C-	March B	March SR	Opt. March C-	March Q
Address Decoder Fault	+	+	-	+	-
Data Retention Fault	-	-	+	+	+
Low Supply Data Retention Fault	-	-	-	+	+
Stuck-at Fault	+	+	+	+	+
Transition Fault	+	+	+	+	+
Write Disturb Fault	-	-	-	-	-
Random Read Fault	-	-	-	-	-
Read Destructive Fault	+	+	+	+	+
Deceptive Read Destructive Fault	-	-	+	+	+
Incorrect Read Fault	+	+	+	+	+
State Coupling Fault	+	-	+	+	+
Write Disturb Coupling Fault	-	-	-	-	-
Disturb Coupling Fault	+	-	+	+	+
Incorrect Read Coupling Fault	+	-	+	+	+
Read Destructive Coupling Fault	+	-	+	+	+
Transition Coupling Fault	+	-	+	+	-
Test Time	10N	17N	14N	12N	10N

We refer to the proposed sequence as Extended March C-. The Extended March C- has a test time of $12N$, resulting in an improvement of 15% compared to March SR [5]. Furthermore, compared to March SR, the extended March C- has a better fault coverage by detecting address decoder faults. To detect the address decoder faults, a set of test operations should be applied to the memory in a designed address order. For example, first, in ascending address order, the test sequence should verify the content of every SRAM cell and then force a transition of the stored value. After that, in descending address order, the inverted value of every SRAM cell should be verified by another read operation. Therefore, if more than one memory address leads to the same SRAM cell, it is going to be detected either by the first read in the ascending address order or by the second read in the descending order. It should be noted that, to detect an address decoder error completely, the example test operations should be applied once again in the opposite address order. In Extended March C-, the $\Uparrow (\mathrm{R1W0}) \Downarrow (\mathrm{R0W1}) \Downarrow (\mathrm{R1R1W0})$ operation detects address decoder errors. However, in March SR, address decoder errors can not be detected because, after each set of operations, such as $\Uparrow (\mathrm{R0, W1, R1, W0})$or $\Downarrow (\mathrm{R1, W0, R0, W1})$, the SRAM cell value is not inverted. Therefore, there are no opportunities to detect address decoder faults in March SR. In order to maintain the test time of March C- (10 N) and yet detect both deceptive read destructive faults and low supply data retention faults, we have developed a novel test sequence, namely March Q, as described below:

$$\text{March Q}-: \quad \begin{array}{l} \Downarrow (\mathrm{W0})(\mathrm{HOLD}) \Uparrow (\mathrm{R0W0W1R1}) \\ (\mathrm{HOLD}) \Uparrow (\mathrm{R1W1W0R0}) \Downarrow (\mathrm{R0}). \end{array}$$

Table I compares the fault coverage and test time of March C-, March SR, March B [17], Extended March C-, and March Q. The test sequences of March C-, March SR, and March B are summarized in Appendix A. "+" or "−" in Table I denote whether the March sequence is able to cover a logic fault model or not. From Table I, it can be observed that the Extended March C- sequence has the best fault coverage, with a test time increase of 20% compared to March C-. March Q detects the deceptive read destructive faults, while maintaining the shortest test time ($10N$). However, March Q cannot detect transition coupling faults. Also, since in March Q all of the SRAM write transition operations are applied with the same address order, it is not capable of detecting address decoder faults (like March SR). Hence, with a similar fault coverage as March SR, the proposed March Q sequence achieves 30% less test time. We conclude that, if transition coupling faults are not major concerns, March Q can be a promising test sequence in scaled technologies.

From the above discussion, we have observed that optimization of the March test to cover all of the emerging faults (i.e., deceptive read destructive faults) will inevitably encounter some test time overhead, such as the 20% test time increase of Extended March C- compared with March C-. In order to reduce the test time, DFT techniques can be explored. In the following section, we propose a DFT circuit to complement proper March test sequences for reduction of test time while maintaining the best fault coverage.

IV. DOUBLE SENSING: A DFT TECHNIQUE TO REDUCE TEST TIME

As discussed in the previous section, it is difficult to optimize the March test to improve the fault coverage without trading off the test time. In this section, a DFT circuit is proposed that reduces the test time without affecting the fault coverage. The DFT technique, which we refer to as double sensing, is applied to test the read stability of SRAM cells. The idea of double sensing technique is to replace the consecutive read operations applied to detect deceptive read destructive faults in some test sequences by a single read operation. In this single read operation, the double sensing technique pushes the unstable SRAM cells to flip and then detects the fault. Therefore, employing the proposed DFT circuit, memory test time can be improved significantly. Moreover, the overhead of the proposed DFT circuit is negligible. It should be noted that, during test, the DFT technique is applied only to detect deceptive read destructive faults, which as shown in Section II-B has a high probability of occurrence due to process variations ($\sim$2% of cells as $\sigma_{\mathrm{vt0}} = 80$ mv). It relies on the proper March test sequence, such as Extended March C-, to detect all other types of faults.

In the remainder of this section, first the circuit design of the double sensing technique is discussed. Then the sizing of some critical transistors is explained. Simulation is performed to validate the benefits of the double sensing technique. Finally, the overhead of the technique is evaluated.

A. Double Sensing Circuit

The basic idea of double sensing is to have parallel SAs to sample the bit-lines twice during one read cycle. Fig. 5 shows the SRAM array with the embedded double sensing circuit. There are two major modifications. First, two PMOSs per column are inserted, which is highlighted in Fig. 5 as mismatch generation PMOSs. The function of these two PMOSs is to generate some upset which shakes the stability of the SRAM cell during the read operation. The detailed operation will be discussed later in this section. Second, another SA per column is attached to bit-lines, as shown in Fig. 5. In the test mode, the first SA is fired in the same way as the conventional scheme,

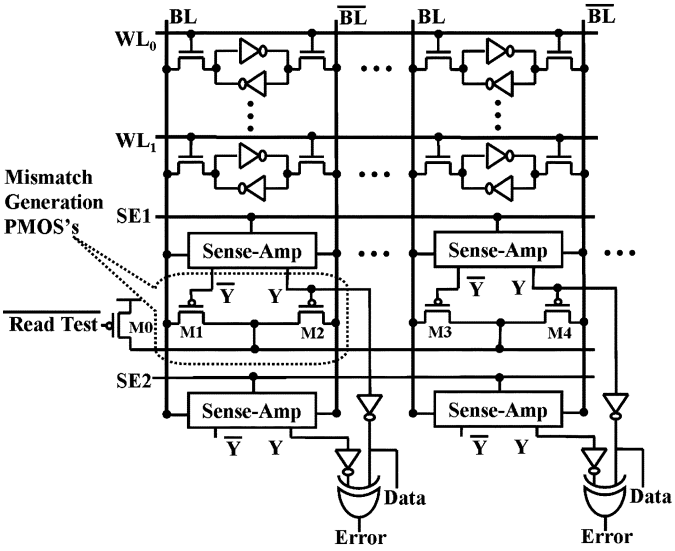


Fig. 5. Proposed double-sensing DFT circuit.

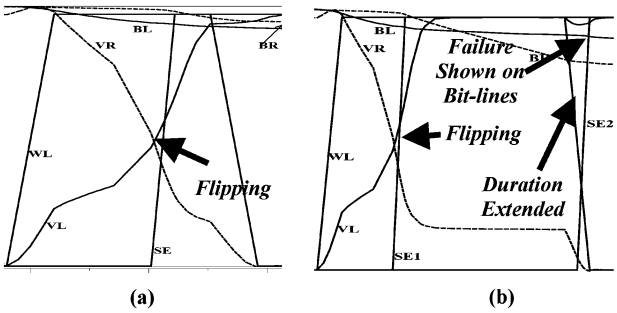


Fig. 6. (a) Flipping read failure waveforms with normal word-line timing. (b) Waveforms with the extended word-line duration.

while the second one is for delayed sensing. Comparing the output of the two SAs by an XOR gate shown in Fig. 5, it is possible to determine whether the SRAM cell has flipped or not during the read operation.

In SRAM array structures, there is a considerable amount of capacitance (transistor diffusion capacitances and interconnect capacitances) on bit-lines. Therefore, the bit-lines respond to the flipping of the SRAM cell slowly due to the considerable amount of charge required to develop an opposite bit-line differential. Hence, to achieve better detection capability, the added second SA has to be fired as late as possible during a read cycle. However, in conventional memory access timing, the word-line does not remain active much longer after firing of the SA (the first SA). Fig. 6(a) summarizes this scenario. As the waveform shows, the SRAM value flips (VR and VL flips) during the read operation. While, the differential gap on bit-lines (BR and BL) remains roughly the same at the end of word-line access time (WL goes down). Therefore, with the conventional SRAM timing, it is difficult to detect the failure by the second SA. To achieve better detection capability of the second SA, the SRAM access timing during the read test mode is modified to extend the word-line activation duration. More importantly, some disturbance is deliberately generated during the read test to shake the stability of the accessed SRAM cell. This accelerates the detection.

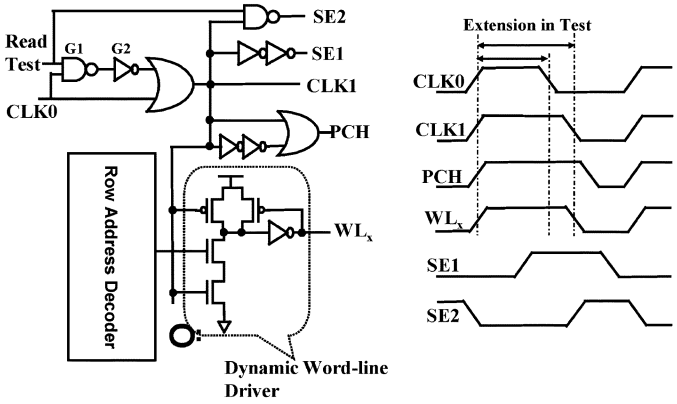


Fig. 7. Word-line extension circuit and timing diagrams.

Modification of SRAM Access Timing in the Read Test: In order to extend the word-line activation time, some modifications in word-line driver circuits are required. Fig. 7 shows the modified word-line driver and other control signal generation circuits. In the read test mode (in Fig. 7, Read Test is high), delay elements (G1 and G2) are activated. The system clock (CLK0) is OR-ed with the delayed CLK0 to generate the local clock (CLK1). This modifies the duty cycle of CLK1, which finally extends the word-line activation duration. The word-line (WLx) is generated by a dynamic driver buffering the row address decoder output and is clocked by the local clock (CLK1). The amount of word-line extension can be adjusted by changing the size of gates G1 and G2. The generation of the other control signals is also shown in Fig. 7. The local clock (CLK1) is OR-ed with its own delayed version to generate the precharge enable signal (PCH). When PCH is low, the bit-lines are precharged to get ready for the next read or write operation. CLK1 is delayed to enable the first SA (SE1). Also, in the test mode, SE2, which enables the second SA, is generated by inverting CLK1. Therefore, the second SA is fired right at the time when the word-line is disabled.

With the proposed circuit, the word-line extension is achieved to test for deceptive read and destructive fault. Fig. 6(b) shows the scenario when the word-line duration is extended and the SRAM value flips during the read operation. As shown in this waveform, due to the WL duration extension, the bit-lines (BL and BR) have developed enough opposite voltage differential, which can be sensed by the second SA. Therefore, the flip in the SRAM cell can be detected in one cycle.

Disturbance Generation: During the read test, some disturbance is generated by the mismatch generation PMOSs shown in Fig. 5. The disturbance shakes the stability of the accessed SRAM cell and pushes it to fail early if it is not sufficiently robust. Also, the inserted PMOSs accelerate the response time of the bit-lines by pulling up the already discharged bit-line if the SRAM flips during the read operation. The disturbance generation avoids excessive extension of the word-line activation duration, which is limited by the clock cycle time as well as by the drive strength of the precharge circuits.

The detailed operation of the disturbance generation is as follows. During the read operation in test, before the firing of the first SA, the SRAM read operation works exactly the same as the conventional scheme. After firing of the first SA, one of the

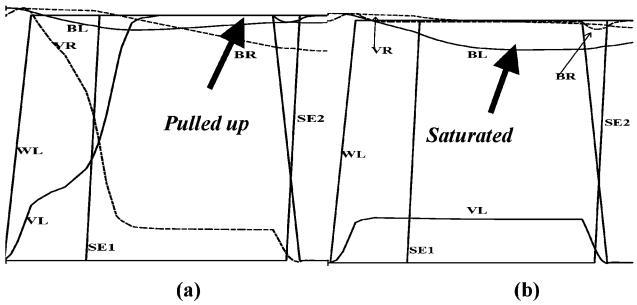


Fig. 8. Double-sensing waveforms of (a) a faulty cell and (b) a robust cell.

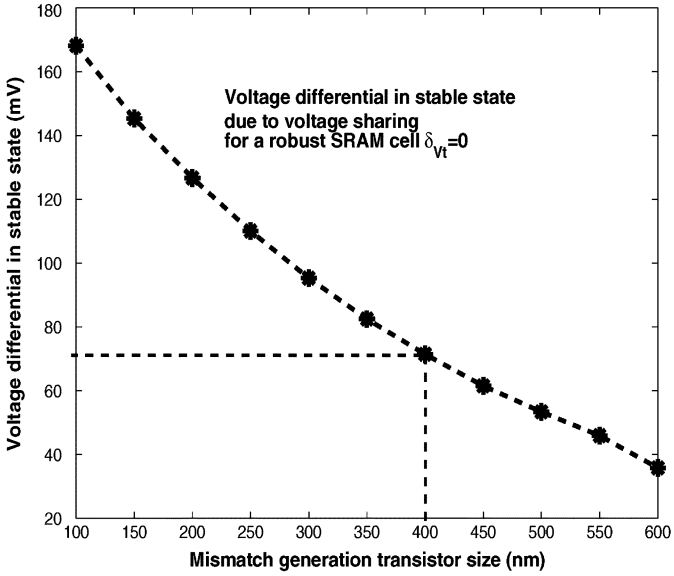


Fig. 9. Voltage differential versus transistor size (50-nm technology [14]).

mismatch generation PMOSs (M1 or M2 in Fig. 5) is turned on according to the output of the first SA. The turned-on PMOS disturbs the SRAM cell by injecting some extra current to the already discharged bit-line and the internal node of the accessed cell storing the value "0." Therefore, it pushes the SRAM cell to flip if it is not sufficiently robust. If the SRAM cell flips during the read operation, the initial discharged bit-line is pulled up by the turned-on mismatch generation PMOS. The other bit-line is discharged through the pull-down network within the accessed SRAM cell because the internal voltage corresponding to this bit-line is flipped to "0" now. This scenario is shown in Fig. 8(a). As the waveform shows, the initial discharged bit-line (BL) is pulled up soon after the SRAM cell flips. The other bit-line (BR) starts to discharge after the cell flips. Therefore, with the two bit-lines going toward opposite directions, the detection of deceptive read destructive faults is accelerated. On the other hand, if the SRAM cell is stable during read operation, the generated disturbance will not flip the cell value. The turned-on PMOS forms a voltage sharing with the already discharged bit-line and the pull-down network within the access SRAM cell. Therefore, the voltage value of the discharged bit-line saturates at some intermediate value depending on the strength of the turned-on PMOS. This scenario is shown in Fig. 8(b), where a robust SRAM is accessed. As the waveform shows, the initially discharged bit-line (BL) continues to discharge with the turning on of the disturbance generation. Finally, the bit-line (BL) settles down at some voltage value.

In the normal mode of operation, the disturbance generation is turned off by shutting down the supply of the mismatch generation PMOSs. In Fig. 5, M0 disconnects the PMOSs from the supply when the READ TEST signal is disabled.

B. Proper Sizing of the Double-Sensing Circuit

From the above discussion, it is observed that the strength of the mismatch generation transistors (M1 and M2) is critical to the correct functionality of the DFT circuit. A larger mismatch generation transistor generates a stronger upset to the accessed SRAM cell. Also, if the cell flips during test, a larger mismatch generation transistor pulls up the initially discharged bit-line faster. On the other hand, if a robust cell is accessed, due to voltage sharing, a larger mismatch generation transistor may significantly reduce the voltage differential between the two bit-lines, as shown in Fig. 8(b).

To avoid oversizing the mismatch generation PMOSs, simulations are performed to see the impact of the PMOS size on the saturated bit-line voltage differential. The simulation is conducted with a predicted 50-nm technology [14]. Fig. 9 shows how the saturated bit-line voltage differential reduces as the size of the mismatch generation transistor increases when a robust SRAM cell is accessed. In Fig. 9, the X axis shows the size of the mismatch generation PMOSs from 100 to 600 nm. The Y axis shows the saturated bit-line voltage differential when a robust SRAM cell is accessed. As the mismatch generation transistor gets larger, there is less bit-line voltage differential for the second SA input. Therefore, the output of the second SA is less immune to noise or SA offset voltage variations. To ensure a sufficiently large bit-line voltage differential at the time of firing the second SA, we have chosen the size of the mismatch generation PMOS to be 400 nm, as shown in Fig. 9 in the following simulations.

In real test applications, the size of the mismatch-generation PMOSs also determines how conservative the test is. Intuitively, a strong disturbance generated by large mismatch generation PMOSs will push the marginally vulnerable SRAM cells to fail in the test. However, those cells may not fail in the normal operation mode without the disturbance. Therefore, the optimal size should be carefully designed considering the technology and the test criteria. It should be noted here, based on our simulation, that the generated disturbance does not increase the occurrence of flipping read failures significantly.

C. SRAM Test With the Double Sensing Technique

Combining the idea of word-line duration extension and disturbance generation, simulations are performed to evaluate the benefits of the double sensing technique. With the chosen size for the mismatch generation transistors, Monte Carlo simulations are performed to observe how the detection capability of double sensing improves with extension of the word-line activation duration. In our experiments, the word-line activation duration for the normal operation mode is assumed to be 150 ps. As shown in Fig. 10, for a specific distribution of transistor V_t variation ($\sigma_{\mathrm{vt0}} = 80$ mv, which is a conservative value for

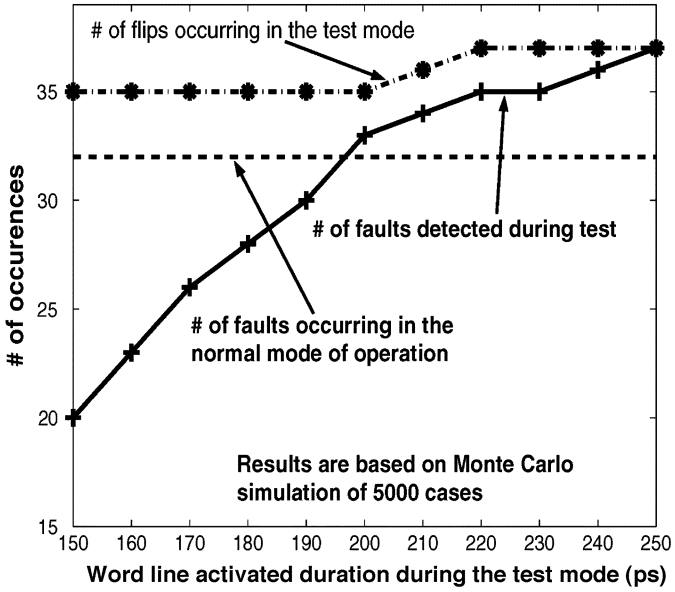


Fig. 10. Detection ability versus word-line extension.

50-nm technology [14]), among 5000 cases, there are 32 deceptive read and destructive faults in the normal operation mode. In the test mode, with the disturbance generation turned on and the word-line extended, there is a negligible increase (three out of 5000 cases) in the occurrence of deceptive read destructive faults. This increase is mainly due to the introduction of the disturbance. The number of fault occurrences is insensitive to the extension of the word-line duration, as shown in Fig. 10. This is due to the fact that, if an SRAM cell is robust, the voltage sharing between the bit-line and the zero-storing internal node of the cell does not lead to a flipping, no matter how long the word-line remains high. However, if the cell is potentially faulty, the voltage of the internal node storing "0" will increase above the cell trip point soon after the word-line is activated. On the other hand, as observed from Fig. 10, the extension of the word-line duration greatly improves the detection capability of double sensing. With the chosen size for the mismatch generation transistor, a reasonable extension (from 150 to 200 ps in Fig. 10) is enough to detect all of the deceptive read destructive faults within a single read operation during test.

D. Overhead of Double Sensing

To estimate the overhead of the proposed DFT circuit on memory performance, power, and area, a SRAM block of 128 rows and 32 columns with one SA per column is considered in a 50-nm predictive process technology [14]. Table II summarizes the overhead of the proposed DFT circuit in normal mode of operation.

The memory access time has several components, including address decoder delay, word-line driver delay, memory cell access delay, SA, and output driver delay. The proposed DFT has no impact on address decoder and word-line driver delay. However, double sensing slightly increases the capacitance of the bit-lines due to gate capacitance of the second SA inputs

TABLE II
OVERHEAD OF THE PROPOSED DFT CIRCUIT

	Access time (ps)	Power (mW)			Area (# of trans.)
		Read	Write	Standby	
Without DFT	300	53.10	55.08	51.61	24896
With proposed DFT	308	53.12	55.11	51.61	25184
Overhead (%)	2.7	0.04	0.06	0	1.15

and diffusion capacitances of the mismatch-generation PMOSs. This extra loading increases the SRAM cell access delay. The DFT circuit also adds some extra capacitance on the output of the first SA. Considering the percentage contribution of each of these delays in the overall memory access time, the overall delay increase is only 2.7%. The power overhead is due to the extra capacitive load of the double-sensing circuit in the normal mode of operation. However, a significant fraction of total memory power is due to leakage caused by a large number of idle memory cells, which is not affected by the proposed DFT circuit. The power overhead values are shown in Table II for read/write operations and standby mode. The area overhead of the proposed DFT is also small (1.15%) since the area of a memory is dominated by the SRAM cells. The power and area estimation results shown here exclude the I/O circuits and the decoders. Therefore, if the whole memory system is considered, since the I/O circuit and the decoders are not affected by the DFT, the percentage of power and area overhead is even less.

V. APPLICATION OF DOUBLE SENSING

The double-sensing technique is applied mainly to detect deceptive read destructive faults, which achieves a test time reduction. Also, in the normal operation mode, double sensing can be activated to complement the error correction code (ECC) scheme in order to detect random read faults. In this section, we have discussed the applications of the double sensing technique.

A. Test Time Improvement by Double Sensing

As discussed in the previous section, the double-sensing technique detects all deceptive read destructive faults within one read operation during test. Cooperating with the proper March test sequences, such as Extended March C-, the proposed double-sensing technique simplifies the consecutive read operations performed on SRAM cells to detect deceptive read destructive faults. This reduces the test time of Extended March C- to $10N$, which is an improvement of 17% in test time. Compared with March SR, which has a similar fault coverage as Extended March C-, using both the proposed Extended March C- and the double-sensing DFT, the test time is reduced from $14N$ to $10N$, which corresponds to an improvement of 29% in test time. Moreover, a better fault coverage is achieved by detecting address decoder faults that are ignored in March SR.

B. Online Detection and Correction of Random Read Faults Using Double Sensing

During memory manufacturing test, random read faults are detected probabilistically, which implies that some of the faults remain undetected. Hence, during the normal mode of operation, random read faults may still occur. Furthermore, the occurrence of random read fault strongly depends on the supply voltage level. If the supply voltage is relatively low at the time the memory cell is accessed (for example, due to some di/dt voltage drop in the supply line), the voltage differential on bit-lines may not be large enough as designed. The reduced voltage differential is less immune to noise, which may lead to an unpredictable SA output. This is the failure mechanism for random read fault. Conventionally, ECC is utilized to detect/correct these faults as well as soft errors [18]. However, the detection and correction capability of the most prevalently used ECC, namely SECDED (Single Error Correction and Double Error Detection), is limited to 1 b of correction and 2 b of detection, with six parity bits per 32 b in a word [18]. When the ECC scheme detects two or more error bits in a word, the ECC scheme fails and the current data block has to be refreshed from higher level cache or main memory. This process takes a significant amount of time [18]. Furthermore, with technology scaling, the soft error rate within the memory subsystem increases [18]. Therefore, an online testing technique should be applied to detect and correct random read faults more effectively to avoid a more expensive ECC scheme.

For the SRAM array used in embedded memories, the timing requirements are less stringent. The word-line remains activated for some amount of time after the firing of the SA [19]. With such SRAM timing, the proposed double-sensing DFT circuit (Fig. 5) can also be utilized in the online mode to detect and correct errors due to random read faults. This considerably improves the detection and correction capability of ECC. During the normal mode of operation ($\overline{\text{Read Test}}$ is set to V_{DD} in Fig. 5), the mismatch-generation PMOS transistors are disabled. Thus, the proposed circuit simply samples the bit-lines twice in a read operation if the second SA is enabled (SE2 signal is still generated in Fig. 7).

In the cases of random read faults, the access transistors are not strong enough to discharge the bit-line so as to develop enough voltage differentials. Therefore, SA input differentials may be disturbed by noise. If the firing of the SA (SE1) is delayed during the word-line activated duration, a longer time is provided for the bit-lines to be discharged. Therefore, more voltage differential is developed on bit-lines to ensure correct function of the SAs. However, the memory performance is degraded due to the delay of the SA firing. This is not acceptable in the design.

With the proposed double-sensing technique, the first SA is still fired to maintain the performance of the memory. As shown in Fig. 11, the output of the first SA is fed into the ECC checker and the memory data bus. Meanwhile, the output of the second SA is stored in a buffer. If the ECC checker detects that there are two errors (and therefore fails to correct the errors) or if there is more than one error flag generated by the proposed DFT circuit (Fig. 5), the content of the buffer (where the output of the second SA is stored) is fed to the ECC checker and the data bus in the next cycle. Since the output generated by the second SA is less susceptible to errors, there is a high probability that the system receives the correct data only with a delay penalty of one cycle. Therefore, with the proposed scheme, the tolerable fault density (due to random read faults) increases considerably.

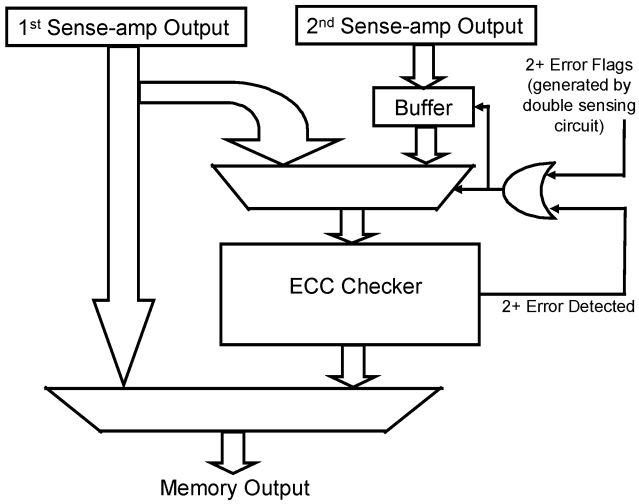


Fig. 11. Online detection and correction with double sensing.

TABLE III
MARCH TEST SEQUENCES

March C- [17]:	⇳ (W0) ⇑ (R0 W1) ⇑ (R1 W0) ⇓ (R0 W1) ⇓ (R1 W0) ⇳ (R0)
March SR [5]:	⇓ (W0) ⇑ (R0,W1,R1,W0) ⇓ (R0,R0) ⇑ (W1) ⇓ (R1,W0,R0,W1) ⇑ (R1,R1)
March B [17]:	⇳ (W0) ⇑ (R0,W1,R1,W0,R0,W1) ⇑ (R1,W0,W1) ⇓ (R1,W0,W1,W0) ⇓ (R0, W1,W0)

VI. CONCLUSION

With technology scaling, process variations result in functional failures in memory systems. In this work, physical failure mechanisms in SRAM due to individual transistor V_t variations are analyzed and classified into the established logic fault models. March test sequences are compared and optimized to target these emerging failure mechanisms. In addition, a DFT circuit based on double sensing of bit-lines is proposed to complement the March test so as to minimize the test time. With both the Double Sensing circuit and the proposed Extended March C- test sequence, the memory test time is reduced by 29% compared to the existing test method with similar fault coverage. The design overhead of the double-sensing circuit is negligible. Also, the double-sensing circuit improves the reliability and performance of the memory subsystem if it is used in online testing.

APPENDIX A

The March test sequences used for comparison in Section III are summarized in Table III.

REFERENCES

[1] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. Design Automation Conf.*, Jun. 2003, pp. 338–342.

[2] *The National Roadmap for Semiconductors*, 2000. Semiconductor Industry Association.

[3] Q. Chen, H. Mahmoodi, S. Bhunia, and K. Roy, "Modeling and testing of SRAM for new failure mechanisms due to process variations in nanoscale CMOS," in *Proc. VLSI Test Symp.*, sec. 8b, May 2005, pp. 292–297.

[4] S. Hamdioui, Z. Al-Ars, A. J. Van De Goor, and M. Rodgers, "Linked faults in random access memories: Concept, fault models, test algorithms, and industrial results," *IEEE Trans. Computer-Aided Design Integr. Circuits Syst.*, vol. 23, no. 5, pp. 737–757, May 2004.

[5] S. Hamdioui and A. J. Van De Goor, "Experimental analysis of spot defects in SRAMs: Realistic fault models and tests," in *Proc. 9th Asian Test Symp.*, Dec. 2000, pp. 131–138.

[6] ——, "Efficient tests for realistic faults in dual-port SRAMs," *IEEE Trans. Comput.*, vol. 51, no. 5, pp. 460–473, May 2002.

[7] S. Mukhopadhyay, H. Mahmoodi-Meimand, and K. Roy, "Modeling and estimation of failure probability due to parameter variations in nanoscale SRAM's for yield enhancement," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2004, pp. 64–67.

[8] A. Agarwal, B. Paul, and K. Roy, "A novel fault tolerant cache to improve yield in nanometer technologies," in *Proc. 10th IEEE On-Line Testing Symp.*, Jul. 2004, pp. 149–154.

[9] S. J. Lovett, G. A. Gibbs, and A. Pancholy, "Yield and matching implications for static RAM memory array sense-amplifier design," *IEEE J. Solid State Circuits*, vol. 35, no. 8, pp. 1200–1204, Aug. 2000.

[10] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE J. Solid State Circuits*, vol. 39, no. 7, pp. 1148–1158, Jul. 2004.

[11] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE J. Solid State Circuits*, vol. 36, no. 4, pp. 658–665, Apr. 2001.

[12] Y. Taur and T. K. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[13] C. Neau, "Modified MIT 50 nm devices," Ph.D. dissertation, Sch. Electric. Comput. Eng. Purdue Univ., West Lafayette, IN, 2004.

[14] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Statistical design and optimization of SRAM cell for yield enhancement," in *Proc. Int. Conf. Computer Aided Design*, Nov. 2004, pp. 10–13.

[15] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *Proc. Int. Conf. Computer Aided Design*, Nov. 2003, pp. 900–907.

[16] H. Chang and S. S. Sapatnekar, "Statistical timing analysis considering spatial correlations using a single PERT-like traversal," in *Proc. Int. Conf. Computer Aided Design*, Nov. 2003, pp. 621–625.

[17] M. L. Bushnell, *Essentials of Electronic Testing for Digital, Memory, and Mixed-Signal VLSI Circuits*. Norwell, MA: Kluwer, 2000.

[18] V. Degalahal, R. Ramanarayanan, N. Vijaykrishnan, Y. Xie, and M. J. Irwin, "The effect of threshold voltages on the soft error rate," in *Proc. 5th Int. Symp. Quality Electronic Design*, 2004, pp. 503–508.

[19] D. Malone, P. Bunce, J. DellaPietro, J. Davis, J. Dawson, T. Knips, D. Plass, P. Pritzlaff, and K. Reyer, "Design validation of 0.18 um 1 GHz cache and register arrays," in *Proc. Custom Integrated Circuit Conf.*, May 2000, pp. 295–298.

Qikai Chen (S'05) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 2003. He is currently working toward the Ph.D. degree at Purdue University, West Lafayette, IN.

He was with Micron Technology in the summer of 2005, working on flash memory. His research interests include SRAM design and test, low-power and robust circuit design using submicrometer CMOS, and circuit/device co-design in the nanometer regime.

Mr. Chen was a recipient of the Ross Fellowship in 2003.

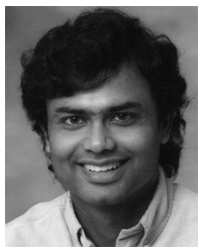
Hamid Mahmoodi (S'00) received the B.S. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 1998, the M.S. degree in electrical and computer engineering from the University of Tehran, Tehran, Iran, in 2000, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2005.

He joined the Electrical and Computer Engineering faculty, San Francisco State University, San Francisco, CA, in 2005, where he is currently an Assistant Professor His research interests include low-power, robust, and high-performance circuit design for nanoscale technologies. He has more than 45 refereed publications in journals and conferences.

Dr. Mahmoodi was a recipient of the Best Paper Award of the 2004 International Conference on Computer Design.

Swarup Bhunia (S'00) received the B.S. degree from Jadavpur University, Calcutta, India, and the M.S. degree from the Indian Institute of Technology (IIT), Kharagpur. He is currently working toward the Ph.D. degree at Purdue University, West Lafayette, IN.

He has worked in the EDA industry on RTL synthesis and verification for approximately three years. His research interests include design methodologies for high-performance low-power testable VLSI system, defect-based testing, noise analysis, and noise-aware design.



Kaushik Roy (S'83–M'90–SM'95–F'02) received the B.Tech. degree in electronics and electrical communications engineering from the Indian Institute of Technology, Kharagpur, India, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 1990.

He was with the Semiconductor Process and Design Center of Texas Instruments, Dallas, TX, where he was involved with FPGA architecture development and low-power circuit design. He joined the Electrical and Computer Engineering Faculty, Purdue University, West Lafayette, IN, in 1993, where he is currently a Professor and holds the Roscoe H. George Professor of Electrical and Computer Engineering Chair. His research interests include VLSI design/computer-aided design for nanoscale silicon and nonsilicon technologies, low-power electronics for portable computing and wireless communications, VLSI testing and verification, and reconfigurable computing. He has published more than 300 papers in refereed journals and conferences, holds eight patents, and is a coauthor of two books on low-power CMOS VLSI design. He is the Chief Technical Advisor of Zenasis Inc. and a Research Visionary Board Member of Motorola Laboratories (2002). He was a Guest Editor for a special issue of the *IEE Proceedings Computers and Digital Techniques* (July 2002).

Dr. Roy was the recipient of the National Science Foundation Career Development Award in 1995, the IBM Faculty Partnership Award, the AT&T/Lucent Foundation Award, the 2005 SRC Technical Excellence Award, the SRC Inventors Award, and Best Paper Awards at the 1997 International Test Conference, the 2000 IEEE International Symposium on Quality of IC Design, and the 2005 IEEE Circuits and Systems Society Outstanding Young Author Award. He has been on the Editorial Board of *IEEE Design and Test,* the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He was the Gest Editor for a Special Issue on Low-Power VLSI of *IEEE Design and Test* (1994) and the IEEE TRANSACTIONS ON VLSI SYSTEMS (June 2000).