A Finite-Oxide Thickness Based Analytical Model for Negative Bias Temperature Instability

Sanjay V. Kumar, Chris H. Kim, and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455
sanjay, chriskim, sachin@umn.edu

***Abstract*— Negative Bias Temperature Instability (NBTI) in PMOS transistors has become a serious reliability concern in present day digital circuit design. With continued technology scaling, and reducing oxide thickness, it has become imperative to accurately determine its effects on temporal circuit degradation, and thereby ensure reliable operation for a finite period of time. A reaction-diffusion (R-D) based framework is developed for determining the number of interface traps as a function of time, for both the DC (static NBTI) and the AC (dynamic NBTI) stress cases. The effects of finite oxide thickness, and the influence of trap generation and annealing in polysilicon, are incorporated. The model provides a good fit with experimental data, and also provides a satisfying explanation for most of the physical effects associated with the dynamics of NBTI. A generalized framework for estimating the impact of NBTI-induced temporal degradation in present day digital circuits, is also discussed.**

***Key Terms* : Negative Bias Temperature Instability (NBTI), Reaction-Diffusion (R-D) Model, Frequency Independence, Oxide Thickness, Delay.**

I. Introduction

When a PMOS transistor is biased in inversion ($V_{gs} = -V_{dd}$), interface traps are generated due to the dissociation of $Si-H$ bonds along the substrate-oxide interface. The rate of generation of these traps is accelerated by temperature, and the time of applied stress. These traps cause an increase in the threshold voltage (V_{th}), and a reduction in the saturation current (I_{dsat}) of the PMOS transistors. This effect, known as Negative Bias Temperature Instability (NBTI), has become a significant reliability issue in high-performance digital IC design, especially in sub-130nm technologies [1]–[6]. An increase in V_{th} causes the circuit delay to degrade, and when this degradation exceeds a certain amount, the circuit may fail to meet its timing specifications.

Experiments have shown that the application of a negative bias ($V_{gs} = -V_{dd}$) on a PMOS transistor leads to the generation of interface traps, while removal of the bias ($V_{gs} = 0$) causes a reduction in the number of interface traps due to annealing [2]–[5], [7]–[11]. Thus, the impact of NBTI on the PMOS transistor depends on the sequence of stress and relaxation applied to the gate. Since a digital circuit consists of millions of nodes with differing signal probabilities and activity factors, asymmetric levels of degradation are experienced by various timing paths. The exact amount of degradation must be determined using a model that estimates the amount of NBTI-induced shift in the various parameters of the circuit that affect the delay. This metric can then be used to design circuits with appropriate guard-bands, such that they remain reliable over the desired lifetime of operation, despite temporal degradation.

Over the past years, there have been many attempts to model the NBTI effect, based on theories, such as reaction-diffusion, dispersive diffusion, hole trapping. The reaction-diffusion (R-D) theory [12], [13] has commonly been used to model NBTI, leading to various long-term models for circuit degradation [4], [14]–[16]. However, alternative views among researchers exist, particularly about the inability of the R-D model to explain some key phenomena, as detailed in [17]–[21]. This has led to alternative models such as [17], [22]–[26], as well as efforts to resolve the controversy between the R-D model theory and the hole trapping theory [27]–[30]. While, this area is still under active research, the domain of our work is restricted to NBTI modeling based on the R-D theory.

This paper compares the existing models for predicting long term effects of aging on circuit reliability, within the R-D framework [4], [14]–[16], and finds that these models do not successfully explain the experimentally observed results. In this regard, we first sketch an outline for the basic requirements of any NBTI model, based on observations from a wide realm of experimental data. Further, most of these models assume that the oxide thickness (d_{ox}) is infinite, which is particularly not valid in sub-65nm technologies, where d_{ox} is of the order of a nanometer. Hence, the effect of interface trap generation and recombination in polysilicon must be considered while developing a model. Numerical simulations are also performed to illustrate the drawbacks of existing models based on the R-D theory, and to highlight the importance of considering the effect of finite oxide thickness.

Accordingly, we propose an R-D based model for NBTI that does not consider the oxide to be infinitely thick. The results show that the model can resolve several inconsistencies, noted with the reaction-diffusion theory for NBTI generation and recombination, as observed in [17]–[20]. Further, the model can also explain the widely distinct experimentally observed results in [20], [31], [32]. Implications of the model, and its usage in determining the long-term impact of NBTI on digital circuit degradation after three years of operation are discussed. Besides the actual analytical modeling and the framework for estimating the degradation of digital circuits, our contribution also involves providing a better understanding of the empirical constant ξ, as used in [4], and has been misinterpreted as being universal.

The paper is organized as follows. Section II outlines the

previous work in NBTI modeling, and their shortcomings. Based on these drawbacks, we outline a set of guidelines that can be used to verify the correctness of an NBTI model. Section III describes the R-D model equations, while Section IV presents a solution to the first stress phase, or the DC stress case of NBTI action. In Section V, we outline a numerical simulation framework for the first stress and recovery phases, thereby showing the origin for some of the key drawbacks of the R-D based model in [4], as well as highlighting the role of finite oxide thickness in long term recovery. Section VI then provides a detailed derivation of the model for the first recovery phase. Simulation results and comparison with experimental data are shown in Section VII. We use the stress and recovery models derived for a single stress and relaxation phase, and extend this to a multi-cycle framework in Section VIII. Section IX then shows how this model can be used to estimate the impact of NBTI on the delay degradation of digital circuits, followed by inferences in Section X.

II. Previous Works and their Shortcomings

In this section, we present the drawbacks of the existing NBTI models based on the R-D theory, in literature. We then proceed to outline a set of requirements that an NBTI model must adhere to, in order to be able to account for the physics of interface trap generation and recombination. The Reaction-Diffusion model was first used in [12] to physically explain the mechanism of negative bias stress (NBS) in p-channel MOS memory transistors, based on the activation energy of electrochemical reactions. Several years later, a detailed mathematical solution to the R-D model was presented by [13]. Subsequently, [4], [10], [11], [33] have used the R-D model to describe the NBTI effect in present day PMOS devices.

The analytical model for NBTI in [4] by Alam provides a simple means to estimate the number of interface traps for a single stress phase, followed by a relaxation phase, under the assumption of infinite oxide thickness. The model does not capture the rapid decrease in the concentration of hydrogen initially, and predicts a 50% reduction in V_{th} when the relaxation time is equal to the stress time. The fit with experimental data (Fig. 3-page 2 of [4]) is not very accurate, especially during the initial part of recovery. We will show later on in Section V that this is due to two reasons:

- The use of a single fixed value of ξ = 0.58 for modeling the back-diffusing front during recovery, whereas in reality ξ varies with time.
- Finite oxide thicknesses, and a higher diffusion rate of H_2 in the oxide, as compared with polysilicon.

The work in [14] provides a multi-cycle analytical model for NBTI, with the framework for the first stress and relaxation phases being built upon the work in [4]. The model demonstrates the widely observed relation that the amount of trap generation over a large period of time is independent of the actual frequency of operation, known as frequency independence [4], [9], [10]. The framework also provides an analytical proof for frequency independence, and a method for estimating the delay of digital circuits, after ten years of degradation. However, the model in [14] does not provide a good fit with experimental data, particularly during the initial few seconds of recovery. Further, the analytical modeling is derived under the assumption of infinite oxide thickness, which is not valid in current process technologies. Our work extends the modeling in [14] to remove the limitations listed above.

The work in [15] is also based on an infinite oxide thickness assumption. To capture the rapid decrease in the number of interface traps during the initial stages of recovery, the model lumps a constant δ. The value of δ is used to fit with experimental data, and no analytical means of computing this value is provided. Further, the shape of the curve around the 1000-1500s region in Fig. 4 of the paper does not fit well with experimental data, from [34]. The above method however is insightful, and leads to a case, where a two-level model for the recovery phase: one for recovery in the oxide, and another for recovery in polysilicon, may be required for accurate modeling, as explained in [35].

Accordingly, the work in [16] attempts to incorporate the effects of finite oxide thickness, and the differing rates of diffusion of H_2 in oxide, and poly, and thereby provides a comprehensive multi-cycle model. The work in [16] concurs with [14] in showing frequency independence analytically. The model provides an excellent fit with experimental data from [35], and shows more recovery for a higher d_{ox} value, which is consistent with experimental observations in [35].

However, the value of ξ in the model in [16] is deemed to be universal, and this can lead to unexpected results as follows. For instance, the recovery phase of the model in [16] for the d_{ox} = 1.2nm case is examined, for a single stress phase of 10000s, followed by continuous recovery for a long period of time. It is expected that the amount of recovery must continue to increase, with time, leading to near complete recovery at infinite time [36]. However, an evaluation of the model shows that the recovery curve reaches a minimum at around 40000s, and continues to increase beyond that time. A similar behavior is seen for the d_{ox} = 2.2nm case, with the minimum occurring at around 20000s, and the deviation from the minimum value is larger here. This may lead to unexpected behavior, and the minimum may shift toward a lower time point, for lower stress periods, and higher oxide thicknesses.

A. Guidelines for an NBTI Model

Based on the drawbacks identified from these models, as well as observations from several publications such as [19], [20], [24], we present some key guidelines for an NBTI model as follows:

1) The model must predict that the number of interface traps increases rapidly with time initially, as explained in [10], [37], and asymptotically lead to a $N_{IT}(t) \propto t^{\frac{1}{6}}$ relationship, (assuming that the diffusing species are neutral hydrogen molecules), as experimentally observed in [1], [5], [35].
2) The model must be able to capture the "fast initial recovery phase" that is of the order of a second [20], during which recovery is higher.
3) The model must predict higher fractional recovery for a PMOS device with a larger t_{ox}, for the same duration

of stress, as observed in [35]. This is because, a larger d_{ox} implies a larger number of fast diffusing hydrogen molecules in the oxide, and hence implies higher amounts of annealing.

4) For an AC stress case where the stress duration is equal to the relaxation time period, the model must predict larger fractional recovery, with lower stress times [20]. Previous works using an NBTI model [4], [15] and numerical solutions of the model in [17], [19] all predict 50% recovery, when the ratio of the relaxation time to the stress time is equal to one, irrespective of the actual duration of the stress time.
5) The model must predict some form of frequency independence, i.e., the number of interface traps generated must approximately be the same asymptotically, irrespective of the frequency of operation. Although, the exact range of frequencies over which this phenomenon holds good is still not very clear, some form of frequency independence is widely observed in the 1Hz - 1MHz range [4], [34] and has recently been shown to exist over the entire range of 1Hz - 2GHz in [38].

B. A Note on OTFM and UFM Techniques and Validity of the R-D Theory

Two current state-of-the-art techniques to measure the impact of NBTI on V_{th} during recovery include OTFM (On-the-Fly Measurement) which estimates ΔV_{th} by measuring $|\frac{\Delta I_d}{I_{d_0}}|$, and UFM (Ultra-Fast V_{th} Measurement) which estimates the intrinsic NBTI and V_{th} degradation directly. UFM-based techniques, which can measure the V_{th} degradation during the recovery phase, within 1μs after removal of the stress, have been employed in [18], [19]. Experimental results show that there is a uniform recovery of V_{th} during the relaxation phase, with an almost identical amount of fractional recovery in every decade. Subsequently, [17], [19] show results comparing the large differences between an R-D theory-based model for recovery, and the experimental data, suggesting that the R-D mechanism does not provide a satisfactory explanation for the physical action during recovery. Further, [17] explains the various drawbacks of the R-D theory-based analytical model proposed by Alam in [4], such as:

1) 50% recovery in V_{th} predicted after τ seconds of recovery, following τ seconds of stress, irrespective of the value of τ, whereas experimental results show a dependence on τ, particularly with smaller values of τ producing larger fractional recovery.
2) Numerical simulations of the R-D model predict 100% recovery, whereas [4] predicts only around 75% recovery, as $t \rightarrow \infty$.
3) Poor fit during the beginning of the recovery phase ($t \ll \tau$), and for $t \gg \tau$.

The authors in [17] hence propose a dispersive transport based model for trap generation and recovery. Further, the works in [22], [25], [26], [39] support a bulk trapping-detrapping based model, instead of a reaction-diffusion based model. However, [29] distinguishes the gate dielectrics into two types (Type I and Type II) depending on whether they are PNO (plasma nitrided oxides) or TNO (thermal nitrided oxides), and explains the discrepancy between the bulk trapping and the R-D models, for each of these types. Recently, [40] highlights the differences between an OTFM and a UFM-based technique for analyzing the impact of NBTI. The above work also shows that the R-D theory is consistent with the experimental results obtained using OTFM techniques, and the log-like recovery (equal recovery in every decade) observed in [18], [19] is consistent with a UFM-based technique. The authors in [40] also state that the log-based recovery of V_{th} observed in [19] is due to the inappropriate usage of the quasi-state relationship:

$$\Delta V_{th} = \frac{q \Delta N_{IT}}{C_{ox}} \tag{1}$$

to ultrafast transient conditions. Further [40] explains the drawbacks in using a UFM-based technique, and strongly supports the validity of the reaction-diffusion theory for predicting the impact of NBTI correctly.

It must be noted that our model is presented under the above assumption that the R-D theory provides a valid and satisfying explanation for interface trap generation and recombination. Our work seeks seeks to provide a better understanding of the R-D mechanism, thereby improving upon the drawbacks in previous (R-D based) works, as listed in the beginning of this section. Further, it must be noted that our goal is to build a modeling mechanism for NBTI action that can be used to predict the impact on the timing degradation of digital circuits after several years of operation. Hence, a fast asymptotically accurate model, as opposed to a slow cycle accurate model that requires extensive numerical simulations, is of utmost utility.

III. Reaction Diffusion Model for NBTI Action

In this section, we describe the framework of the Reaction-Diffusion (R-D) model, used to develop an analytical model for NBTI action. The R-D model is solved assuming that alternate periods of stress and relaxation, each of equal duration τ, are applied to the gate of a PMOS device, whose source and bulk are tied to V_{dd} while the drain is grounded, as shown in Fig. 1. It must be noted that the derivation is valid, with minor changes in the limits of integration, for any arbitrary sequence of stress and relaxation. However, since the special case of a square wave-like sequence of "alternating" stress and relaxation (also called AC stress in the NBTI literature) is frequently used in experimentation, we consider this case.

A. Reaction-Diffusion Model

The R-D model is used to annotate the process of interface trap generation and hydrogen diffusion, which is governed by the following chemical equations:

$$\begin{aligned} Si - H + h^+ &\rightarrow Si^+ + H \\ H + H &\rightarrow H_2 \end{aligned} \tag{2}$$

where the holes in the channel interact with the weak $Si-H$ bonds, thereby releasing neutral hydrogen atoms, and leaving behind interface traps. Hydrogen atoms combine to form hydrogen molecules, which diffuse into the oxide.

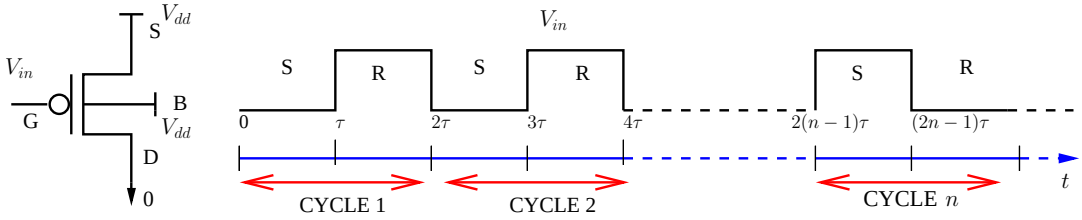


Fig. 1. Input waveform applied to the gate of the PMOS transistor to simulate alternate stress (S) and relaxation (R) phases of equal duration τ.

According to the R-D model, the rate of generation of interface traps initially depends on the rate of dissociation of the Si-H bonds (which is controlled by the forward rate constant, k_f) and the local self-annealing process (which is governed by the reverse rate constant, k_r). This constitutes the **reaction phase** in the R-D model. Thus, we have:

$$\frac{dN_{IT}}{dt} = k_f[N_0 - N_{IT}] - k_r N_{IT} N_H^0 \tag{3}$$

where N_{IT} is the number of interface traps, N_0 is the maximum density of $Si-H$ bonds and N_H^0 is the density of hydrogen atoms at the substrate-oxide interface. After sufficient trap generation, the rate of generation of traps is limited by the diffusion of hydrogen molecules[1]. The rate of growth of interface traps is controlled by the diffusion of hydrogen molecules away from the surface as:

$$\frac{dN_{IT}}{dt} = \phi_{N_{H_2}} \tag{4}$$

where $\phi_{N_{H_2}}$ is the flow of diffusion of H_2 from the interface to oxide/poly. Hence, when diffusion is limited to the oxide, it follows the equation:

$$\frac{dN_{IT}}{dt} = -D_{ox}\frac{dN_{H_2}}{dx} \tag{5}$$

where D_{ox} represents the diffusion coefficient in the oxide, while D_p is that in polysilicon.

Using Fick's second law of diffusion, the rate of change in concentration of the hydrogen molecules inside the oxide is given by:

$$\frac{dN_{H_2}}{dt} = D_{ox}\frac{d^2N_{H_2}}{dx^2} \text{ for } 0 < x \leq d_{ox} \tag{6}$$

where N_{H_2} is the concentration of hydrogen molecules at a distance x from the interface at time t, (while $N_{H_2}^0$, at the substrate-oxide interface)[2]. This constitutes the **diffusion phase** in the R-D model. In order to find a coupling relation between N_H^0 in the reaction-phase equation in (3) and $N_{H_2}^0$ in the diffusion-phase equation, we use the mass action law:

$$N_{H_2}^0 = k_H (N_H^0)^2 \tag{7}$$

since two hydrogen atoms can combine to form a hydrogen molecule with the rate constant k_H [10], [41].

[1]Initial works assumed diffusion of hydrogen atoms, although it is now widely conjectured that hydrogen molecular diffusion occurs [5], [10], [35].

[2]We will represent $N_{H_2}^0(t)$ and $N_H^0(t)$ as $N_{H_2}^0$ and N_H^0, respectively, except in cases where the the value of t is not obvious within the context.

B. Solution to the Reaction Phase

During the initial reaction phase, the concentration of hydrogen atoms and interface traps are both very low, and there is virtually no reverse reaction. Hence, the number of interface traps increases with time linearly as:

$$N_{IT}(t) = k_f N_0 t \tag{8}$$

The linear dependence of N_{IT} on time t correlates with results from numerical simulations in [5], [41]. This process lasts for a very short time (around 1ms). Gradually, the process of interface trap generation begins to slow down due to the increasing concentration of hydrogen molecules, and the reverse reaction. The process then attains a quasi-equilibrium [42], and subsequently becomes diffusion limited.

Fig. 2 shows results from our numerical simulation setup (described later on in Section V), showing the three regimes namely:

1) Reaction phase which lasts less than a millisecond, during which N_{IT} increases linearly with time, as seen from Fig. 2.
2) Quasi-equilibrium phase during which the interface trap count does not increase.
3) Rate-limiting diffusion phase during which the mechanism is diffusion limited.

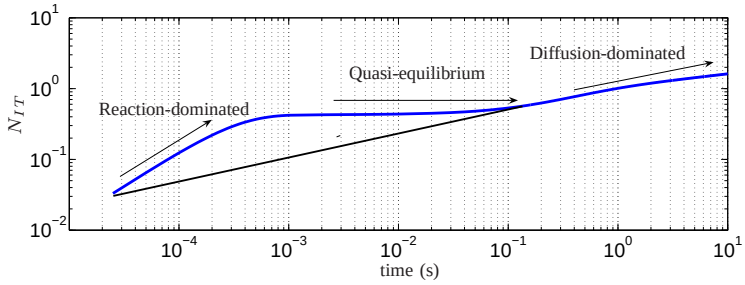


Fig. 2. Results of numerical simulation showing the three regimes of interface trap generation, during the DC stress phase.

The reaction phase is ignored in the final model, for reasons that will become apparent at the end of Section IV-A.

C. Diffusion Phase

During this phase, the diffusion of hydrogen molecules becomes the rate limiting factor. Since the number of interface traps now grows rather slowly with time, the left hand side in (3) is approximated as zero. The initial density of Si-H

bonds is larger than the number of interface traps that are generated, so that $N_0 - N_{IT} \approx N_0$. This leads to the following approximation for the reaction equation:

$$\frac{k_f N_0}{k_r} \approx N_{IT} N_H^0 \tag{9}$$

We initially solve the diffusion equation for the first stress and relaxation phases, and provide a method to extend the solution to the subsequent phases, in Section VIII.

IV. The First Stress Phase

The first stress phase occurs from time t = 0s to τ, as indicated in Fig. 1. During this stage, the PMOS device is under negative bias stress, and hence, generation of interface traps occurs. The stress phase consists of two components, namely diffusion in oxide and diffusion in polysilicon, leading to two analytical expressions, respectively.

A. Diffusion in Oxide

The number of interface traps increases with time rapidly initially, as given by (8), before reaching quasi-equilibrium, and eventually the mechanism becomes diffusion-limited. At this point, the rate of generation of hydrogen is rather slow, and therefore, diffusion within the oxide, described by (6), can be approximated as:

$$D_{ox} \frac{d^2 N_{H_2}^x(t)}{dx^2} = 0 \tag{10}$$

This implies that $N_{H_2}^x(t)$ is an affine function of x, where x is the extent to which the front has diffused at a given time t. The diffusion front can be approximated as shown in Fig. 3, which plots the front at various time points, during the diffusion process. The concentration of hydrogen molecules is highest at the interface, where the traps are generated, and gradually decreases as hydrogen diffuses into the oxide, as illustrated in Fig. 3(c). The hydrogen concentration at the interface is denoted by $N_{H_2}^0$, and can be approximated as zero at a point known as the *diffusion front*, which we will denote as $x_d(t)$: this is the extent to which the diffusing species has penetrated, at time t, into the oxide[3]. Therefore, we have:

$$\frac{dN_{H_2}}{dx} = -\frac{N_{H_2}^0}{x_d(t)} \tag{11}$$

and from the triangular approximation in Fig. 3(c),

$$N_{H_2}^x(t) = N_{H_2}^0 - \left[\frac{N_{H_2}^0}{x_d(t)}\right] x \tag{12}$$

Due to the one-one correspondence between the interface traps and the H species, the total density of interface traps must equal the total density of hydrogen atoms (or twice the number of hydrogen molecules) in the oxide. Therefore,

$$N_{IT}(t) = 2 \int_{x=0}^{x=x_d(t)} N_{H_2}^x(t) dx \tag{13}$$

[3]This is consistent with the right half of Fig. 4a in [4]: the curve there looks (deceptively) more rounded, but this is because the y-axis is on a log scale, and on a linear y-axis, the triangle is a reasonable assumption.

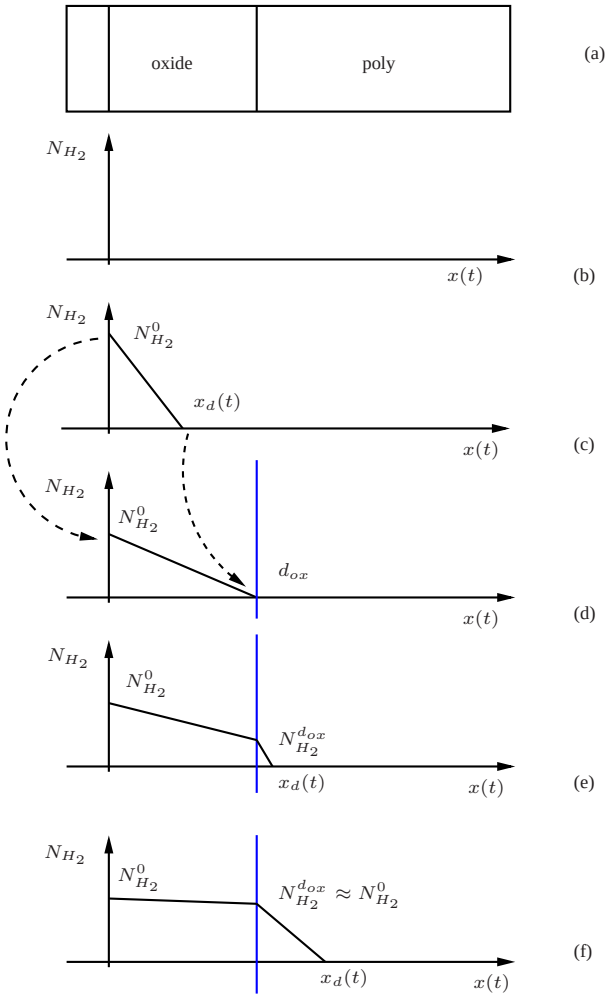


Fig. 3. Diffusion front for the first stress phase: (a) shows the cross section of the PMOS transistor: $x > 0$ denotes the direction of the oxide-poly. (b) shows the front at time $t = 0$, and the hydrogen concentration is 0. (c)-(f) show the front during the first stress phase. (c) shows the triangular approximation of the diffusion front in the oxide, with the peak denoted by $N_{H_2}^0$, while the tip of the front is at $x_d(t)$. (d) shows the front at the oxide-poly boundary, i.e., when $x_d(t) = d_{ox}$, and the subsequent decrease in the peak concentration. (e) shows the front extending into poly, while (f) shows that since $D_{ox} \gg D_p$, the front can be approximated as a rectangle in oxide, followed by a triangle in poly, i.e., $N_{H_2}^{d_{ox}} \approx N_{H_2}^0$.

The value of the above integral is simply twice the area of the triangle enclosed by the diffusion front in Fig. 3(c). Therefore,

$$N_{IT}(t) = N_{H_2}^0 x_d(t) \tag{14}$$

The above equations can be expressed equivalently in terms of N_H^0 using (7). Hence,

$$N_{IT}(t) = k_H (N_H^0)^2 x_d(t) \tag{15}$$

The approximation comes about because the reaction rate is fast enough that uncombined N_H^0 are sparse: this is supported by the fact that practically, diffusion is seen to be due to H_2 and not H. The above equation relates the number of interface traps to the number of hydrogen species at the interface. We may now substitute (15) in the LHS of (5), and (11) in the

RHS of (5), and further use (7) to obtain:

$$k_H(N_H^0)^2\frac{dx_d(t)}{dt} = D_{ox}\frac{k_H(N_H^0)^2}{x_d(t)}$$
$$\text{i.e., } x_d(t)dx_d(t) = D_{ox}dt \quad (16)$$

Integrating this, we obtain:

$$x_d(t) = \sqrt{2D_{ox}t} \quad (17)$$

and using this in (15), we get:

$$N_{IT}(t) = k_H(N_H^0)^2\sqrt{2D_{ox}t} \quad (18)$$

Finally, we substitute the above relation in (9) to obtain:

$$N_{IT}(t) = \left(\frac{k_f N_0\sqrt{k_H}}{k_r}\right)^{\frac{2}{3}}(2D_{ox}t)^{\frac{1}{6}} = k_{IT}\,(2D_{ox}t)^{\frac{1}{6}} \quad (19)$$

where $k_{IT} = \left(\frac{k_f N_0\sqrt{k_H}}{k_r}\right)^{\frac{2}{3}}$.

The above equation is valid until the tip of the diffusion front has reached the oxide-poly interface, as shown in Fig. 3(d). The time at which this occurs is denoted by t_1, and can be computed by substituting $x_d(t) = d_{ox}$ in (17) to obtain:

$$t_1 = \frac{{d_{ox}}^2}{2D_{ox}} \quad (20)$$

Typically, t_1 is of the order of a second for current technologies, considering the values of the oxide thickness, and D_{ox}. The number of interface traps for the first stress phase can thus be expressed as:

$$N_{IT}(t, 0 < t \leq t_1) = k_{IT}x_d(t)^{\frac{1}{3}} \quad (21)$$

where $x_d(t) = \sqrt{2D_{ox}t}$.

It must be noted that we ignore the reaction phase equation given by (8), which captures the rapid initial rise in the number of interface traps. Fig. 2 shows the extrapolated shape of the curve (using dotted lines) from a numerical simulation, for the case where the reaction phase is ignored in the model, and merely the diffusion phase is considered. The results show that ignoring the reaction and equilibrium phases leads to an underestimation in N_{IT} initially, as shown in Fig. 2. However, the mechanism is clearly diffusion limited, and we are interested in determining the impact of NBTI after a few years of operation. Hence, an underestimation in the number of interface traps for up to 1s does not affect the overall accuracy of the model, or the long-term shape of the N_{IT} curve.

B. Diffusion in Poly

Assuming that τ is greater than t_1 (the case where $\tau < t_1$ is handled later), the diffusion front moves into polysilicon as well, as shown in Fig. 3(e), although the diffusion coefficient for H_2 in poly (denoted as D_p), is lower than that in the oxide [35]. The detailed derivation is presented in Appendix A, and only the end result is shown here. Thus, from (21) and (53), the number of interface traps for the first stress phase is given by:

$$N_{IT}(t, 0 < t \leq t_1) = k_{IT}(2D_{ox}t)^{\frac{1}{6}}$$
$$N_{IT}(t, t_1 < t \leq \tau) = k_{IT}\Big[d_{ox}(1 + f(t)) + \sqrt{2D_p(t - t_1)}f(t)\Big]^{\frac{1}{3}} \quad (22)$$

$$f(t) = \left[\frac{D_{ox}\sqrt{2D_p(t - t_1)}}{D_{ox}\sqrt{2D_p(t - t_1)} + D_p d_{ox}}\right] \approx 1 \text{ for } t > t_1 \quad (23)$$

where the first equation accounts for diffusion in the oxide leading to a rapid stress phase, followed by the second equation which involves diffusion in poly, and therefore, a slower stress phase.

Using the above equations, it is easy to obtain an analytical expression for the number of interface traps for the static NBTI stress case, or the DC stress case as follows:

$$N_{IT_{DC}}(t, 0 < t \leq t_1) = k_{IT}(2D_{ox}t)^{\frac{1}{6}}$$
$$N_{IT_{DC}}(t, t > t_1) = k_{IT}\Big[d_{ox}(1 + f(t)) + \sqrt{2D_p(t - t_1)}f(t)\Big]^{\frac{1}{3}} \quad (24)$$

Simulation results for the DC stress case, using the above model are shown in Section VII - Fig. 12.

V. Numerical Simulation for the First Stress and Recovery Phases

Before deriving an analytical model for the first recovery phase as shown in Fig. 1, we present a detailed numerical analysis and solution to this case. This section aims to identify the origin of the drawbacks of the recovery modeling in [4], and argues that these are not necessarily a limitation of the R-D mechanism itself, as contended in [17]. Accordingly, a modified R-D model for recovery, based on the model in [4] is developed in Section VI. It must be noted that numerical simulation is only used to aid the reader in understanding the development of the actual mathematical model for the recovery phase. The employment of such a numerical simulation-based model is prohibitively computationally intensive, particularly in a multi-cycle framework to estimate the asymptotic impact of NBTI on transistor threshold voltage after three years ($\approx 10^{17}$ cycles at a frequency of 1GHz) of operation.

We present a numerical solution framework for the R-D model equations, described in Section III-A. We provide an in-depth analysis of the recovery modeling in [4], and show that the value of the back-diffusion coefficient $\xi = 0.5$, as used in [4] is not universal, and ξ is actually based on curve-fitting. We argue that the poor fit between the analytical model in [4] and measured data is partly due to the misinterpretation of the value of ξ as being universal, and not the R-D model itself.

We then explore the impact of using a two-region model considering the finite thickness of the gate-oxide, and a higher value of the diffusion constant in oxide, as compared with poly [35]. We show simulation results using this finite-oxide thickness-based model for NBTI recovery, and argue that the model further helps eliminate the previously encountered limitations in using the R-D theory based models.

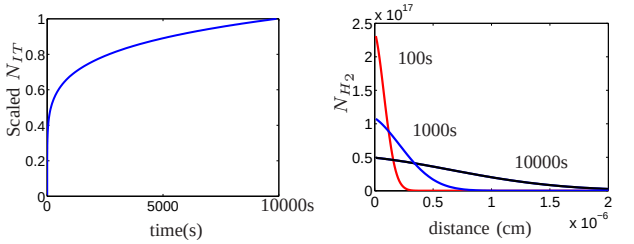
A. Simulation Setup

A backward-Euler numerical solver based on [43] is implemented with adaptive time stepping, using k_f = 4.66s^{-1}, k_r = 4.48e-9cm^{3}s^{-1}, k_H = 1.4e-3s^{-1}, N_0 = 5e12cm^2, and D_{ox} = 4e-17cm^{2}s^{-1}. It must be noted that the exact values

do not influence the time-dependencies [41]. A minimum step-size of 1e-4s is used for the simulations. We assume that there is a one-one correspondence between ΔV_{th} and N_{IT} for each of the cases, and that the y-axis, which denotes the normalized N_{IT} values (marked as "Scaled N_{IT}" in the figures), may also be interpreted as the normalized V_{th} values. The results are shown in the following subsections:

B. DC Stress

We first present the simple case of applying a DC stress on the PMOS transistors for 10000s. Fig. 4(a) shows the growth of N_{IT} with time, while Fig. 4(b) shows the evolution of the diffusion front with time, for t = [100s, 1000s, 10000s]. The tip of the diffusion front grows as $\sqrt{t}$ and the peak concentration decreases, while N_{IT} increases asymptotically as $\propto t^{\frac{1}{6}}$. Both results are consistent with the findings of the analytical model, detailed in Section IV.



(a) N_{IT} for stress phase. (b) Evolution of H_2 diffusion front at 100s, 1000s, and 10000s.

Fig. 4. Trap generation and H_2 diffusion for DC stress.

C. Effect of Stopping Stress

Fig. 5 shows the evolution of the diffusion front where stress was applied until time τ = 10000s, followed by diffusion of existing hydrogen molecules for time $t > \tau$. The results show that the peak concentration of hydrogen at the interface reduces, whereas the tip of the diffusion front continues to grow as $\sqrt{t}$. The shape of the diffusion front, and the decrease in $N^0_{H_2}$ for $t > \tau$ is obvious since there is no further generation or annealing of interface traps, and the increase in the base of the triangular front must be accompanied by a decrease in its height.

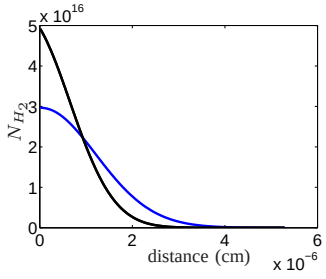


Fig. 5. Evolution of diffusion front for 10000 seconds of stress followed by diffusion of existing species: upper curve shows the front after τ = 10000s, while the lower curve plots the case where diffusion of existing species occurs after 10000s of stress, with a lowering of the peak concentration, and widening of the tip of the diffusion front $x_d(t)$.

Recovery is modeled as a superposition of two mechanisms:

1) Continued diffusion of existing hydrogen molecules away from the interface.
2) Annealing of interface traps, and backward diffusion of hydrogen molecules near the interface.

Thus, we have:

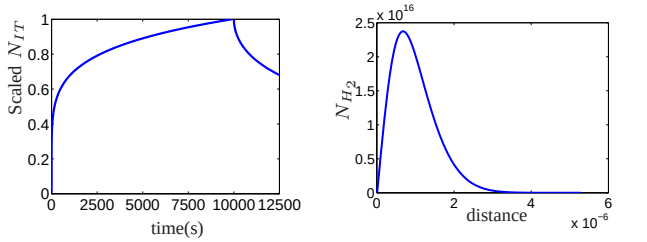
$$N_{IT}(t, t > \tau) = N_{IT}(\tau) - N^*_{IT}(t) \tag{25}$$

where N^*_{IT} is the annealed component.

In the absence of annealing, i.e., if k_r = 0 (along with k_f = 0) during the recovery phase, the profile of hydrogen molecular diffusion must be as shown in Fig. 5. Hence, the area under both curves in Fig. 5 is the same, and is given by:

$$N_{IT}(t, t \geq \tau) \propto x_d(t) N^0_{H_2}(t) = x_d(\tau) N^0_{H_2}(\tau) \tag{26}$$

D. Impact of Annealing



(a) N_{IT} for stress phase. (b) Evolution of H_2 diffusion front.

Fig. 6. Trap generation for AC stress case: 10000s of stress followed by 2500s of recovery.

In order to determine the impact of annealing, we first simulate the case where 10000s of stress followed by 2500s of recovery is applied to the PMOS device. Fig. 6(a) shows the decrease in N_{IT} beyond 10000s, with $N_{IT}(1.25\tau = 12500) = 0.673 N_{IT}(\tau = 10000)$, whereas Fig. 6(b) shows the diffusion front, where there is annealing close to the interface. The peak concentration point moves away from the interface, unlike the diffusion curves in the stress phase, which resemble a right angled triangle. However, the tip of the diffusion front continues to grow further into the oxide.

Fig. 7 shows the diffusion front after 2500s of recovery (Fig. 6(b)), superimposed on the diffusion front for the case where the device is stressed for 10000s, followed by continued diffusion (without annealing) for the remaining 2500s, as explained in Section V-C. The area under the black curve, denoted as **diffusing front** represents $N_{IT}(\tau)$, as explained in (26), whereas the area under the shaded curve (in blue) is $N_{IT}(t > \tau)$, and is denoted as the **existing front**. In Fig. 7, the region under the triangular shape filled with (red) vertical lines, denoted as **backward front** indicates the number of interface traps annealed, given by N^*_{IT}. Assuming that all fronts are triangular, which is reasonably accurate based on Fig. 7, we can write[4]:

[4]Number of interface traps N_{IT} is equal to twice the area under the N_{H_2} curve, from (13).

$$
\begin{aligned}
N_{IT}(\tau) &= N_{H_2}(x=0,t,t>\tau)\sqrt{2D(t+\tau)} \\
N^*_{IT}(t) &= N_{H_2}(x=0,t,t>\tau)x^*(t) \\
N_{IT}(t,t>\tau) &= N_{H_2}(x^*(t),t)\sqrt{2D(t+\tau)} \\
N_{IT}(t,t>\tau) &= N_{IT}(\tau) - N^*_{IT}(t) \qquad (27)
\end{aligned}
$$

where $x^*(t)$ is the point at which the diffusion front during the recovery phase reaches its peak. Unlike the figures in [14], where the authors assume that the peak value occurs at $\Delta \approx 0$, i.e., close to the Si-SiO$_2$ interface, $x^*(t)$ grows with time, i.e., the peak point moves away from the interface, due to forward diffusion of existing hydrogen species.

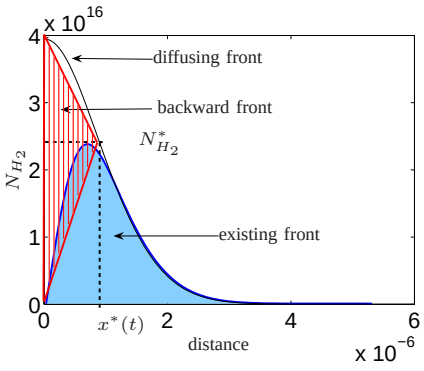
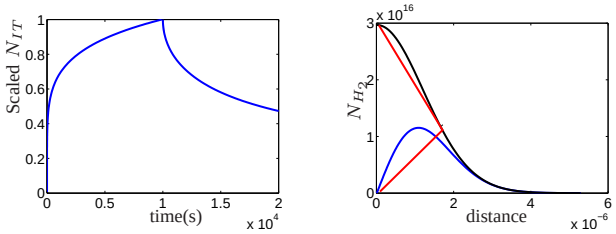


Fig. 7. Diffusion fronts during recovery.

Fig. 8(a) shows the case for τ seconds of stress followed by τ seconds of recovery, where τ = 10000s (as this case is widely used to compare the performance of an analytical model, as well as to demonstrate experimental results). The shape of the fronts indicate that the number of interface traps can be expressed as a difference in the area of the two triangles between the diffusing front, and the backward front, as shown in Fig. 7, and derived in (27). We now derive the analytical modeling in [4] using (27).



(a) N_{IT} for stress-relaxation phases. (b) Evolution of H_2 diffusion front.

Fig. 8. Trap generation for AC stress case - 10000s of stress followed by 10000s of recovery.

Numerical simulations plotted in Fig. 8(a), for this case show that:

$$N_{IT}(2\tau) = 0.47 N_{IT}(\tau) \qquad (28)$$

From Figs. 6(b) and 8(b), we can see that $x^*(t) \propto \sqrt{t}$, and can be written as:

$$x^*(t) = \sqrt{\xi \times 2Dt} \qquad (29)$$

where ξ is the curve-fitting parameter whose value must be determined. Using the above relation in (27), we have:

$$
\begin{aligned}
N_{IT}(\tau) &= N_{H_2}(x=0,t,t>\tau)\sqrt{2D(t+\tau)} \\
N^*_{IT}(t) &= N_{H_2}(x=0,t,t>\tau)\sqrt{2\xi Dt} \\
N_{IT}(t,t>\tau) &= N_{IT}(\tau) - N^*_{IT}(t) \\
&= N_{IT}(\tau) - \frac{N_{IT}(\tau)\sqrt{2\xi Dt}}{\sqrt{2D(t+\tau)}} \\
&= N_{IT}(\tau)\left[1 - \sqrt{\frac{\xi t}{t+\tau}}\right] \qquad (30)
\end{aligned}
$$

which is the equation for recovery in [4]. Substituting the value of $N_{IT}(2\tau)$ from (28) in (30), we have:

$$0.47 N_{IT}(\tau) = N_{IT}(\tau)\left[1 - \sqrt{\frac{\xi\tau}{\tau+\tau}}\right] \qquad (31)$$

from which, we obtain ξ = 0.58, which is the theoretical value of ξ for double sided diffusion, as stated in [4]. However, for simplicity, a fixed value of ξ = 0.5 is used, which results in $N_{IT}(2\tau) = 0.5 N_{IT}(\tau)$.

We now compare the values of the analytical model for recovery using (30) and the results from numerical simulations, for different values of t, with a fixed value of ξ = 0.58. Table I shows the values of $\frac{N_{IT}(t+\tau)}{N_{IT}(\tau)}$, i.e., the fractional recovery numbers during the relaxation phase, computed using numerical simulations, and using the analytical model from (30) with ξ = 0.58, for different values of t, where τ = 10000s. The last

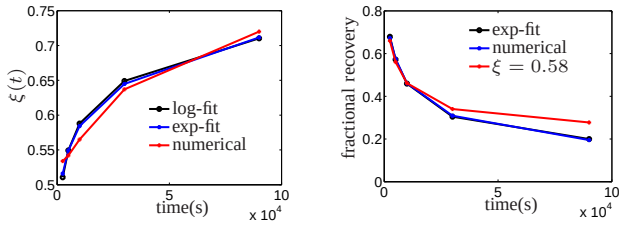
TABLE I

COMPARISON BETWEEN FRACTIONAL RECOVERY NUMBERS OBTAINED THROUGH NUMERICAL SIMULATIONS AND ANALYTICAL MODEL

time (t)	analytical	numerical	new value of ξ
2500	0.659	0.673	0.534
5000	0.560	0.575	0.542
10000	0.468	0.468	0.580
30000	0.340	0.309	0.637

column of Table I recomputes ξ from (30) by substituting the value of $N_{IT}(t+\tau)$ for each case. The results show that ξ is not a constant, and increases with t. However, for $t < \tau$, the difference between numerical and analytical results using ξ = 0.58 is not large. Thus, the discrepancy between numerical simulation results and analytical modeling for the recovery phase, for large values of t, is clearly attributed to the use of a fixed value of ξ, based on curve fitting at one time stamp $t = \tau$. This discrepancy can be resolved by using a curve-fitted expression for ξ, as shown in Fig. 9. Two sample curve fitted expressions and their accuracies are shown in Fig. 9(a), while the corrected model is plotted in Fig. 9(b) along with numerical data, as well as the case where ξ = 0.58 is used. The results indicate that with a time varying ξ, a good fit between numerical and analytical results can be obtained. Such a modified analytical solution from an R-D theory based on [4], with a time varying ξ does indeed converge well with numerical simulation results. It must be noted that the curve fitted expression for ξ in Fig. 9 is one of many choices, and is

merely shown to illustrate the usage of a time-varying model for ξ[5].



(a) Curve-fit for ξ. (b) Fractional recovery using time-varying curve-fitted ξ.

Fig. 9. Curve-fitted expressions for time varying ξ using an exponential relation ($\xi = 0.5843\left(\frac{t}{\tau}\right)^{.0897}$) and a log relation ($\xi = .0557log\left(\frac{t}{\tau}\right) + 0.5879$).

Simulation results also show that the value of ξ depends on τ, as well, particularly for smaller values of τ. Hence, any comparison of recovery models with the R-D theory based analytical model expression of [4] must be done using the appropriate value of ξ.

E. Finite Oxide Thickness

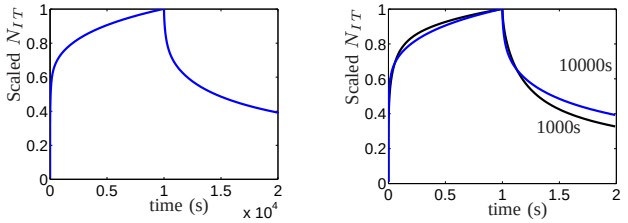
In this section, we propose to account for further discrepancies between the findings from a numerical or an analytical model and experimental data, such as:

1) Experimental results for a single stress phase followed by a single recovery phase show more than 80% recovery in [20] for τ = 1000s, around 60% recovery in [35] for τ = 10000s, and 50% recovery in [4] for τ = 1000s, for devices with an oxide thickness of 1.2nm-1.3nm.
2) Larger fractional recovery for the same value of τ for a higher oxide thickness is seen in [35].
3) Rapid decrease in V_{th} at the beginning of the recovery phase [20], implying a $logt$ behavior for recovery, where equal recovery is observed in every decade [17]–[19][6].

Accordingly, a finite oxide thickness-based two-step model is contended since the diffusion constant of hydrogen in oxide is larger than that in polysilicon ($D_{ox} > D_p$). Although the exact values of D_{ox} and D_p are still widely debated [35], their relative ratio influences the shape of the N_{IT} curve. We perform numerical simulations, using our setup, as described in Section V-A for a case where d_{ox} = 1.3nm. Additional boundary conditions at the oxide-poly interface are added to the numerical simulation setup used for the infinitely thick oxide case, in Section V-A. D_p is assumed to be $0.25D_{ox}$. Fig. 10(a) which plots the simulation results shows that there is approximately 60% recovery after τ seconds of recovery for τ = 10000s, as opposed to Fig. 8(a) which shows 50% recovery.

[5]Both the curve-fitted expressions in Fig. 9 do not guarantee that ξ converges to 1, as $t \rightarrow \infty$, and may require to be further modified for the case of a single stress phase followed by recovery of the device for infinite time, thereby resetting it to be equivalent to an original unstressed device. However, these expressions are merely shown to illustrate the fact that ξ is a function of t, and is not a constant.

[6]It must be noted that [40] has attributed this behavior to an inaccurate way of estimating the impact of NBTI by using UFM techniques.

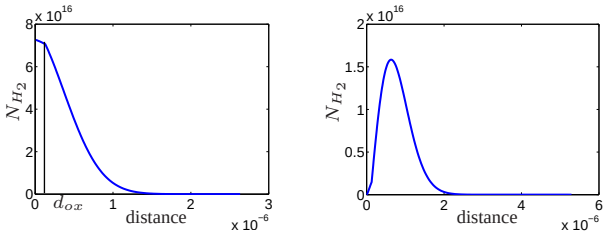


(a) Finite oxide thickness-based model. (b) Fractional recovery for different values of τ.

Fig. 10. Validation of finite oxide thickness-based model.

Fig. 10(b) shows the model for the case of τ = 10000s, and τ = 1000s, with higher fractional recovery for the 1000s case, since more H_2 is contained in the oxide, and rapidly diffuses back to the interface. Unlike the infinite oxide thickness case, which would have incorrectly predicted a fractional recovery of $\approx$ 50% for both τ = 10000s, and τ = 1000s, higher fractional recovery is seen with lower values of τ. The shape of the diffusion profile at the end of the first stress and recovery phases for the case of τ = 10000s, and D_{ox} = 4 D_p are shown in Fig. 11. Fig. 11(a) shows the diffusion of N_{H_2} at the end of the stress phase, with the rectangular shaped front in the oxide, followed by a triangular front in poly. The diffusion profile for recovery in Fig. 11(b) indicates that the fraction of the hydrogen molecules contained in the oxide quickly diffuses backwards during recovery.

Thus, it is clear that a two-region based model for recovery with differing diffusion constants for oxide and poly is necessary to model the recovery phase of NBTI action. Accordingly, we also use two curve fitting constants ξ_1 and ξ_2, for the backward diffusing fronts in oxide and poly, respectively, and determine the values of these constants to match the experimental results. The development of the analytical model for recovery is detailed in the next section.



(a) Diffusion front for stress phase. (b) Diffusion front at the end of the recovery phase.

Fig. 11. Diffusion front considering finite oxide thickness.

VI. Model for The First Recovery Phase

During the recovery phase, the stress applied to the PMOS device is released, as shown in Fig. 1. Some of the hydrogen molecules recombine with Si^+ species, to form Si-H bonds, thereby annealing some of the existing traps. Since the rate of diffusion of hydrogen molecules in the oxide is greater than that in poly, rapid annealing of traps occurs in the oxide, followed by a slow annealing in polysilicon. Accordingly, we

have two stages of recovery in each relaxation phase, that are modeled separately:

A. Recovery in Oxide

The detailed derivation for the first recovery phase of NBTI action is shown in Appendix B. The final equation is of the form:

$$N_{IT}(t+\tau, 0 < t \leq t_2) = \frac{N_{IT}(\tau)}{1+g(\xi_1, t)} \tag{32}$$

where t_2 is the time when the back-diffusion front has reached the oxide-poly interface, and is of the order of less than a second, while:

$$g(\xi_1, t) = \left[\frac{\sqrt{2\xi_1 D_{ox} t}}{2d_{ox} - \sqrt{2D_{ox}t} + \sqrt{2D_p(t+\tau)}}\right] \tag{33}$$

with the value of ξ_1, which is a function of t, τ, and d_{ox}, chosen appropriately using curve fitting, based on the discussion in Sections V-D and V-E.

B. Slow Recovery in Poly

If recovery continues beyond time t_2, the back-diffusion front now enters poly, where its growth is slower, in comparison with that in the oxide ($\equiv$ to the diffusion front during the first stress phase in Fig. 3). Hence, during this phase, the rate of annealing of interface traps reduces. However, by this time, since the oxide is almost completely annealed, only a slow recovery in poly occurs. The diffusion front in poly is triangular, and its peak moves further away from the oxide-poly interface as being proportional to $\sqrt{\xi_2 t}$ where ξ_2 is the curve fitting parameter. The mechanism is similar to recovery for the case of an infinitely thick oxide. Hence, the model derived in Section V-D for the infinite oxide case, can be used here. Thus, we have:

$$N_{IT}(t+\tau, t > t_2) = N_{IT}(\tau + t_2)\left[1 - \sqrt{\frac{\xi_2(t-t_2)}{t+\tau}}\right] \tag{34}$$

for time $\tau + t_2$ to 2τ, where ξ_2 is the curve fitting factor. It must be noted that due to the difference in the coefficients of oxide and poly, and the slow progression of the back-diffusion front in poly, the value of ξ is less than 0.58, and is of the order of around 0.125 for $t < t_0$[7]. Thus, the two-step model for annealing consists of a quick annealing stage where the number of interface traps decreases rapidly in the first few milliseconds to about a second, followed by a slow decrease over the remaining time period.

The model proposed can thus also account for rapid recovery during the beginning of the relaxation stage, due to mechanisms not attributed to a reaction-diffusion process, using the curve fitted value of ξ_1. The authors in [40] argue that the rapid decrease in V_{th} at the beginning of the recovery phase, that does not correspond to a simultaneous decrease in N_{IT}, is an incorrect manifestation of the UFV technique used to measure recovery in PMOS devices. While it is not clear what the actual physical mechanism is, in nanometer scale PMOS devices during actual circuit operation, the use of a curve-fitted ξ_1 helps fit better the results of the model with experimental data, while still adhering to the basic guidelines of the R-D theory.

[7]A time varying ξ_2, as deemed necessary in Section V-D is used to model the impact of a single stress phase, followed by long periods of recovery, in the plots (Fig. 17) shown later on, in Section VII-D.

C. Complete Set of Equations for First Stress and Relaxation Phase

The equations for the first stress and relaxation phase can be summarized as follows:

$$\begin{aligned}
N_{IT}(t, 0 < t \leq t_1) &= k_{IT}(2D_{ox}t)^{\frac{1}{6}} \\
N_{IT}(t, t_1 < t \leq \tau) &= k_{IT}\Big[d_{ox}(1+f(t)) + \\
&\quad \sqrt{2D_p(t-t_1)}f(t)\Big]^{\frac{1}{3}} \\
N_{IT}(t+\tau, 0 < t \leq t_2) &= \frac{N_{IT}(\tau)}{1+g(\xi_1, t)} \\
N_{IT}(t+\tau, t_2 < t \leq \tau) &= N_{IT}(\tau + t_2)\left[1 - \sqrt{\frac{\xi_2(t-t_2)}{t+\tau}}\right]
\end{aligned} \tag{35}$$

VII. Simulation Results and Sanity Check Plots

In this section, we compare the results of our model with the requirements outlined in Section II.

A. DC Stress

The plot for a DC stress case, for a PMOS transistor with d_{ox} = 1.2nm is obtained using (24), and is shown in Fig. 12. The plot consists of three significant phases:

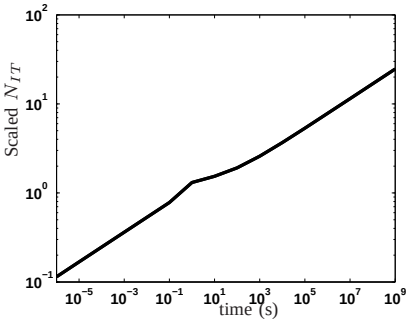


Fig. 12. Plot of DC stress for d_{ox} = 1.2nm. The curve plots the normalized interface trap values for k_{IT} = 1.

1) The initial phase of $t < 0.1$s, during which the reaction phase is dominant. It must be noted that this phase has been not been explicitly modeled in (35), and (21) is used for $t \geq 0$, as has been explained in the end of Section IV-A, using Fig. 2.
2) The transient phase of 0.1s $\leq t <$ 10s, during which the process is dominated by diffusion in the oxide.
3) The final phase, for large values of t, over which the mechanism is dominated by diffusion in poly.

It follows from the shape of the log-log plot in Fig. 12, that as t increases, the number of interface traps asymptotically

approaches a $t^{\frac{1}{6}}$ relationship, which satisfies the first guideline outlined at the end of Section II. It must be noted that in the analytical model for DC stress, and hence the plots in Fig. 12, we ignore the reaction and the quasi equilibrium phases of interface trap generation, for reasons already explained in Section III-B.

B. AC Stress (Single Stress phase followed by a single relaxation phase)

The plot in Fig. 13 shows the simulation results for the number of interface traps generated for a single stress phase, followed by a relaxation phase, each of duration τ = 10000s, using (35), for a PMOS device whose oxide thickness (d_{ox}) is 1.2nm. These match the values used in the experimental setup from [35]. The values of ξ_1 and ξ_2 are chosen based on curve fitting, with $\xi_1 >> \xi_2$. The results of our simulation are shown in Fig. 13. The curve shows a good fit with experimental data from [35], [44]. The accurate fit with experimental data, particularly during the recovery phase, satisfies the second requirement outlined in Section II.

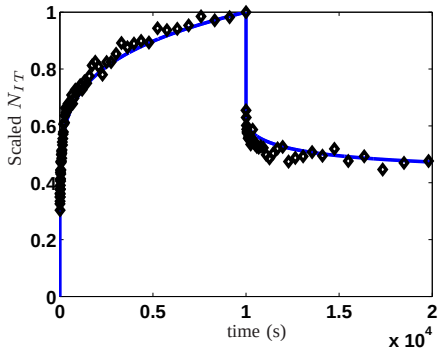


Fig. 13. Plot of first stress and recovery phases for τ = 10000s, and d_{ox} = 1.2nm, with experimental data from [35], [44], shown in ⋄, on a linear scale.

Recent publications [18], [19] have motivated the plotting of stress and relaxation data, on a semi-log scale, to compare the accuracy of the fit, over the broad spectrum of time constants. The fit with experimental data from [35] on a semi-log scale is shown in Fig. 14. Fig. 14(a) shows the plot for the first stress and recovery phases, while Fig. 14(b), for the recovery phase only. The fit for our model is not very accurate, during the beginning of the stress phase, as seen from Fig. 14(a), and our model shows a higher exponent as opposed to experimental data. Recently published works [29], [30], [41] have shown that this is nevertheless consistent with a H_2 diffusion based R-D model, and attribute this discrepancy in short-term measurements to the assumption that H-to-H_2 conversion is extremely fast, which may not be realistic [41]. A detailed analysis of the H↔H_2 conversion has been incorporated into an analytical model recently by [29], and the fit of the model with the experimental data indeed verifies that this is true. The shape of the plots from [41] are similar to that shown in Fig. 14, with measurement data showing an initial slope of $t^{\frac{1}{3}}$, whereas the R-D model solution using only H_2 diffusion predicts a $t^{\frac{1}{6}}$ behavior. However, for the purposes of circuit delay degradation estimation and optimization, we are more concerned about long term effects of aging after a few years of circuit operation under various conditions, rather than actual cycle accurate values. In this context, the accuracy of the plot toward the end of the stress phase, and the asymptotic fit is more important, since this governs the shape of the next recovery phase, and the subsequent stress phases.

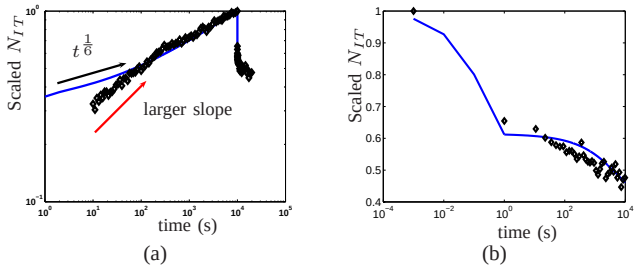


Fig. 14. Plot of first stress and recovery phases for τ = 10000s, and d_{ox} = 1.2nm, with experimental data from [35], [44], shown in ⋄, on a log scale: (a) shows the plot for both the phases, while (b) shows the plot for the recovery phase only, as a function of the time of recovery ($t - \tau$).

C. Effect of thicker oxides

Experimental results have shown that as the oxide thickness increases, greater amount of recovery is expected. We verify this by simulating the case of d_{ox} = 2.2nm, and τ = 10000s. While the d_{ox} = 1.2nm case showed ≈60% recovery after τ seconds of relaxation, we expect a higher fractional recovery for this case, since more N_{H_2} is contained in the oxide, and hence diffuses back faster. The results are shown in Fig. 15, and expectedly there is 80% recovery after τ seconds of relaxation. The results match well with experimental data from [35], thereby satisfying the third requirement in Section II.

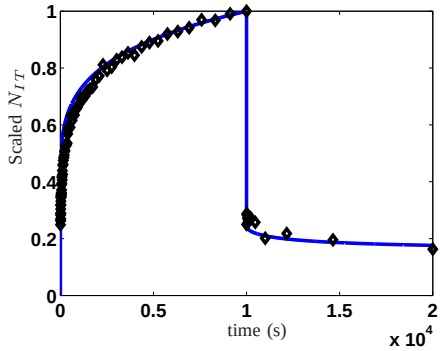


Fig. 15. Plot of first stress and recovery phases for τ = 10000s, and d_{ox} = 2.2nm, with experimental data from [35], [44], shown in ⋄, on a linear scale.

D. Effect of lower stress times on the amount of recovery

Previous solutions to the R-D model ignored the effect of finite oxide thickness, and the difference in the diffusion rates in polysilicon and the oxide. Hence, these results always

showed 50% recovery, when the ratio of recovery time to stress time was one, independent of the stress time. However, experimental results [20] show that a higher fractional V_{th} recovery is observed for lower stress times. We verify this by plotting the results for the case of t_{ox} = 1.2nm, with stress times of 10000s and 1000s, respectively, in Fig. 16.

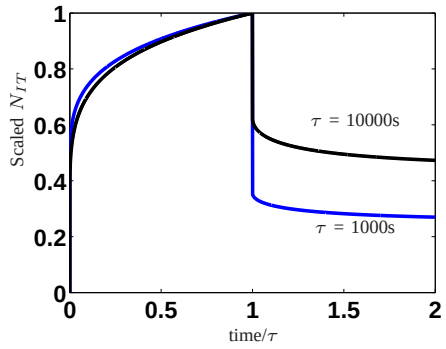


Fig. 16. Plot of the first stress and recovery phase for τ = 10000s, and τ = 1000s, showing the effect of reduced stress times.

Further, we also use compare the results of our model with experimental data from [20], for the case of a single stress phase followed by variable amounts of recovery, for different values of τ. Fig. 17 shows the case where a single stress phase was followed by 100 seconds of recovery for four cases of stress times: 1000s, 100s, 10s, and 1s, respectively, for a 1.3nm oxide case.

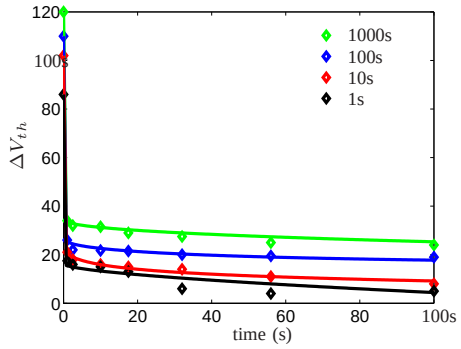


Fig. 17. Experimental data from [20] compared with model results to demonstrate the effect of reducing τ.

Fig. 17 indicates that our two-stage model for recovery with two sets of curve fitted constants ξ_1 and ξ_2 provide a reasonably accurate fit with the experimental results. Some key findings are:

1) The plots in [45] shows results where a 1000s of NBTI stress on a PMOS device with an oxide thickness of 1.3nm causes a threshold voltage shift of 30mV. Subsequent recovery causes an approximate 50% reduction in the amount of V_{th} degradation. However, the results in [20] show approximately 120mV increase in V_{th} with 1000s of stress, and a large amount of recovery as well, after 100s of relaxation.
2) The curve fitted value of ξ_1 is largest for the case of 1000s of stress, and decreases with a reduction in the value of τ.
3) A single value of ξ_2 suffices for the τ = 1000s and τ = 100s cases of stress, followed by 100s of recovery, (since t =100s is $\leq \tau$ for these two cases). However, a curve fitted expression for ξ_2 of the form $\xi_{2_0}(\frac{t}{\tau})^{\alpha}$ is used for the cases of 10s and 1s of stress followed by continuous recovery for 100s, since $t >> \tau$.
4) The value of ξ_2 decreases with τ, as well. This can be explained as follows. For the case of 1000s of stress, the tip of the diffusion front is well into the polysilicon region, implying that the base of the triangular diffusion front is large, and its height relatively narrower. Hence, with 100s of recovery, the back diffusion front moves deeper into the poly region with its narrower height - as compared with the 100s case, implying a larger ξ_2 for a larger τ.

Thus, our model satisfies the guidelines outlined in Section II (the last observation about frequency independence is deferred to Section VIII-C), and provides reasonably accurate fits with experimental data. We now present the extension of our single cycle model, to a multi-cycle operation, i.e., we calculate the number of interface traps for any k^{th} stress or relaxation phase, assuming the input pattern in Fig. 1.

VIII. Extension for Multi-cycle and High-frequency Operation

The detailed derivation for the second stress and recovery phases are shown in Appendix C. The plot for the first two stress and relaxation phases is shown in Fig. 18.

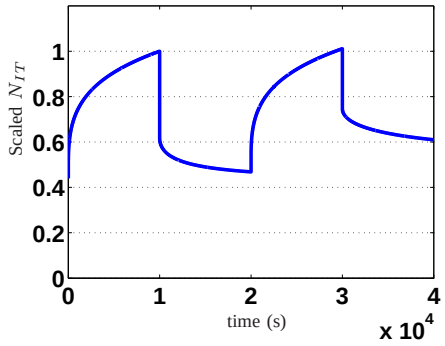


Fig. 18. Plot of the first two stress and recovery phases for τ = 10000s, and d_{ox} = 1.2nm.

The figure shows the number of interface traps rapidly increasing during the beginning of the second stress phase, because of rapid dissociation of the $Si - H$ bonds, which is consistent with the results in [4]. Recovery during the second relaxation phase is expectedly less than that during the first relaxation phase, since the peak concentration has now decreased, due to further diffusion of hydrogen molecules into the poly region.

A. Comparison with Experimental Results

We also compare the results of our multicycle model with some published experimental results. Fig. VIII-A shows the model results for the first stress phase, first recovery phase, as well as the second stress phase for a 1.3nm oxide thickness case, and τ = 1000s. Experimental results from [45] for this case indicate a 50% recovery after τ seconds of relaxation. Fig. VIII-A shows that the fit is reasonably accurate.

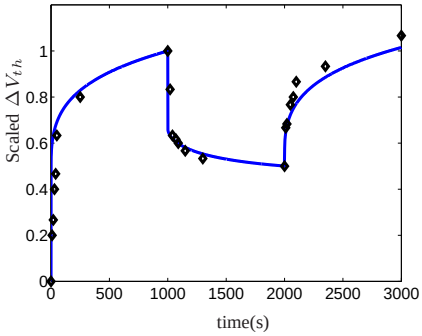


Fig. 19. Comparison of experimental data and model results for subsequent stress and relaxation phases.

B. Final Simplified Model and Range of Operation

For a multi-cycle periodic operation, where an AC stress is applied on the PMOS device, with stress time being the same as relaxation time, both being equal to τ, as shown in Fig. 1, we obtain the following expressions for the $(n+1)^{th}$ cycle, consisting of stress from time $2n\tau$ to $(2n+1)\tau$, and relaxation from time $(2n+1)\tau$ to $2(n+1)\tau$, respectively:
Stress Phase:

$$\begin{aligned} N_{IT}(2n\tau+t, 0<t\leq t_1) &= k_{IT}\left[\left(\frac{N_{IT}(2n\tau)}{k_{IT}}\right)^6 + 2D_{ox}t\right]^{\frac{1}{6}} \\ N_{IT}(2n\tau+t, t_1<t\leq \tau) &= k_{IT}\left[\sqrt{\left(\frac{N_{IT}(2n\tau)}{k_{IT}}\right)^6 + (2d_{ox})^2} + \right. \\ &\quad \left.\sqrt{2D_p(t-t_1)}\right]^{\frac{1}{3}} \end{aligned} \quad (36)$$

Relaxation Phase:

$$\begin{aligned} N_{IT}((2n+1)\tau+t, 0<t\leq t_2) &= \frac{N_{IT}((2n+1)\tau)}{1+h_1(\xi_1,t)} \\ N_{IT}((2n+1)\tau+t, t_2<t\leq \tau) &= N_{IT}((2n+1)\tau+t_2)\left[1-h_2(\xi_2,t)\right] \end{aligned}$$

$$\begin{aligned} \text{where } h_1(\xi_1,t) &= \left[\frac{\sqrt{\xi_1\times 2D_{ox}t}}{2d_{ox}-\sqrt{2D_{ox}t}+\sqrt{2D_p(t+(2n+1)\tau)}}\right] \\ h_2(\xi_2,t) &= \left[\sqrt{\frac{\xi_2(t-t_2)}{t+(2n+1)\tau}}\right] \end{aligned} \quad (37)$$

The above model is valid for $\tau > t_1$ and $\tau > t_2$, i.e., for $\tau >$ 1s. Simulation results using this model for τ = 10000s, for 10 years of operation are shown in Fig. 20. The results show that the number of traps produced by AC stress is about 0.7 times that produced by a DC stress. The shape of the curves also indicates that the asymptotic slopes of the two stress cases are the same. This is suggestive of the fact that AC stress can be modeled as a linear function of DC stress, for long term estimates, as explained in Section IX.

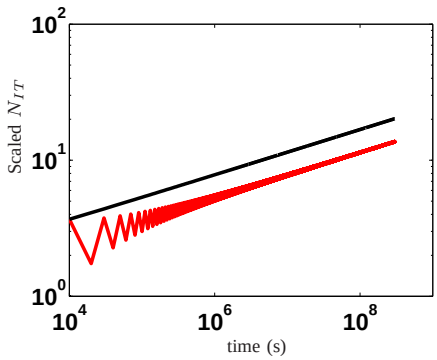


Fig. 20. Plot showing interface trap generation for τ = 10000s for AC and DC stress cases up to 10 years of operation, on a log-log scale.

C. NBTI Model for High Frequency Operation

For high frequency operation, the above multicycle model cannot be used, due to the underlying assumptions about the shape of the diffusion front, and the various approximations made during the course of the derivations. However, for a 1GHz frequency operation, it is computationally infeasible to compute the interface trap concentration on a cycle accurate basis for 10 years of operation, amounting to $\approx 10^{17}$ cycles, either using analytical models or through simulations. Hence we seek transformations of high frequency waveforms into extremely low frequency waveforms (say, of the order of $\leq$ 1Hz), thereby obtaining tractable and fairly accurate asymptotic estimates with a large speed-up. In this regard, we explore a key property of the dynamics of interface trap generation, namely, frequency independence.

Experimental results have shown that the number of interface traps, measured after a large duration of time is approximately the same irrespective of the actual frequency of the input AC waveform being applied [3], [4], [10], [14], [16], implying identical asymptotic N_{IT} estimates. This property is known as **frequency independence**. Although several differing experimental results have been observed, recent experiments have shown that this holds good over the 1Hz-1GHz bandwidth [38], which seconds the analytical findings in [16]. However, as we move closer to DC, some form of frequency dependence is expected. We verify this phenomenon by plotting the number of interface traps up to 10^6s for five different τ values differing by an order each, ranging from 1s to 10000s. The values are compared with the DC case as well, and the plots are shown in Fig. 21. The results show that with increasing τ, the N_{IT} curves tend to become closer. Hence, for τ = 1s, some form of frequency independence can be assumed to hold good asymptotically.

Thus, on the basis of experimental data from [38], and the trend seen in Fig. 21, we conclude that the interface trap count determined for τ = 1s, asymptotically equals the number for a case where τ = 1ns, over t_{life}, where t_{life} is the lifetime of the circuit, and is assumed to be 10 years of operation:

$$N_{IT}(t=t_{\text{life}}, \tau=1\text{s}) \approx N_{IT}(t=t_{\text{life}}, \tau=1\text{ns}) \quad (38)$$

Thus, we can use our multi-cycle model derived in the previous

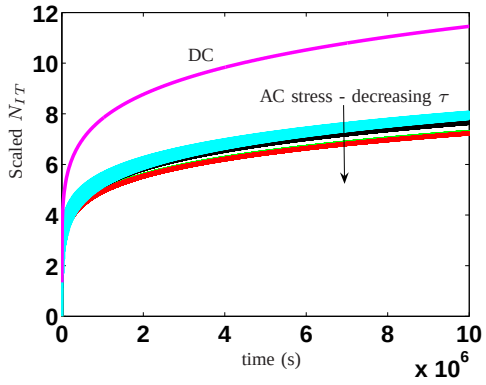


Fig. 21. Plot showing interface trap generation for different time periods, along with the DC stress case, to demonstrate frequency independence.

subsection, with τ = 1s, to estimate the impact of NBTI on gigascale circuits.

IX. A Framework for Estimating the Impact of NBTI on Circuit Delay

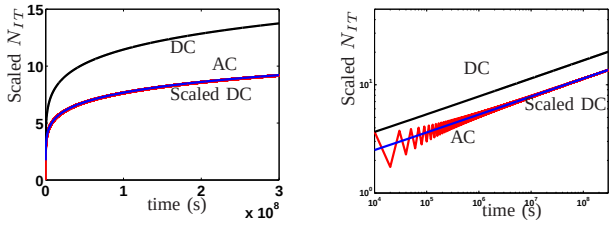
In this section, we present a framework for using the NBTI model to estimate the temporal delay degradation of digital circuits over 10 years of operation. We use the method described in [46], where the authors claim that AC NBTI can be represented as being asymptotically equal to some α times DC NBTI, where α represents the ratio between the number of interface traps for the AC and DC stress cases:

$$\alpha = \frac{N_{IT_{AC}}(t = t_{\text{life}} = N\tau)}{N_{IT_{DC}}(t, t = t_{\text{life}})} \tag{39}$$

where N denotes the number of half cycles, each of duration τ, in 10 years of operation. Accordingly, AC stress can be approximated as:

$$N_{IT_{AC}}(t, \tau < 1s) \equiv \alpha N_{IT_{DC}}(t) \tag{40}$$

where $N_{IT_{AC}}(t)$ is the number of interface traps due to AC stress, and $N_{IT_{DC}}(t)$, that due to DC stress, at time t. We verify this method graphically by plotting the actual AC waveform and the scaled DC waveform, where α is the ratio of the number of interface traps computed after 10 years of operation, for τ = 10000s, in Figs. 22(a) and (b). A good fit in the linear plot (Fig. 22(a)) guarantees correct estimates, for the circuit lifetime, ranging over the 1 year-10 year period.



(a) Linear Scale (b) Log Scale

Fig. 22. Plot showing AC stress represented as an equivalent scaled DC stress. The two curves almost perfectly overlap.

Although, the equivalent DC stress model may not provide an exact upper bound, especially, over the first few stress and relaxation phases, and may not show the exact transient response initially, the overall fit is fairly accurate for asymptotic NBTI estimates, over a period of time, as large as 10 years, as seen from Fig. 22(b). Since reliability estimates do not require cycle accurate behavior of the number of interface traps, the scaled DC model is simple and sufficient.

The above method in conjunction with frequency independence can be used to estimate the number of interface traps as follows:

1) Convert the high frequency waveforms to equivalent 1Hz waveforms, by using the SPAF method outlined in [14] or otherwise.
2) Calculate the number of interface traps up to 10 years of operation, for the 1Hz square waveform, and the DC waveform using the model.
3) Compute the value of α, and use the scaled DC model as an approximate temporal estimate of the number of interface traps, at various time stamps.
4) Repeat this method for waveforms of different duty cycles, and compute the value of α in each case, to obtain a simple look-up table of α versus signal probability (such that ΔV_{th} for each signal probability = some α times the ΔV_{th} for DC stress), as described in [14], or even a smooth curve fitting-based model, as desired.
5) Compute the number of interface traps and the V_{th} degradation at any desired time stamp, for any signal probability, using this scaled DC model.

Since N_{IT} is linearly proportional to V_{th}, experimental results can be used to compute this ratio, and the N_{IT} numbers can accordingly be converted to V_{th} values. We present a generic framework in our work, and hence, simply work with normalized N_{IT} values. A plot of V_{th} versus the probability that a PMOS device is stressed, computed using the method outlined above, is shown in Fig. 23. The figure shows an initial steep rise, since N_{IT} and ΔV_{th} are $\propto t^{\frac{1}{6}}$. A lookup table built using this figure can then be used to determine the sensitivity of gate delays to temporal degradation caused by aging, and thereby shifts in timing numbers can be estimated.

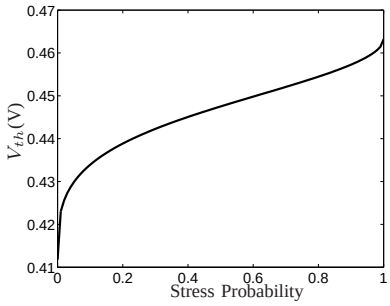


Fig. 23. PMOS V_{th}, after three years of aging, as a function of the probability that the transistor is stressed.

X. Conclusion

NBTI (Negative Bias Temperature Instability) is a growing threat to temporal circuit reliability and hence its accurate

estimation is essential for suitably guard-banding our designs. The dynamics of interface trap generation and annealing depend on a large number of complex factors, which can be analytically captured using the framework of Reaction-Diffusion (R-D) model. Existing NBTI models fail to account for all of these factors, particularly the effect of finite oxide thickness, and the role of the reaction phase during recovery, thereby leading to poor scalability, or an inaccurate fit with experimental data. We propose a new model for estimating the number of interface traps and suitably account for these effects in our model. A framework for using this model in a multi-cycle gigahertz operation is proposed, which can be used to estimate the temporal delay degradation of digital circuits.

REFERENCES

[1] V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, "Impact of Negative Bias Temperature Instability on Digital Circuit Reliability," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 248–253, April 2002.

[2] A. T. Krishnan, V. Reddy, S. Chakravarthi, J. Rodriguez, S. John, and S. Krishnan, "NBTI Impact on Transistor and Circuit: Models, Mechanisms and Scaling Effects," in *IEEE International Electronic Devices Meeting*, pp. 14.5.1–14.5.4, December 2003.

[3] D. K. Schroder, "Negative Bias Temperature Instability: Physics, Materials, Process, and Circuit Issues," in *IEEE Solid-State Circuit Society*, August 2005. Available at `http://www.ewh.ieee.org/r5/denver/sscs/Presentations/2005.08.Schroder.pdf`.

[4] M. A. Alam, "A Critical Examination of the Mechanics of Dynamic NBTI for pMOSFETs," in *IEEE International Electronic Devices Meeting*, pp. 14.4.1–14.4.4, December 2003.

[5] S. Chakravarthi, A. T. Krishnan, V. Reddy, C. Machala, and S. Krishnan, "A Comprehensive Framework for Predictive Modeling of Negative Bias Temperature Instability," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 273–282, April 2004.

[6] J. G. Massey, "NBTI: What We Know and What We Need to Know - A Tutorial Addressing the Current Understanding and Challenges for the Future," in *IEEE International Integrated Reliability Workshop Final Report*, pp. 199–211, October 2004.

[7] B. Zhu, J. S. Suehle, Y. Chen, and J. B. Bernstein, "Negative Bias Temperature Instability of Deep Sub-micron p-MOSFETs Under Pulsed Bias Stress," in *IEEE International Integrated Reliability Workshop Final Report*, pp. 125–129, October 2002.

[8] M. Ershov, R. Lindley, S. Saxena, A. Shibkov, S. Minehane, J. Babcock, S. Winters, H. Karbasi, T. Yamashita, P. Clifton, and M. Redford, "Transient Effects and Characterization Methodology of Negative Bias Temperature Instability in pMOS Transistors," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 606–607, April 2003.

[9] G. Chen, M. F. Li, C. H. Ang, J. Z. Zheng, and D. L. Kwong, "Dynamic NBTI of p-MOS Transistors and its Impact on MOSFET Scaling," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 196–202, April 2003.

[10] M. A. Alam, "On the Reliability of Micro-electronic Devices: An Introductory Lecture on Negative Bias Temperature Instability," in *Nanotechnology 501 Lecture Series*, September 2005. Available at `http://www.nanohub.org/resources/?id=193`.

[11] M. A. Alam and S. Mahapatra, "A Comprehensive Model of PMOS NBTI Degradation," *Journal of Microelectronics Reliability*, vol. 45, pp. 71–81, August 2004.

[12] K. O. Jeppson and C. M. Svensson, "Negative Bias Stress of MOS Devices at High Electric Fields and Degradation of MNOS Devices," *Journal of Applied Physics*, vol. 48, pp. 2004–2014, May 1977.

[13] S. Ogawa and N. Shiono, "Generalized Diffusion-Reaction Model for the Low-Field Charge-Buildup Instability at the Si-SiO2 interface," *Journal of Applied Physics*, vol. 51, pp. 4128–4230, February 1995.

[14] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "An Analytical Model for Negative Bias Temperature Instability (NBTI)," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 493–496, November 2006.

[15] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and Minimization of PMOS NBTI Effect for Robust Nanometer Design," in *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 1047–1052, July 2006.

[16] S. Bhardwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, "Predictive Modeling of the NBTI Effect for Reliable Design," in *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 189–192, September 2006.

[17] T. Grasser, W. Gos, V. Sverdlov, and B. Kaczer, "The Universality of NBTI Relaxation and its Implications for Modeling and Characterization," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 268–280, April 2007.

[18] H. Reisenger, O. Blank, W. Heinrigs, A. Muhlhoff, W. Gustin, and C. Schlunder, "Analysis of NBTI Degradation and Recovery Behavior Based on Ultra Fast Vt Measurements," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 448–453, April 2006.

[19] H. Reisenger, O. Blank, W. Heinrigs, W. Gustin, and C. Schlunder, "A Comparison of Very Fast to Very Slow Components in Degradation and Recovery due to NBTI and Bulk Hole Trapping to Existing Physical Models," *IEEE Transactions on Devices and Materials Reliability*, vol. 7, pp. 119–129, March 2007.

[20] C. Shen, M. F. Li, C. E. Foo, T. Yang, D. M. Huang, A. Yap, G. S. Samudra, and Y.-C. Yeo, "Characterization and Physical Origin of Fast Vth Transient in NBTI of pMOSFETs with SiON Dielectric," in *IEEE International Electronic Devices Meeting*, pp. 333–336, November 2006.

[21] J. H. Stathis and S. Zafar, "The Negative Bias Temperature Instability in MOS Devices: A Review," *Journal of Microelectronics Reliability*, vol. 46, pp. 270–286, February–April 2006.

[22] C. R. Parthasarathy, M. Denais, V. Huard, G. Ribes, E. Vincent, and A. Bravaix, "New Insights into Recovery Characteristics Post NBTI Stress," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 471–477, April 2006.

[23] S. Zafar, B. H. Lee, J. Stathis, A. Callegari, and T. Ning, "A Model for Negative Bias Temperature Instability (NBTI) in Oxide and High k pFETs," in *Digest of Symposium on VLSI Technology*, pp. 208–209, June 2004.

[24] B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, and M. Goodwin, "Disorder-Controlled Kinetics Model for NBTI and its Experimental Verification," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 381–387, April 2005.

[25] V. Huard, C. R. Parthasarathy, C. Guerin, and M. Denais, "Physical Modeling of Negative Bias Temperature Instabilities for Predictive Exploration," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 733–734, April 2006.

[26] V. Huard, M. Denais, and C. Parthasarathy, "NBTI Degradation: From Physical Mechanisms to Modeling," *Journal of Microelectronics Reliability*, vol. 46, pp. 1–23, January 2006.

[27] S. Mahapatra, K. Ahmed, S. Varghese, A. E. Islam, G. Gupta, L. Madhav, D. Saha, and M. A. Alam, "On the Physical Mechanism of NBTI in Silicon Oxynitride p-MOSFETs: Can Differences in Insulator Processing Conditions Resolve the Interface Trap Generation Versus Hole Trapping Controversy?," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 1–9, April 2007.

[28] A. E. Islam, H. Kufluoglu, D. Varghese, and M. A. Alam, "Critical Analysis of Short-Term Negative Bias Temperature Instability Measurements: Explaining the Effect of Time-Zero Delay for on-the-fly Measurements," *Applied Physics Letters*, vol. 90, pp. 3505–3508, February 2007.

[29] A. E. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra, and M. A. Alam, "Recent Issues in Negative Bias Temperature Instability: Initial Degradation, Field Dependence of Interface Trap Generation, Hole Trapping Effects, and Relaxation," *IEEE Transactions on Electron Devices*, vol. 54, pp. 2143–2154, September 2007.

[30] J. H. Lee, W. H. Wu, A. E. Islam, M. A. Alam, and A. Oates, "Separation Method of Hole Trapping and Interface Trap Generation and Their Roles in NBTI Reaction-Diffusion Model," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 745–746, April 2008.

[31] J. Keane *et al.*, "An On-Chip NBTI Sensor for Measuring PMOS Threshold Voltage Degradation," in *Proceedings of the ACM International Symposium on Low Power Electronics and Design*, pp. 189–194, August 2007.

[32] T.-H. Kim, J. Liu, R. Persaud, and C. H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor fo r Measuring Frequency Degradation of Digital Circuits," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 874–880, April 2008.

[33] S. Mahapatra, P. B. Kumar, and M. A. Alam, "Investigation and Modeling of Interface and Bulk Trap Generation During Negative Bias

Temperature Instability of p-MOSFETs," *IEEE Transactions on Electron Devices*, vol. 51, pp. 1371–1379, September 2004.
[34] G. Chen, K. Y. Chuah, M. F. Li, D. S. H. Chan, C. H. Ang, J. Z. Cheng, Y. Jin, and D. L. Kwong, "Dynamic NBTI of PMOS Transistors and its Impact on Device Lifetime," in *Proceedings of the IEEE International Reliability Physics Symposium*, pp. 196 – 200, April 2003.
[35] A. T. Krishnan, C. Chancelor, S. Chakravarthi, P. E. Nicollian, V. Reddy, and A. Varghese, "Material Dependence of Hydrogen Diffusion: Implication for NBTI Degradation," in *IEEE International Electronic Devices Meeting*, pp. 688–691, December 2005.
[36] S. Rangan, N. Mielke, and E. C. C. Yeh, "Universal Recovery Behavior of Negative Bias Temperature Instability in PMOSFETs," in *IEEE International Electronic Devices Meeting*, pp. 14.3.1–14.3.4, December 2003.
[37] N. K. Jha and V. R. Rao, "A New Oxide-Trap Assisted NBTI Model," *IEEE Electron Device Letters*, vol. 26, pp. 687–689, September 2005.
[38] R. Fernandez, B. Kaczer, A. Nackaerts, S. Demuynck, R. Rodriguez, M. Nafria, and G. Groeseneken, "AC NBTI Studied in the 1 Hz – 2 GHz Range on Dedicated on-chip CMOS circuits," in *IEEE International Electronic Devices Meeting*, pp. 337–340, December 2006.
[39] M. Denais, V. Huard, C. Parthasarathy, G. Ribes, F. Perrier, N. Revil, and A. Bravaix, "Oxide Field Dependence of Interface Trap Generation During Negative Bias Temperature Instability in PMOS," in *IEEE International Integrated Reliability Workshop Final Report*, pp. 109–112, October 2004.
[40] A. E. Islam, E. N. Kumar, H. Das, S. Purawat, V. Maheta, H. Aono, E. Murakami, S. Mahapatra, and M. A. Alam, "Theory and Practice of On-The-Fly and Ultra-Fast-Vt Measurements for NBTI Degradation: Challenges and Opportunities," in *IEEE International Electronic Devices Meeting*, pp. 805–808, December 2007.
[41] H. Kufluoglu and M. A. Alam, "A Generalized Reaction-Diffusion Model with Explicit H-H_2 Dynamics for Negative-Bias Temperature Instability (NBTI) Degradation," *IEEE Transactions on Electron Devices*, vol. 5, pp. 1101–1107, May 2007.
[42] S. Mahapatra, P. B. Kumar, T. R. Dalei, D. Saha, and M. A. Alam, "Mechanism of Negative Bias Temperature Instability in CMOS Devices: Degradation, Recovery and Impact of Nitrogen," in *IEEE International Electronic Devices Meeting*, pp. 105–108, December 2004.
[43] B. Zhang. Private Communication.
[44] A. T. Krishnan. Private Communication.
[45] M. F. Li, G. Chen, C. Shen, X. P. Wang, H. Y. Yu, Y.-C. Yeo, and D. L. Kwong, "Dynamic bias temperature instability in ulthathin siO_2 and hfO_2 metal oxide semiconductor field effect transistors and its impact on device lifetime," *Japanese Journal of Applied Physics*.
[46] K. Kang, H. Kufluoglu, K. Roy, and M. A. Alam, "Impact of Negative Bias Temperature Instability in Nano-Scale SRAM Array: Modeling and Analysis," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, pp. 1770–1781, October 2007.

XI. Appendix A - First Stress Phase: Diffusion in Poly

In this section, we provide the details of the derivation for computing the interface traps during the first stress phase, on account of diffusion in polysilicon layer.

The rate of change in concentration of the hydrogen molecules inside poly is given by:

$$\frac{dN_{H_2}}{dt} = D_p \frac{d^2 N_{H_2}}{dx^2} \text{ for } x > d_{ox} \tag{41}$$

which is similar to the equation for oxide in (6). Assuming steady state diffusion, as in the case with the oxide (10) in Section IV, the above expression can also be approximated as:

$$D_p \frac{d^2 N^x_{H_2}(t)}{dx^2} = 0 \tag{42}$$

implying that the diffusion front in poly is also linear. The diffusion front assumes a quadrilateral shape inside the oxide, followed by a triangle in poly, with the tip of the diffusion front being at some $x_d(t) > d_{ox}$. Hence, we have:

$$\begin{aligned} \phi_{N_{H_2}} &= -D_p \frac{dN_{H_2}}{dx} \\ \frac{dN_{H_2}}{dx} &= \frac{N^{d_{ox}}_{H_2}}{x_d - d_{ox}} \end{aligned} \tag{43}$$

for $x > d_{ox}$ and

$$N^x_{H_2}, x > d_{ox} = N^{d_{ox}}_{H_2} - \frac{N^{d_{ox}}_{H_2}}{x_d - d_{ox}}(x - d_{ox}) \tag{44}$$

For large values of t, i.e., $t \gg t_1$, the shape of the plot can be approximated as a rectangle in oxide, followed by a triangle in poly, since the oxide thickness is of the order of a few angstroms, and $D_{ox} \gg D_p$. We verify this analytically by computing $N^{t_{ox}}_{H_2}$ as a function of $N^0_{H_2}$, as follows:

The number of interface traps is equal to the integral from (13), which is equal to the area under the curve in Fig. 3(e), as follows:

$$N_{IT}(t, t > t_1) = \left[d_{ox} \left(N^0_{H_2} + N^{d_{ox}}_{H_2} \right) + N^{d_{ox}}_{H_2} (x_d - d_{ox}) \right] \tag{45}$$

where $N^{d_{ox}}_{H_2}$ is the hydrogen molecular concentration at the oxide. Differentiating, with respect to time, and ignoring the $\frac{dN^{d_{ox}}_{H_2}}{dt}$ component, since $N^{d_{ox}}_{H_2}$ is a slowly decreasing function of time[8], we have:

$$\frac{dN_{IT}}{dt} \approx N^{d_{ox}}_{H_2} \frac{dx_d}{dt} \tag{46}$$

From (43) and (46), we have:

$$(x_d - d_{ox}) dx = D_P dt \tag{47}$$

Integrating, and using initial conditions, i.e., $x_d(t_1) = d_{ox}$, we have:

$$x_d = d_{ox} + \sqrt{2D_p(t - t_1)} \tag{48}$$

We now use the diffusion equation in (4) to compute the value of $N^{d_{ox}}_{H_2}$ in (45). Along the oxide-poly interface, the outgoing flux from the oxide is equal to the incoming flux into poly. Therefore, we have:

$$\phi^{d_{ox}}_{N_{H_2}} = D_{ox} \frac{dN_{H_2}}{dx} = D_p \frac{dN_{H_2}}{dx} \tag{49}$$

at $x = d_{ox}$. Since, N_{H_2} is a linear function of x, we have, at the interface:

$$D_{ox} \frac{\left(N^0_{H_2} - N^{d_{ox}}_{H_2} \right)}{d_{ox}} = D_p \frac{N^{d_{ox}}_{H_2}}{x_d - d_{ox}} \tag{50}$$

Substituting and simplifying, we have:

$$\begin{aligned} N^{d_{ox}}_{H_2} &= N^0_{H_2} \left[\frac{D_{ox}\sqrt{2D_p(t - t_1)}}{D_{ox}\sqrt{2D_p(t - t_1)} + D_p d_{ox}} \right] \\ &= N^0_{H_2} f(t) \text{ for brevity} \end{aligned} \tag{51}$$

It is easy to see that for $t \gg t_1$, the value of $N^{d_{ox}}_{H_2}$ almost becomes equal to $N^0_{H_2}$. The diffusion front is shown in

[8]The value of $N^{d_{ox}}_{H_2}$ and $N^0_{H_2}$ are determined by the rate of generation of interface traps at the surface (increases as $\sim t^{\frac{1}{6}}$), and the rate of diffusion of hydrogen molecules at the tip of the diffusion front (decreases as $\sim \sqrt{t}$), causing N_{H_2} to be a slowly decreasing function of time.

Fig. 3(f) for this case. The front almost becomes a rectangle in the oxide followed by a right angled triangle in poly. Using (51) in (45), we have:

$$N_{IT}(t, t_1 < t < \tau) = \Big[d_{ox} N^0_{H_2}(1 + f(t)) + N^0_{H_2}\sqrt{2D_p(t - t_1)} f(t)\Big] \quad (52)$$

Lumping the terms in (52), we have:

$$x_{\text{equiv}}(t, t_1 < t \leq \tau) = d_{ox}(1 + f(t)) + \sqrt{2D_p(t - t_1)} f(t) \quad (53)$$

where x_{equiv} represents the tip of an equivalent triangular front that has the same area. This step is performed such that the expression resembles the form in (21). Thus, we have the final expression:

$$\begin{aligned} N_{IT}(t, 0 < t \leq t_1) &= k_{IT}(2D_{ox}t)^{\frac{1}{6}} \\ N_{IT}(t, t_1 < t \leq \tau) &= k_{IT}\Big[d_{ox}(1 + f(t)) + \sqrt{2D_p(t - t_1)} f(t)\Big]^{\frac{1}{3}} \end{aligned} \quad (54)$$

XII. Appendix B - First Relaxation Phase: Recovery in Oxide

In this section, we describe the detailed derivation for the oxide recovery phase of NBTI action, during the first relaxation phase. During this stage, rapid annealing of interface traps occurs, and $N_{IT}(t)$ decreases significantly. It is vital to model this phase explicitly, to consider the impact of recovery during the time lag between the end of stress and the first time of recovery measurement[9].

Recovery in oxide consists of two sub-phases, namely, a reaction phase and a diffusion phase. During the reaction phase, we have from (3):

$$\frac{dN_{IT}}{dt} = -k_r N_{IT} N^0_H \quad (55)$$

where k_f is zero since there is no trap generation. The hydrogen concentration decreases exponentially during the beginning of the recovery phase, as shown in Fig. 24. A decrease in the concentration of interface traps occurs during this process. However, the reaction phase lasts only a few milliseconds, as seen from the simulation results. As the hydrogen concentration remains almost constant, diffusion becomes the dominant physical mechanism. During this diffusion phase, annealing of interface traps near the interface, followed by back-diffusion of existing hydrogen molecular species in the oxide occurs. For simplicity in modeling, we combine the reaction phase and the diffusion phase into a single stage of modeling as follows:

We model the rapid annealing of interface traps inside the oxide, which occurs from time τ to $\tau + t_2$, where t_2 is the time at which annealing proceeds into poly. Let us model the events at the interface as a superposition of two effects: "forward" diffusion, away from the interface, and "reverse" diffusion,

[9]For instance, undesired recovery during the time lag between end of stress and the first time of measurement which was not modeled previously, incorrectly led researchers to believe that the dynamics of N_{IT} generation followed a $t^{\frac{1}{4}}$ dependence, instead of the actual $t^{\frac{1}{6}}$ dependence [10].

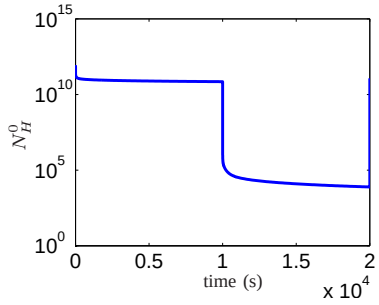


Fig. 24. Hydrogen concentration at the oxide-substrate interface during the first stress and recovery phases, showing the rapid decrease in N^0_H at the beginning of the recovery phase.

toward the interface; the latter anneals the interface traps, as explained in Section V-D. During this condition, the diffusion of existing species continues as $x(t + \tau) \propto \sqrt{2D_p(t + \tau)}$ inside poly, while the peak of the diffusion front decreases from $N^0_{H_2}$ to $N^{\Delta}_{H_2}$, as shown in Fig. 25(c), for some $\Delta \leq d_{ox}$.

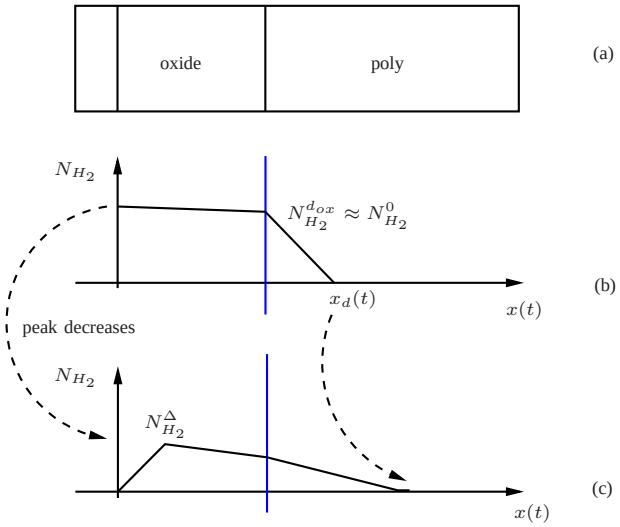


Fig. 25. Diffusion front for the first recovery phase: (a) shows the cross section of the PMOS transistor, (b) shows the front at time τ, i.e., at the end of the first stress phase, while (c) shows the front at time $\tau + t$, into the first recovery phase.

We may approximate the hydrogen concentration in the oxide as being a triangle plus a quadrilateral: at time $(\tau + t)$, it goes from 0 at $x = 0$, to $N^{\Delta}_{H_2}(\tau + t)$ at $x = \Delta$, for some $\Delta \leq d_{ox}$. The hydrogen molecular concentration follows a right angled triangle profile in poly, since there is no effect of annealing here yet, with the concentration being $N^{d_{ox}}_{H_2} \approx N^{\Delta}_{H_2}(\tau + t)$ at the oxide-poly interface, and decreasing to 0 again at $x_d(\tau + t)$. During this phase, the rate of decrease of interface traps can be assumed to be low, and is hence approximated as 0. Using the same notation as [4], page 3, we have:

$$\frac{dN_{IT}}{dt} \approx 0 = -k_r(N^0_{IT} - N^*_{IT})(N^0_H - N^*_H) \quad (56)$$

Since the residual number of interface traps, $(N^0_{IT} - N^*_{IT})$, is significantly larger than zero, it must mean that the residual

hydrogen concentration at the interface, $(N_H^0 - N_H^*)$ must be near-zero. Denoting the number of annealed traps as $N_{IT}^*(\tau + t)$, we can express the net number of interface traps during the relaxation phase as the original number of traps, minus the number of annealed traps:

$$N_{IT}(\tau + t) = N_{IT}(\tau) - N_{IT}^*(\tau + t) \tag{57}$$

The number of interface traps annealed due to backward diffusion [4] can be expressed as:

$$N_{IT}^*(\tau + t) = N_{H_2}^{\Delta}\sqrt{\xi_1 \times 2D_{ox}t} \tag{58}$$

Intuitively, this can be considered to be equivalent to a triangle whose height is given by $N_{H_2}^{\Delta}$, and the backward diffusion front beginning at time τ is given from [4] as:

$$x^*(t) = \sqrt{2\xi_1 D_{ox}t} \tag{59}$$

ξ_1 is a parameter that captures the effect of two-sided diffusion, and its original value is of the order of ≈ 0.58 [4]. However, in order to account for the exponential decrease in the interface trap concentration during the reaction phase of the first recovery phase, using a single analytical model, ξ_1 is set to a large number, and its exact value is determined through curve fitting.

Based on the argument in Section IV, the total number of interface traps is given by the area enclosed under the quadrilateral plus the triangles in Fig. 25(c) as:

$$N_{IT}(t + \tau) \approx N_{H_2}^{\Delta}\left(2d_{ox} - \Delta + \sqrt{2D_p(t + \tau)}\right) \tag{60}$$

where Δ, i.e., the location of the peak concentration of hydrogen molecules during recovery (follows the dynamics of the diffusion front for stress phase, and hence from (17)) increases with time as:

$$\Delta = \sqrt{2D_{ox}t} \tag{61}$$

The tip of the diffusion front $x_d(t)$, computed from (53), is approximately at:

$$x_d(t) = d_{ox} + \sqrt{2D_p(t + \tau)} \tag{62}$$

Solving for $N_{H_2}^{\Delta}$ in (60), we have:

$$N_{H_2}^* = \frac{N_{IT}(t + \tau)}{2d_{ox} - \sqrt{2D_{ox}t} + \sqrt{2D_p(t + \tau)}} \tag{63}$$

Since, the number of interface traps is given by the difference between the number of traps at τ, and the number of traps annealed, we have:

$$N_{IT}(t + \tau) = N_{IT}(\tau) - N_{IT}^*(t + \tau) \tag{64}$$

Substituting for $N_{IT}^*(t + \tau)$, and simplifying, we have:

$$N_{IT}(t + \tau) = N_{IT}(\tau) - N_{IT}(t + \tau)g(\xi_1, t) \tag{65}$$

where for brevity, $g(\xi_1, t) = \left[\frac{\sqrt{2\xi_1 D_{ox}t}}{2d_{ox} - \sqrt{2D_{ox}t} + \sqrt{2D_p(t + \tau)}}\right]$ (66)

Simplifying, we have:

$$N_{IT}(t + \tau, 0 < t \leq t_2) = \frac{N_{IT}(\tau)}{1 + g(\xi_1, t)} \tag{67}$$

This process continues until time t_2, when the back-diffusion front has reached the oxide-poly interface.

XIII. Appendix C - Second Stress and Recovery Phases

A. Second Stress Phase

For the second stress phase, we use boundary conditions at time 2τ, to determine the tip of the effective diffusion front. We solve for $x_{\text{eff}}(2\tau)$ by assuming an equivalent front which has diffused from time 0 to 2τ, and has the same interface trap concentration as $N_{IT}(2\tau)$:

$$N_{IT}(2\tau) = k_{IT}x_{\text{eff}}(2\tau)^{\frac{1}{3}} \tag{68}$$

The integral for $x_d(t)$ from (16) is now solved with the limits modified, to obtain:

$$x_d(t) = \sqrt{x_{\text{eff}}(2\tau)^2 + 2D_{ox}t} \tag{69}$$

instead of (17). This equation can be used in (21) to estimate the rapid increase in interface traps due to diffusion inside the oxide for the second stress phase as follows:

$$N_{IT}(t + 2\tau, 0 < t \leq t_1) = k_{IT}[2D_{ox}t + x_{\text{eff}}(2\tau)^2]^{\frac{1}{6}} \tag{70}$$

This process continues until time t_1, beyond which diffusion occurs in poly. Diffusion inside poly can be computed using the method outlined in the previous section, and the number of interface traps is approximated as:

$$\begin{aligned} N_{IT}(t + 2\tau, t_1 < t \leq \tau) &= k_{IT}\Big[\sqrt{((1 + f(t))d_{ox})^2 + x_{\text{eff}}(2\tau)^2} \\ &\quad + f(t)\sqrt{2D_p(t - t_1)}\Big]^{\frac{1}{3}} \end{aligned} \tag{71}$$

for time $2\tau + t_1$ to 3τ. For large values of t, $f(t) \approx 1$. Hence, we can approximate the above expression as:

$$\begin{aligned} N_{IT}(t + 2\tau, t_1 < t \leq \tau) &= k_{IT}\Big[\sqrt{2{d_{ox}}^2 + x_{\text{eff}}(2\tau)^2} + \\ &\quad \sqrt{2D_p(t - t_1)}\Big]^{\frac{1}{3}} \end{aligned} \tag{72}$$

As a sanity check, setting $x_{\text{eff}}(2\tau)$ in (70), we obtain:

$$N_{IT}(t) = k_{IT}[2D_{ox}t]^{\frac{1}{6}} \tag{73}$$

which is the equation for the interface trap generation inside the oxide for the first stress phase, from (19). Similarly, setting $t = t_1$ and therefore $f(t) = 0$ in (71), we have:

$$\begin{aligned} N_{IT}(t_1 + 2\tau) &= k_{IT}[d_{ox}^2 + x_{\text{eff}}(2\tau)^2]^{\frac{1}{6}} \\ &= k_{IT}[2D_{ox}t + x_{\text{eff}}(2\tau)^2]^{\frac{1}{6}} \end{aligned} \tag{74}$$

which is the equation for interface trap generation inside the oxide during the stress phase, from (70).

B. Second Recovery Phase

Recovery modeling for the second relaxation phase is similar to that in the first relaxation phase. We assume that by this time, the diffusion front has recovered to its original shape of almost a rectangle in the oxide, followed by a triangle in poly (Fig. 3(f)). The above assumption has been verified through numerical simulations to be valid for large values of $\tau >$1s. Accordingly, the front for the second recovery phase is similar to that in (62) and (53), and is given by:

$$x_d(3\tau) \approx d_{ox} + \sqrt{2D_p(3\tau)} \tag{75}$$

During the second recovery phase, the tip of the existing front is away from the interface, and hence grows as:

$$x_d(t+3\tau) = d_{ox} + \sqrt{2D_p(3\tau+t)} \tag{76}$$

However, rapid annealing occurs near the interface, causing a decrease in the number of interface traps. Accordingly, we have the equation:

$$N_{IT}(t+3\tau, 0 < t \leq t_2) = \frac{N_{IT}(3\tau)}{1+g'(t)} \tag{77}$$

$$\text{where } g'(t) = \left[\frac{\xi_1\sqrt{2D_{ox}t}}{2d_{ox} - \sqrt{2D_{ox}t} + \sqrt{2D_p(t+3\tau)}}\right] \tag{78}$$

for time 3τ to $3\tau + t_2$, which is similar to the expression for the first recovery phase, in the oxide, given by (67). Modeling for the slow recovery phase is similar to that derived in the previous section, and the final expression is given by:

$$N_{IT}(t+3\tau, t_2 < t \leq \tau) = N_{IT}(3\tau+t_2)\left[1 - \sqrt{\frac{\xi_2(t-t_2)}{t+3\tau}}\right] \tag{79}$$

for time $3\tau + t_1$ to 4τ.