Modeling and Testing of SRAM for New Failure Mechanisms due to Process Variations in Nanoscale CMOS

Qikai Chen, Hamid Mahmoodi, Swarup Bhunia, and Kaushik Roy
School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
{qikaichen, mahmoodi, bhunias, kaushik}@purdue.edu

Abstract

In this paper, we have made a complete analysis of the emerging SRAM failure mechanisms due to process variations and mapped them to fault models. We have proposed two efficient test solutions for the process variation related failures in SRAM: (a) modification of March sequence, and (b) a novel low-overhead DFT circuit to complement the March test for an overall test time reduction of 29%, compared to the existing test technique with similar fault coverage.

Keywords: DFT, Failure Mechanism, March Test, Process Variation, SRAM

1. INTRODUCTION

As silicon industry moves towards the end of the technology roadmap, controlling the fabrication of scaled devices is becoming a great challenge. Both inter-die and intra-die device parameter variations are expected to be significantly large in sub-90nm technology generations [1]. Intra-die variations due to random dopant fluctuations result in parameter mismatch for transistors in a die [1]. These atomic-level fluctuations are most pronounced in minimum-geometry transistors commonly used in area-constrained circuits such as memories.

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 primarily due to inter-die variations. With random dopant fluctuations, similar transistors on a die may have different strengths (different threshold voltages). Such variations affect the stability of 6-transistor SRAM cells (Fig. 1). Moreover, the variations also affect proper functionality of sense-amplifiers. Hence, design and test of SRAM in scaled technologies are emerging as major challenges. To develop efficient memory test solutions in nanoscale regimes, a complete investigation of failure mechanisms (resulting from inter-die and intra-die process variations) in a memory subsystem is required. Conventionally, March test is the prevalent test technique for SRAM [3-5]. However, alternative or complementary low-cost test methods need to be developed to detect and diagnose emerging failures of memory subsystems in nanoscale CMOS.

In this paper, we have analyzed physical mechanisms of logic faults under process variations in all major components of a memory subsystem and presented efficient techniques to test for these failures. In particular, the main contributions are:

- A complete analysis of emerging failure mechanisms under process variations in a memory subsystem and development of required logic level fault models (Section 2).
- Development of two optimized March test sequences to achieve better fault coverage (considering the failures due to process variations) with minimum increase in test time (Section 3).
- A novel design for test (DFT) circuit, referred to as *Double Sensing*, to reduce test application time for March test with minimal overhead (Section 4).

2. FAILURE MECHANISMS AND FAULT MODELS

To compare logical behavior of faulty memories against good ones, modeling the physical failure mechanisms as logic level fault models is required. This section first investigates the failure mechanisms under process variations in a memory system, comprising SRAM cells, sense-amplifiers, and address decoders. Then, those physical failure mechanisms are mapped to the logic level fault models.

In [3], 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*; *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* and *Incorrect Read Coupling Fault*).

2.1 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, 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 [7]. Therefore, in this work, we focus on V_t variations. V_t's of transistors are considered to be independent random variables. The V_t shift is considered to be a zero mean Gaussian distribution, with [8]:

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

where, σvt0 is the standard deviation of Vt shift of a minimum-sized transistor. σvt0 depends on the doping concentration and the oxide thickness [8]. L_{min} and W_{min} are the minimal length and width of a transistor in the corresponding technology.

- *SRAM Cell Failure Mechanisms*

Process variations in SRAM cells (Fig. 1) may result in [6]:

1. A decrease in the current (due to weak, i.e. high V_t, access transistors, i.e. AXL and AXR in Fig. 1) that discharges the bit-lines through the access transistors during the read operation. This mechanism leads to less voltage difference between the bit-lines when the sense-amplifier samples them, which may result in a wrong evaluation in the sense-amplifier. We refer to this failure as *access failure*.
2. Increased disturbance to the cell (due to strong, i.e. low V_t, access transistors) during the read operation, which flips the cell value. This failure is referred as *flipping read failure*.

Proceedings of the 23rd IEEE VLSI Test Symposium (VTS'05)
1093-0167/05 $ 20.00 **IEEE**

IEEE COMPUTER SOCIETY

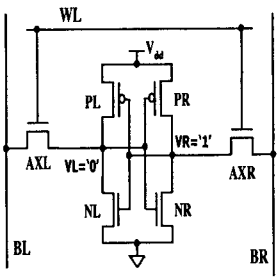

Fig. 1: 6T SRAM cell

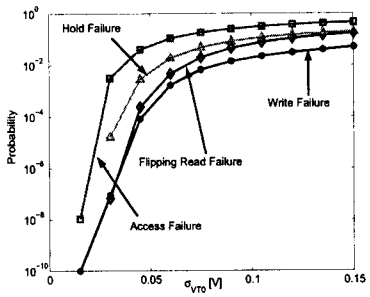

Fig. 2: Failure probability vs. σ_{vt0} [6]

3. Unsuccessful write operation due to the deviation of the strength of the access transistors and the trip point of the cross-coupled inverters. We refer to this failure as *write failure*.
4. Instability of an SRAM cell in holding its content even when it is not accessed, due to excessive mismatch in the cross-coupled inverters (especially when the supply voltage is lowered in standby mode 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 variations (σ_{vt0}) are plotted [6]. The probabilities are estimated by Monte Carlo simulations for a 50nm predictive technology [10]. As shown in Fig. 2, the failures are more likely to occur with increasing V_t variations.

- *Sense-amplifier Failure Mechanisms*

Functional failures in sense-amplifier (Fig. 3) 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 sense-amplifier functional failure. Unbalanced V_t's of M2 and M3 modify the amount of current on each side, which discharges the node Y and $\overline{Y}$ during the evaluation period. This causes a shift in the offset voltage (the minimum voltage difference between two inputs for the sense-amplifier to evaluate correctly). Fig. 4 shows the distribution of the offset voltage of a sense-amplifier and the distribution of the voltage difference between the two bit-lines at the time when they are sampled by the sense-amplifier. As shown in Fig. 4, under process variations, although some of the SRAM cells can develop enough voltage difference on bit-lines during read operation, the sense-amplifier output can still be wrong due to offset voltage shift. The probability that the output of a sense-amplifier (shared by a column) is incorrect can be estimated as:

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

where, $f_{SenseAmp}(x)$ is the probability density function (PDF) of the distribution of the sense-amplifier offset voltage; $F_{SRAMCell}(x)$ is the cummulative density function (CDF) of the distribution of the bit-line voltage difference; and n is the number of SRAM cells per column. This PDF or CDF can be evaluated by Monte-Carlo simulations or semi-analytical methods as described in [6].

- *Address Decoder Failure Mechanisms*

Under process variations, 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. Probability that the address decoder will fail to meet the delay target can be obtained from statistical delay distribution of the decoder.

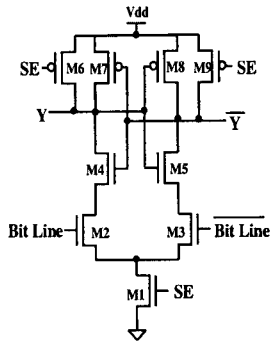

Fig. 3: Sense-amplifier

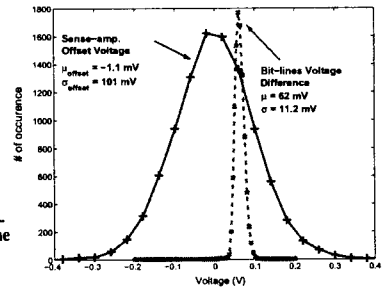

Fig. 4: Distributions of sense-amp offset & bit-line voltage difference (σ_{vt0} = 80mv; 50nm process [10])

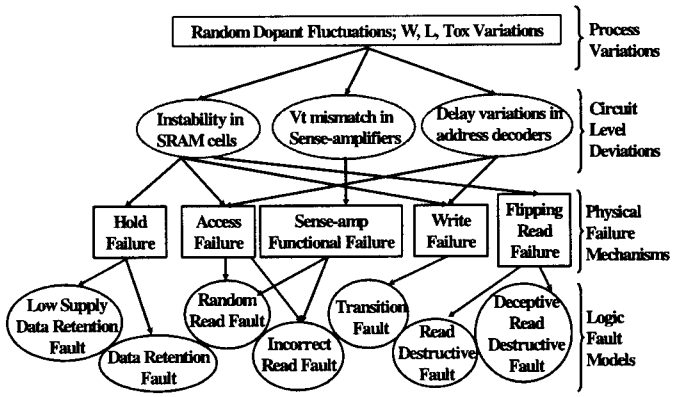

Fig. 5: Failure mechanisms and logic fault models

2.2 Fault Models

Fig. 5 summarizes the failure mechanisms in a memory under process variations and shows mapping of the failures to the established logic fault models. Among the failure mechanisms: (1) SRAM access failures and sense-amplier 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*, according to the time the cell flips; (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 as *low supply data retention fault* to describe flipping failures occuring due to application of low supply voltage in the standby mode when the memory is not accessed.

By mapping the process variation related failure mechanisms to logic fault models, memory test can be designed to target failures in nano-scale SRAMs. According to our simulation results (Fig. 2), the flipping read failure in SRAM cells has a high probability of occurrence (~2% of cells). Most of the flipping read failures show themselves as *deceptive read destructive faults*, which is overlooked in conventional memory test. Moreover, with process variations, *low supply data retention faults* are very likely to occur (about 4%, as shown in Fig. 2). This type of faults is also not detecatble with a conventaional March test. Therefore, in the remainder of this paper, we propose techniques to address the detection of these logical faults efficiently through optimization of March test and a novel low-cost DFT circuit.

IEEE COMPUTER SOCIETY

3. MARCH TEST OPTIMIZATION

March test is prevalently applied to SRAM test [9]. 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 these operations of one March element are applied to a certain address before proceeding to the next address, either in an increasing (denoted by ⇑ before a March element) or a decreasing (denoted by ⇓) order.

To cover the traditional single cell and coupling fault models, March C- is popularly used as a base sequence [9]. 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 intra-die process variations. In [4], a test sequence named March SR is proposed that covers the *deceptive read destructive faults*. However, March SR ignores address decoder faults. Furthermore, the sequence has a test time of 14N (N being 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 do not cover *data retention faults* and *low supply data retention faults*. However, 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 Vt mismatches resulting 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 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 lower supply for a specified time (at least one cycle); and then bring it back to normal supply. The March C- sequence with the above-mentioned modifications is presented as follows:

Optimized March C- : ⇕ (W0) ⇑ (R0 W1) ⇑ (R1 W0) ⇓ (R0 W1) (HOLD) ⇓ (R1 R1 W0) (HOLD) ⇕ (R0 R0)

We refer to this proposed sequence as *Optimized March C-*. The optimized March C- has a test time of 12N, resulting in an improvement of 15% compared to March SR [4]. We assume that the time required for *HOLD* is insignificant compared to the time required for one March operation to all the addresses. Furthermore, the optimized March C- has better fault coverage than March SR (see Table 1), since it is capable of detecting address decoder faults.

In order to maintain the test time of March C- (10N) 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:

March Q : ⇓ (W0) (HOLD) ⇑ (R0 W0 W1 R1) (HOLD) ⇑ (R1 W1 W0 R0) ⇓ (R0)

Table 1 compares the fault coverage and test time of March C-, March SR, March B [9], optimized March C-, and March Q. "+" or "-" denote whether the March sequence is able to cover a logic fault model or not, and "+-" means that the test sequence can cover a fault model with some probability. From Table 1, it can be observed that the optimized 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 can statistically detect only half of the *transition coupling faults* and is not capable of detecting *address decoder faults* (like March SR). Hence, with similar fault coverage as March SR, the proposed March Q sequence has a 30% less test time. We conclude that if *transition coupling faults* are not of major concerns, March Q can be a promising test sequence in scaled technologies.

Table 1: 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	+	+	+	+	+
Random Read Fault	+-	+-	+-	+-	+-
Read Destructive Fault	+	+	+	+	+
Deceptive Read Destructive Fault	-	-	+	+	+
Incorrect Read Fault	+	+	+	+	+
State Coupling Fault	+	-	+	+	+
Disturb Coupling Fault	+	-	+	+	+
Incorrect Read Coupling Fault	+	-	+	+	+
Read Destructive Coupling Fault	+	-	+	+	+
Transition Coupling Fault	+	+-	+	+	+-
Test Time	**10N**	**17N**	**14N**	**12N**	**10N**

Optimization of March test through algorithmic techniques to minimize the overhead in test time while achieving the best possible fault coverage is a challenging problem. Therefore, novel DFT techniques to complement conventional March test can be promising not only to have a better fault coverage but also to minimize the test time. In the following section, we propose a DFT circuit to complement March sequences for minimization of test time while maintaining the best fault coverage.

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

As discussed in the previous section, under process variations, 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. Using this DFT with a proper March test sequence, such as optimized March C- or March SR, the consecutive read operations to detect *deceptive read destructive faults* can be replaced with a single read operation. Therefore, memory test time can be improved significantly.

IEEE COMPUTER SOCIETY

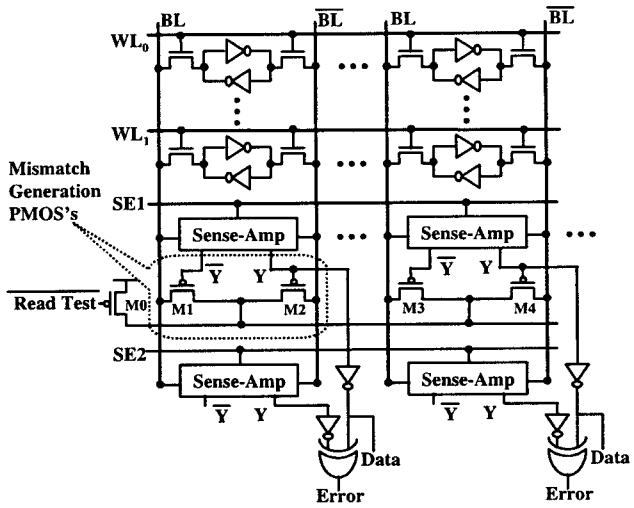

Fig. 6: Proposed double sensing DFT circuit

4.1 Double Sensing Circuit

The basic idea of double sensing is to have parallel sense-amplifiers to sample the bit-lines twice during a read cycle as shown in Fig. 6. The first sensing is performed in the same way as the conventional sensing, while the second one is a delayed sensing. The added second sense-amplifier 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 sense-amplifier. Furthermore, in SRAM structures, there is considerable amount of capacitance (transistor diffusion capacitances and interconnect capacitances) associated with bit-lines. Therefore, although flipping read failures occur during the time word-line is activated, in most cases, there is not enough time for bit-lines to get discharged so as to show the failures (*deceptive read and destructive faults*). Fig. 7(a) shows a flipping read failure during a normal SRAM read. In order to detect *deceptive read and destructive fault* within one cycle in the at-speed test mode, word-line duration needs to be extended (Fig. 7(b)). With the extended word-line, a second sense-amplifier is used to sample the bit-lines again at the end of the extended word-line duration in a read operation (SE2 in Fig. 7(b)).

In order to extend the word-line activation time, some modifications in word-line driver circuits are required. Fig. 8 shows the word-line extension and other control signal generation circuits. In the *read test* mode (Read Test, in Fig. 8, is high), the system clock (CLK0) is OR-ed with the delayed CLK0 to generate the local clock (CLK1). This modifies the duty cycle of CLK1, extending the word-line activation duration. The word-line (WL_x) 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. 8. CLK1 is inverted (in the test mode) to generate the enabling signal for the second sense-amplifier (SE2). Therefore, the second sense-amplifier 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*. Therefore, with the second sense-amplifier, bit-lines can be sampled twice in a read cycle to detect the fault (Fig. 7(b)). However, word-line extension requires upsized pre-charge transistors for the bit-lines to finish pre-charging within the reduced pre-charging time. The upsized

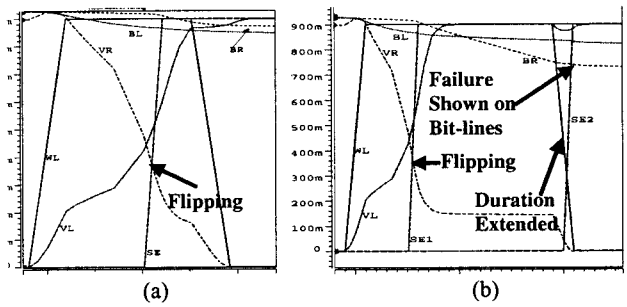

Fig 7: (a) Flipping read failure waveforms with normal word-line timing (b) waveforms with the extended word-line duration

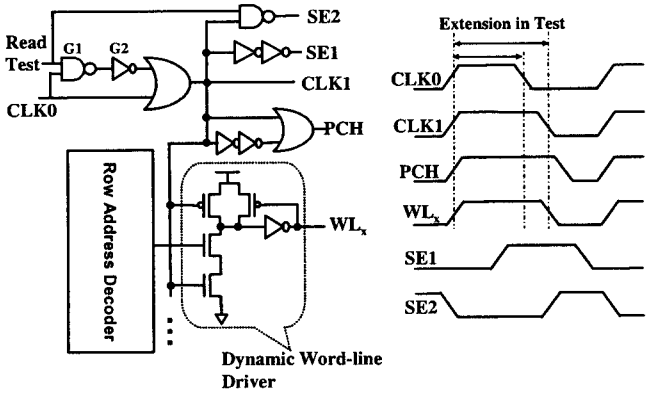

Fig. 8: Word-line extension circuit and timing diagrams

pre-charge transistors are therefore over-designed for the normal mode of operation (since in normal operation, the same transistors are still used to pre-charge without extending the word-line timing). This over-design results in area and power overhead. However, this overhead is reduced as the minimum required word-line extension is reduced. In order to reduce the required word-line extension, two mismatch generation transistors (per column) are introduced into the SRAM array ((M1 & M2) or (M3 & M4) in Fig. 6). The *double sensing* DFT circuit is shown in Fig. 6. Details of the *double sensing* circuit are further explained below.

During a read operation in the test mode, if a cell is affected by flipping read failure, the cell starts discharging the bit-lines in an opposite direction after it flips. If the flipped memory cell can establish enough voltage differentials on the bit-lines in an opposite direction by the time the second sense-amplifier is fired, the flipped data is detected. Therefore, comparing the outputs of the first and the second sense-amplifiers (with the XOR gate in Fig. 6), a *deceptive read destructive fault* is detected when a mismatch between the two outputs occurs. However, since there is considerable amount of capacitance associated with the bit-lines, it takes time for a flipped cell to change the voltage differential on the bit-lines to an opposite direction. In order to pull up the already discharged bit-line, a PMOS transistor is turned on, connecting this bit-line to VDD. This reduces the time required for the flipped cell to develop enough opposite voltage differentials on the bit-lines. The inserted mismatch generation transistors do not disturb the operation of the first sense-amplifier, because, before the firing of the first sense-amplifier, the outputs of the first sense-amplifier are pre-charged to VDD, and therefore, the PMOS mismatch transistors are both OFF. After the firing of the first-sense amplifier, M1 or M2 is turned on based on the result of the first sense-amplifier. For example, if a faulty SRAM cell storing '0' (VL='0' in Fig. 1) is accessed, the bit-line (BL in Fig. 9(a)) is initially discharged and has a lower voltage compared to the

IEEE COMPUTER SOCIETY

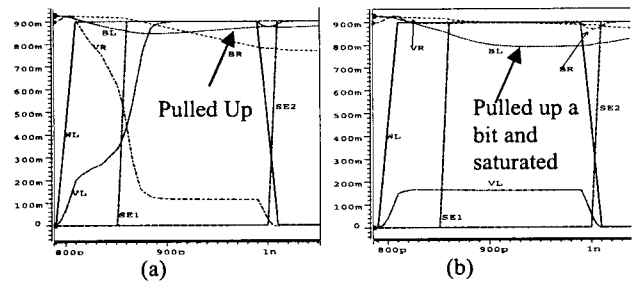

Fig. 9: Waveforms of (a) a faulty cell and (b) a robust cell

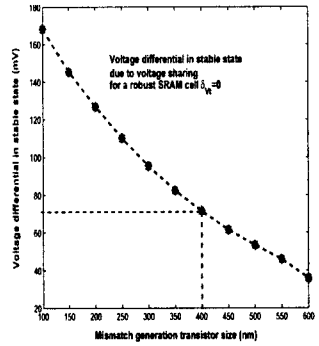

Fig. 10: Voltage differential vs. transistor size (50nm technology [10])

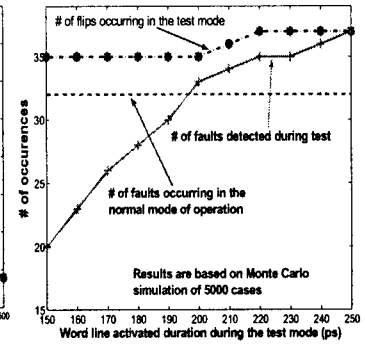

Fig. 11: Detection ability vs. word-line extension

bit-line (BR). The output $\overline{Y}$ of the first sense-amplifier switches to low, turning M1 (Fig. 6) ON. When M1 is ON, the bit-line (BL) is connected to VDD. Therefore, there is further disturbance on the accessed cell due to voltage sharing between the turned-on mismatch generation PMOS (M1 in Fig. 6) and the pull-down network in the memory cell (NL in Fig. 1). With the generated mismatch, if the memory cell is potentially faulty or unstable, the cell flips much earlier during the read cycle. When the cell flips, M1 will charge the bit-line (BL) back up to VDD while the bit-line (BR) gets discharged through the pull-down network within the accessed SRAM cell since the cell is now storing '1'. Fig. 9(a) shows the waveforms of the proposed circuit when a faulty SRAM cell is accessed. Since the voltages of the two bit-lines are going towards opposite directions; the development of opposite voltage differential is accelerated after the cell flips during read. Therefore, by using the mismatch generation transistors, the required amount of extension of the word-line duration to detect *deceptive read and destructive faults* is reduced.

On the other hand, if the cell is robust enough, the cell does not flip even after turning on the mismatch generation transistor. Fig. 9(b) shows the waveforms if a robust SRAM cell is accessed. For a robust cell (no flipping read failure), the memory cell continues to discharge the bit-lines in the same direction after firing of the first sense-amplifier. With the turning on of the mismatch generation transistor, the already discharged bit-line (BL in the case of Fig. 9(b)) may be pulled up a bit due to the voltage sharing effect; however, by properly sizing the mismatch generation transistors, the bit-line voltage differential at the time of firing the second sense-amplifier can be kept more than the differential at the time of firing the first sense-amplifier. This will be further discussed in the following section. Therefore, for a robust memory cell, the data read by the second sense-amplifier is the same as the output of the first sense-amplifier (no flipping during read), and the test algorithm compares the output value with the expected value to detect other types of faults in the memory. Hence, the proposed double sensing circuit enables detection of *deceptive read destructive faults* within a single cycle in at-speed test.

4.2 Proper Sizing of the DFT Circuit and Experimental Results

From the above discussion, we see that the strength of the mismatch generation transistors (M1 & M2) is critical to the correct function of the whole DFT circuit. A stronger mismatch generation transistor (larger size) can pull up the initially discharged bit-line faster (when the cell flips during the read operation). Hence, less extension is required for the word-line duration. On the other hand, due to voltage sharing, a large mismatch generation transistor significantly increases the voltage of the initially discharged bit-line (BL in Fig. 9(b)), if a robust cell is accessed. Therefore, the bit-line voltage differential at the time of firing the second sense-amplifier is reduced. This may result in an incorrect evaluation for the second sense-amplifier (voltage differential on bit-lines may be below the offset voltage of the second sense-amplifier.) Fig. 10 shows how the settled-down bit-line voltage differential reduces as the size of the mismatch generation transistor increases when a robust SRAM cell is accessed. As the mismatch generation transistor gets larger, there is less bit-line voltage differential for the second sense-amplifier input. Therefore, the output of the second sense-amplifier is less immune to noise or sense-amplifier offset voltage variations. To ensure the bit-line voltage differential to be large enough at the time of firing the second sense-amplifier, we have chosen the size of the mismatch generation PMOS to be 400nm as shown in Fig. 10. This sizing ensures a bit-line voltage differential of 70mv for the second sense-amplifier which is larger than the target bit-line voltage differential for the first sense-amplifier (50mv).

With an optimal size for the mismatch generation transistor, Monte-Carlo simulations are performed to observe how the detection ability of the *double sensing* improves with extension of the word-line activation duration. In our simulations, the word-line activation duration for the normal mode of operation is assumed to be 150ps. Simulation results (Fig. 11) show that for a specific distribution of transistor V_t variation (σ_{vt0} = 80mv, a conservative value for 50nm technology), among 5000 cases, there are 32 *deceptive read and destructive faults* in the normal mode of operation (mismatch generation transistors are turned off and the word-line duration is set at its normal value of 150ps). In the test mode, with the mismatch generation turned on and word-line duration extended, there is a negligible increase (0.06% of the cells tested) in the number of *deceptive read destructive fault* occurrences. This increase is mainly due to the introduction of the mismatch generation circuit, and the number of faults is insensitive to the extension of the word-line duration as shown in Fig. 11. That 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 zero-storing internal node of the cell will increase above the cell trip point soon after the word-line is activated.

On the other hand, as observed from Fig. 11, the extension of the word-line duration greatly improves the detection capability of double sensing. With the chosen size for the mismatch generation transistor (400nm), a reasonable extension (from 150ps to 200ps in Fig. 11) is enough to detect all the *deceptive read destructive faults*. In order to avoid a possible yield loss (i.e. test should not detect non-faulty memory cells under the normal mode of operation as faulty cells), an extended word-line duration that is just long enough to cover all the faults under normal mode of operation (200ps as shown in Fig. 11) is selected.

IEEE COMPUTER SOCIETY

4.3 Test Time Improvement by Double Sensing

By extending the word-line for read operations during test and optimal sizing of the mismatch transistors, all *deceptive read destructive faults* are detected. Cooperating with the proper March test sequence, such as March SR or optimized March C-, the proposed *double sensing* technique can remove the consecutive read operations performed on SRAM cells to detect *deceptive read destructive faults* (resulting from flipping read failures under process variations). This reduces the test time of March SR and optimized March C- to 12N and 10N, corresponding to an improvement of 14% and 17% in test time (when double sensing is used with March SR or optimized March C-), respectively. Compared with March SR, using both the proposed optimized March C- and the double sensing DFT, the test time is reduced from 14N to 10N with a corresponding improvement of 29% in test time. Moreover, a better fault coverage is achieved by detecting address decoder faults that are ignored in March SR. Therefore, with the proposed DFT circuit, memory test can be conducted more efficiently in terms of test time.

4.4 Failure Diagnosis

With double sensing, it is also possible to diagnose some of the failure mechanisms under process variations from the logic behavior of faulty cells during March test. An SRAM cell can be faulty due to *access failure*, which is manifested as less-than-enough voltage differential on bit-lines at the firing of the sense-amplifier. In this case, the sense-amplifier output depends considerably on the noise in bit-line voltage. In double sensing, since the second sense-amplifier is more reliable under noise, most often it reads the correct cell value. Hence, if the output of only the second sense amplifier matches with the expected value (the value written to the memory cell in the previous writing during the March test), we can infer that the failure is most likely caused by *access failure*. On the other hand, in most cases of a *flipping read failure*, the output of only the first sense amplifier matches with the expected value. However, if the memory cell is influenced by a *writing failure*, both sense-amplifier outputs are wrong (since the previous writing has not successfully written the correct value to the memory cell). Therefore, with the help of double sensing, some memory failure mechanisms can be identified during March test. With this diagnostic information, the manufacturing process of a memory design can be adapted so as to minimize occurrence of process-related failures in a memory sub-system. Therefore, the yield is improved.

4.5 Overhead of Double Sensing DFT circuit

To estimate the overhead of the proposed DFT circuit on memory performance, power, and area during the normal mode of operation, an SRAM block of 128 rows and 32 columns with one sense-amplifier per column is considered in a 50nm predictive process technology [10]. Table 2 shows 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 and bit-line delay, sense-amplifier and output driver delay. The proposed DFT has no impact on address decoder and word-line driver delay. *Double sensing* slightly increases the capacitance of the bit-lines due to gate capacitance of the second sense-amplifier inputs and diffusion capacitances of the mismatch generation transistors. This extra loading increases the cell and bit-line delay. The DFT circuit also adds some extra capacitance on the output of the first sense-amplifier. Considering the percentage contribution of each of these delays in the overall memory access time, the overall increase is only 2.7%. Therefore, the impact of the proposed DFT circuit on memory access time is negligible.

Table 2: Overhead of the DFT circuit for a 4k SRAM array

	Access time (ps)	Power (mW)			Area (# of trans.)
		Read	Write	Stdby	
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

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 large number of idle memory cells, which are not affected by the proposed DFT circuit. The power overhead values are shown in Table 2. The area overhead of the proposed DFT is also very small (1.15%) since the area of a memory is dominated by the memory 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. As observed from Table 2, the area and power overheads of the proposed DFT circuit are also negligible.

5. CONCLUSIONS

With technology scaling, process variations result in new functional failures in memory systems. In this work, SRAM physical failure mechanisms caused by process variations are analyzed and classified into the established logic fault models. March test sequences are compared and optimized to target the failure mechanisms introduced by process variations. In addition, a DFT circuit based on double sensing of bit-lines is proposed to cooperate with the March test so as to minimize the test time. With the *double sensing* circuit and the optimized March test sequence, the memory test time is reduced by 29% (compared to March SR) while all the faults induced by process variations are covered.

Acknowledgement

This work is sponsored in part by MARCO GSRC, National Science Foundation (NSF) and Semiconductor Research Cooperation (SRC).

References

[1] S. Borkar et al., "Parameter Variations and Impact on Circuits and Microarchitecture", *DAC 2003*, pp. 338-342.

[2] The National Roadmap for Semiconductors, *S.I.A.*, 2000.

[3] S. Hamdioui et al., "Linked Faults in Random Access Memories: Concept, Fault Models, Test Algorithms, and Industrial Results," *IEEE TCAD*, May 2004.

[4] S. Hamdioui et al., "Experimental analysis of spot defects in SRAMs: Realistic fault models and tests," *ATS*, pp. 131-138.

[5] S. Hamdioui et al., "Efficient tests for realistic faults in dual-port SRAMs," *IEEE Trans. on Computers*, May 2002.

[6] S. Mukhopadhyay et al., "Modeling and estimation of failure probability due to parameter variations in nano-scale SRAMs for yield enhancement", *VLSI Circuit Symposium*, June 2004.

[7] A. J. Bhavnagarwala et al., "The impact of intrinsic device fluctuations on CMOS SRAM cell stability", *JSSC*, April 2001.

[8] Y. Taur et al., "Fundamentals of Modern VLSI Devices", *Cambridge University Press*, 1998.

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

[10] Casey Neau, Modified MIT 50nm devices, Ph.D. dissertation, Purdue University, 2004.