Impact of Partial Resistive Defects and Bias Temperature Instability on SRAM Decoder Reliability

Seyab Khan Mottaqiallah Taouil Said Hamdioui Halil Kukner Praveen Raghavan Francky Catthoor
Computer Engineering Laboratory Delft University of Technology Kapeldreef 75,B-3001 Leuven, Belgium
Mekelweg 4, 2628 CD Delft,The Netherland
{M.S.K.Seyab,M.Taouil,S.Hamdioui}@tudelft.nl
{Halil,Ragha,Francky.catthoor}@imec.be

Abstract—Partial open defects in modern Static Random Access Memory (SRAM) address decoders are one of the main causes of small delays; these are hard to detect and may result in escapes and reliability problems. In addition, Aging failures -such as Bias Temperature Instability (BTI)- may worsen the situation and accelerate the degradation (i.e. increase the delay) and cause sooner field failures. This paper investigates the impact of partial opens and BTI in SRAM address decoders first separately and thereafter in a combined manner. Simulation results show that BTI impact strongly depends on the selected wordline, transistor location and addressing scheme; and it cause sooner field failures. In addition, they show that partial opens, which do not cause hard faults and allow memory operations to pass correctly, contribute up to 23.65% additional delay. Combining these failure mechanisms reveals that the degradation can strongly be worsened and accelerate wear-out; an additional delay of up to 31.20% can be caused. This indicates the importance of incorporating appropriate design-for-reliability/testability schemes in order to guarantee the required lifetime of the memory system.

Index Terms—SRAM decoders, Bias Temperature Instability, Resistive defects, Addressing schemes,

I. INTRODUCTION

Static Random Access Memory (SRAM) system is one of the driving forces in state-of-the-art semiconductor industry [1]. The memories suffer from variability and reliability issues [2]. Variability in the shrunk interconnects cause resistive/open defects. Reliability failure mechanisms, such as BTI (Negative Bias Temperature Instability (NBTI) in PMOS transistors and Positive Bias Temperature Instability (PBTI) in NMOS transistors), degrade transistors performance during the operational lifetime [3,4,6].

From variability perspective, Nigh et al. in [7] reported that newer technologies will suffer from elusive variability defects causing more test escapes that will show up as timing failures during operation. Kundu in [8] argued that the defect mode shifts from shorts/bridges to more resistive/open defects causing more timing failures. Needham reported in [9] that resistive defects are the main cause of field return of Intel microprocessors. On the other, it well recognized that BTI causes threshold voltage shift to MOS transistors during operation and consequently additional gates/circuits delays [10–12,25]. In conclusion, BTI induced delay can accumulate with the delay caused by partial open defects resulting in the acceleration of SRAM field failures.

An SRAM system comprises memory cells array and peripheral circuits, such as address decoders, control, read/write drivers, and sense amplifiers. Much have been published on the separate impacts of resistive/open defects and BTI on the memory cell array [14–18]. For instance, Hamdioui in [13,14] and Dilillo in [15] investigated resistive defect in SRAM cell. Nourivand et al. in [16] analyzed resistive defects in drowsy SRAM cells. Similarly, Kumar et al. in [17] analyzed NBTI impact on Static Noise Margin (SNM) and read stability of the SRAM cell. Kang et al. in [18] investigated BTI impact the on SRAM cell’s SNM, read and write stability, and leakage current. On the other hand, few authors have focused on defects analysis of the address decoders. For instance, Hamdioui et al. in [14] presented analysis of spot defects in SRAM and in [19] identified decoder delay faults due to inter and intra-gate resistive defects. Manoj in [20] proposed a test pattern to identify defects in the decoders. Dilillo et al., in [21] presented a comparative analysis of resistive and open faults in address decoder of the embedded-SRAM.

Although it is known that both resistive open and BTI can cause timing failures and reduce memory reliability, to the best knowledge of the authors there is no published work comparing their impacts and analyzing their combined effect. These are the aspects covered in this paper; therefore, it analyzes the degradation due to BTI and resistive defects on the SRAM decoders. The main contributions of the paper are:

- Investigation of BTI impact on the decoder for different addressing schemes. The addressing schemes include linear, gray, and address-complement.
- Analysis of the impact of resistive defects at different locations in the address decoders. The addressing scheme used in the analysis is linear-up.
- Investigation of the combined impact of BTI and resistive defects on the decoder.

The rest of the paper is organized as follows: Section II first briefly presents the overall SRAM and focuses on decoder;
Bias Temperature Instability (BTI) mechanism takes place inside MOS transistors and causes a threshold shift that translates to additional delay, as described below.

BTI Mechanism

BTI causes threshold voltage ($V_{th}$) increment to MOS transistors. The $V_{th}$ increment in a PMOS transistor that occurs under the negative gate stress is known as NBTI. For a MOS transistor, there are two BTI phases, i.e., the stress phase and the relaxation phase.

In recent times, exhaustive efforts have been put to understand NBTI [4–6]. Kaczer et al. in [5] have analyzed NBTI reasonably well but have not extended their analysis to deal with BTI at a higher level. Alam et al. in [6] have modeled NBTI and presented the overall dynamics of NBTI as a reaction diffusion process. The model is usable at a higher level such as circuit level. Since in this work, BTI analysis is done at the circuit level, model of [6] will be used.

Stress Phase: In the stress phase, the Silicon Hydrogen bonds ($≡$Si-H) break at Silicon-Oxide interface. The broken Silicon bonds ($≡$Si-) remain at the interface (known as interface traps), and the released H atoms/molecules diffuse towards the poly gate. The number of interface traps ($N_{IT}$) generated after applying a stress of time ($t$) is given by [6]:

$$N_{IT}(t) = \left( \frac{N_0 \cdot k_f}{k_r} \right)^{2/3} \cdot \left( \frac{k_H}{k_{H_2}} \right)^{1/3} \cdot (6 \cdot D_{H_2} \cdot t)^{1/6}, \quad (1)$$

where $N_0$, $k_f$, $k_r$, $k_H$, and $k_{H_2}$, represent initial $≡$Si-H density, $≡$Si-H breaking rate, $≡$Si- recovery rate, H to H$_2$ conversion rate, and H$_2$ to H conversion rate inside the oxide layer, respectively. While $D_{H_2}$ is the hydrogen diffusion constant.

Relaxation Phase: In the relaxation phase, there is no $≡$Si-H breaking. However, the H atoms/molecules diffuse back towards the interface and anneal the $≡$Si- bonds. The number of interface traps that do not anneal by the approaching H atoms during the relaxation phase is given by [17]:

$$N_{IT}(t_o + t_r) = \frac{N_{IT}(t_o)}{1 + \sqrt{\frac{\xi t_r}{t_o + t_r}}} \quad (2)$$

where $N_{IT}(t_o)$ is the number of interface traps at the start of the relaxation, $\xi$ is a relaxation coefficient with $\xi=0.5$ [17], $t_o$ is the duration of the previous stress phase and $t_r$ is the relaxation duration.

The $N_{IT}$ opposes the gate stress resulting in the threshold voltage increment ($\Delta V_{th}$). The relation between $N_{IT}$ and $\Delta V_{th}$ is given by [25]:

$$\Delta V_{th} = (1 + m) \cdot q \cdot N_{IT} / C_{ox} \cdot \chi \cdot \gamma, \quad (3)$$

where $m$, $q$, and $C_{ox}$ are the holes/mobility degradation that contribute to the $V_{th}$ increment [10], electron charge, and oxide capacitance, respectively. $\chi$ is a BTI coefficient with a value $\chi=1$ for NBTI and $\chi=0.5$ for PBTI [23]. Additionally, $\gamma$ represents the stress duration with respect to the total input period (i.e., activity factor) of the transistor. The $\gamma$ dependence of the $\Delta V_{th}$ shows that transistors in a gate/circuit that have different stress and relaxation phases will suffer from different degradations.
BTI Model

BTI induced $\Delta V_{th}$ of each individual MOS transistor has its contribution to the additional delay. A generalized formula that relates BTI induced $\Delta V_{th}$ in a transistor to wordline/bitline delay is given by [24,25]:

$$\Delta D = \frac{n \cdot \Delta V_{th}}{(V_{gs} - V_{th})}$$

where $n$ is the velocity saturation index of majority carriers in MOS channels. Since NBTI causes $\Delta V_{th}$ to PMOS transistor and PBTI causes $\Delta V_{th}$ to NMOS transistor, the paper considers the threshold voltage shifts to both types of MOS transistors.

C. Resistive Defects

The resistive defects in decoder of Fig. 1(b) can be classified as followings.

- **Inside pull-up network:** The PMOS transistors between the supply and the final/intermediate output node constitute the pull-up network. $R_{pu}$ is an example of a defect in the pull-up network of the decoders shown Fig. 1(b).

- **Inside pull-down network:** The NMOS transistors between the final/intermediate output node and the ground manifest the pull-down network. There are many possible resistive defect locations in the pull-down network of the decoder (e.g., between each pair of the NMOS transistors). For simplicity, $R_{pd}$ is the only considered resistive defect in pull-down network of the decoder shown in Fig. 1(b) and is adjacent to the ground.

The resistive defects contribute to the delay of wordline activation and or de-activation delays in the following ways. First, the resistance limits the current that (dis-)charges node X. Second, the voltage drop in the resistive defects can be high enough to prevent transitions at node X.

III. SIMULATION SETUP

This section describes the simulation environment and the type of experiments conducted in the paper.

A. Simulation Environment

A netlist of a $5 \times 32$ address decoder (Fig. 1(b) shows the wordline selection for the last word only) has been synthesized using 32nm PTM transistor models [26] and simulated using HSPICE. For the BTI analysis, the impact is augmented to each transistor with Verilog-A modules. Each module generates voltage shift that depends on the activity factor of the transistor. Similarly, in the resistive defect analysis, resistors are inserted at the locations shown in Fig. 1(b); one at a time. The measurements are performed when the signals reach 50% of the $V_{dd}$.

B. Performed Simulations

Three sets of experiments are performed in this work:

1) **BTI Impact experiments:** First, BTI impact on the wordline activation and de-activation delay of a $5 \times 32$ decoder (using linear-up addressing scheme) is investigated. Second, the impact of the location of the affected transistor on the delay is analyzed. Finally, BTI dependency on the workload is addressed while considering realistic addressing schemes.

2) **Resistive open experiments:** The impact of the resistive defect in the pull-up or in the network pull-down network on the delay is investigated.

3) **Combined impact experiments:** In this case, both the impact of BTI as well as that of resistive defect (either in the pull-up or in the pull-down network) is considered in order to measure the increase in the delay.

IV. SIMULATION RESULTS

This section reports and analyzes the results of the first two sets of experiments described in the previous section.

A. Impact of BTI

First, BTI impact on the wordline activation and de-activation are presented. Then, the impact of the location of the affected transistor on the delay is described. Finally, BTI dependency on the workload covered.

**BTI Impact on wordline Delay**

The $\Delta V_{th}$ in MOS transistors due to BTI affect the wordline activation and de-activation delays. Both NBTI and PBTI impact the delays uniquely, i.e., NBTI in PMOS transistors affect the de-activation delay while PBTI in the NMOS transistors affect the activation delay. In addition, different access patterns (or workload) stress the transistors differently, resulting in different delays. We analyze the workload in more detail later. Here, we assume a linear-up addressing scheme (i.e., from address 0 to address 31) for the $5 \times 32$ decoder. The results are shown in Fig. 2; the x-axis presents the number of the accessed row and the y-axis the corresponding delay increase in % . The figure shows that on the average the induced delays in the wordline activation and deactivation are 13% (due to PBTI) and 7.50% (due to NBTI) respectively. Hence, the BTI impact on the wordline activation is more significant. This can be justified by analyzing Fig. 1(b). Normally, the NBTI induced delay is more significant than caused by PBTI.
for an inverter [4]. However, as all the NMOS transistors of the pull-down network are connected in series, their impact accumulate, resulting in higher impact on the wordline activation. The figure also shows an oscillating trend in the impact as we move from lower to higher wordlines/addresses. This is directly related to the way the addressing sequence is selected; in this case linear-up. For instance, moving from address 00001 (WL1) to 0010 (WL2) requires the switching of two transistors in the pull-down network (which are in series), while moving from address 00010 to 00011 (WL3) requires only the switching of a single transistor in the pull-down network; hence less delay.

**BTI Impact of Transistor Locations**

Most of published work assume that the threshold voltage shifts in all the degraded transistors have uniform contribution to the delay increment of the circuit [17,25]. However, in reality, such a contribution strongly depends on the location of the transistor in the circuit.

In order to explore such dependency both for NBTI and PBTI, the impacts of the reference transistors $M_{\text{ref1}}$ and $M_{\text{ref3}}$ (for NBTI), as well as the impacts of reference transistors $M_{\text{ref2}}$ and $M_{\text{ref4}}$ (for PBTI) will be simulated; see Fig. 1(b). For this experiments, obviously we assume that only a single transistor is degraded in order to measure the contribution of each transistor separately.

Fig. 3(a) shows the additional delay caused by NBTI induced $\Delta V_{\text{th}}$ in $M_{\text{ref1}}$ and $M_{\text{ref3}}$. The figure shows that additional delay due to $M_{\text{ref1}}$ is 6.18\% and that due to $M_{\text{ref3}}$ is 5.82\% after $10^5$ s; hence, the contribution of parallel connected transistors to the delay is nearly the same and the uniform approach in [17,25] can be used. In the right side of the figure, the results of PBTI impact in $M_{\text{ref2}}$ and $M_{\text{ref4}}$ are depicted (see Fig. 3(b)). The figure clearly shows the impact is strongly location of the transistor dependent. For example, the induced $\Delta V_{\text{th}}$ by $M_{\text{ref2}}$ cause 2.01\% additional delay, while the same shift in $M_{\text{ref4}}$ causes 1.7x more delay increment. The higher degradation contribution of $M_{\text{ref4}}$ can be explained as follows. The increased $\Delta V_{\text{th}}$ at $M_{\text{ref2}}$ causes a much weaker drive strength at the source of transistor $M_{\text{ref4}}$. Therefore, the uniform approach in [17,25] is not accurate enough for serially connected transistors.

**BTI Impact of Various Workloads**

The $\Delta V_{\text{th}}$, is determined by stress and relaxation durations of the transistors. Therefore, BTI impact is strongly dependent on the applied input patterns (henceforth referred to as workload). This workload dependence is analyzed by applying various the following addressing schemes:

- **Linear-up**: this is binary sequence counting from address 0 to 31 for 5-bit address decoder. The sequence is 0, 1, 2, 3, ... 30, 31.
- **Linear-down**: this is the inverse sequence of linear up; for 5-bit address decoder the sequence is: 31, 30,..., 1, 0.
- **Gray-up**: is the sequence presented by the binary code where two successive values differ in only one bit. For a 5-bit decoder, the sequence is: 0,1,3,2,...17,16.
- **Gray-down**: is the inverse of the Gray-up. For 5-bit decoder, the sequence is 16,17,19,18,..., 1, 0.
- **Address complement-up** [19]: for 5 bit decoder, the binary representation of address sequence is {00000, 11111, 00001, 11110, ..., 01111, 10000} etc. Note that each address is followed by its one’s complement (in bold-face). The sequence is equivalent to the sequence 0, 31, 1, 30,..., 15, 16.
- **Address complement-down**: it is the reverse address sequence of address complement-up.

Fig. 4 shows the obtained results of both wordline activation and wordline de-activation additional delays for the different addressing schemes. The results reveal that on average the degradation in the de-activation delay is about 2x that of the activation delay. In addition, the results indicate that the variation in delay between the considered addressing schemes does not exceed 2\%. Hence, for the considered address decoder, the impact of different addressings is not significant. BTI impact on wordline activation ranges between 6.85\% (for address complement-down addressing scheme) and 8.17\% (for the gray-up addressing scheme), while that on the wordline de-activation ranges between 12.69\% (for...
the address complement-up) and 14.27% (for linear-down addressing).

**B. Impact of resistive opens**

The impact of partial resistive open defects in the pull-up and in the pull-down network of the CMOS address decoder on the timing of the wordlines as described next.

**Inside Pull-up Network**

The resistive defect in the pull-up network of the decoder shown in Fig. 1(b) affects the rising transition at node X, consequently the de-activation of the wordline.

Figure 5(a) shows the wordline transitions for various resistive defects; R_{pu} between 0kΩ and 100kΩ. In general, it shows that the resistance has impacts on both activation and de-activation transitions of the decoder. However, impact on the wordline de-activation is more significant. The larger the resistive defects the worsen the transition delay. For example, the falling transition time increases with 6000% in case R_{pu} = 100kΩ as compared to the defect free case. Higher resistive defects are easy to catch during manufacturing test; however, small resistance defects cause small delays and are hard to detect; they may escape the test and cause reliability problems in the field.

Fig. 5(c) shows the impact of these low resistances present in the pull-up and pull-down networks. Here, we focus on the results related to de-activation delay only (caused by defects in the pull-up network). The figure depicts the delay increment for R_{pu} between 0 and 5 kΩ. It shows that the delay increments is linear proportional to the resistive defect value; the (de-activation) delay increases approximately with a rate of 3.74%/kΩ (18.70%/5kΩ).

**Inside Pull-down Network**

The resistive defect in the pull-down network of the decoder shown in Fig. 1(b) affects activation delay of the wordlines. Simulation results are depicted in Fig. 5(b): the ascending resistances have significant impact on the wordline activation. For example, at 100 kΩ the resistance causes even a transition failure, i.e., the pull-down network is not strong enough to drive node X to zero; therefore, the inverter fails to drive the wordline high. Fig. 5(c) shows the impact of the low lower resistance R_{pd} between 0 and 5 kΩ. The delay increments shows a linear behavior with increasing pull-down resistance values; the delay increases with a rate of 4.73%/kΩ (23.65%/5kΩ).

**V. BTI AND RESISTIVE DEFECT IMPACT ON SRAM DECODER**

This section focuses on the last set of experiments described in Section III; it analyzes the combined impact of BTI and the resistive defects (both in pull-up and pull-down network), while considering linear-up addressing scheme for the decoder shown in Fig. 1(b).

Figure 6 shows the simulation results for: (a) BTI combined with a resistive defect R_{pu} in the pull-up network and (b) BTI combined with a resistive defect R_{pd} in the pull-down network. Note that the value of R_{pu} and R_{pd} considered in the experiment range from 0 and 5 kΩ (same values as in Fig. 5(c)); these value do not cause hard faults (such as a failing access of the memory), but cause small delays that may escape manufacturing test.

The figure shows that BTI combined with either pull-up or pull-down resistance have mutually exclusive impact on the wordline activation and de-activation delays. For example, for R_{pu}=R_{pd}=0, the delay increment in the wordline activation is
only 13.20%, and in the wordline de-activation is 7.50%. The BTI induced delays are consistent with results presented in the previous section. The figure also shows that higher values of of $R_{pd}$ or $R_{pu}$ increase activation and de-activation wordline delays. However, the increment rate in the activation delay due to $R_{pd}$ increment is higher than that of de-activation delay due to $R_{pu}$. Increasing the value of the resistive defect $R_{pd}$ (i.e. $R_{pd} = 0 \rightarrow 5k\Omega$) increases the wordline activation delay from about 13.20% to 31.20%. Similarly, the increment in the pull-up resistance (i.e. $R_{pu} = 0 \rightarrow 5k\Omega$) cause the de-activation delay to reach 22.15%. These trends are also consistent with what we found in the previous section.

In conclusion, it can be stated that when BTI and resistance are considered simultaneously, the mean delay increment approaches 30% for the wordline activation. The impact is quite significant; hence BTI can accelerate the field failures of an SRAM system in the presence of small open defects that are hard to detect at manufacturing test phase.

VI. CONCLUSION

The analysis presented in this paper for BIT and resistive defects clearly shows that the life time degradation in the presence of small open defects in the memory decoders accelerate the failures. This will become even severe for the future smaller technologies, where the 2011 ITRS is expecting higher and even scary failure rates of $10^{-2}$. As the simulation results showed, depending on the resistor location and the workload, the additional delay can reach 30%. Therefore, it is extremely important to address the degradation of the decoder performance in the presence of small defects both at the design stage and during operation. At the design stage, the test technique should be developed to capture the smaller resistive defects in the decoder while stressing the decoder (acceleration of aging). However, if the defects escape the test, it should be mitigated during the operation by dynamic techniques such as frequency reduction, $V_{dd}$ adjustment and adaptive body biasing.

REFERENCES

[26] Predictive Technology Model "http://ptm.asu.edu/"