On Quasi-Saturation of Negative Bias Temperature Degradation

M. A. Alam and H. Kufluoglu
School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47906

Abstract

Five different models have been proposed in recent years to interpret the quasi-saturation of interface trap generation in PMOS Transistors due to Negative Bias Temperature Instability (NBTI). We use both analytical and numerical methods to capture the essence of these models and show that these models predict very different temperature, thickness, and voltage dependence regarding the onset of quasi-saturation. These predictions should be compared with experiments to uniquely identify the origin of quasi-saturation of NBTI characteristics. The physical mechanism of quasi-saturation determines the extrapolated lifetime of PMOS transistors and has important implications for reliability of Silicon Integrated Circuits.

Keywords: MOSFET, NBTI, Reliability Theory and Modeling, Reaction-diffusion model

1. Introduction and Background:

Since late 1990s, Negative Bias Temperature Instability (NBTI) of PMOS transistors has become the leading reliability concern for Silicon ICs [1-2]. The prevailing view of NBTI degradation (see Fig. 1) is that when PMOS is stressed in inversion by applying a negative gate voltage (V_{GS}), Si-H bonds at the Si/SiO_2 interface capture holes from the inversion layer. Every hole removes one of the covalent electrons holding Si and H together. The weakened bond is easily broken at high temperatures. These broken bonds act like interface traps: they increase the threshold voltage and reduce transistor drive current. If the threshold voltage shift due to NBTI degradation becomes larger than the design margin of an IC, the timing among various sub-circuits will be incorrect. And this would lead to functional failure of the IC.

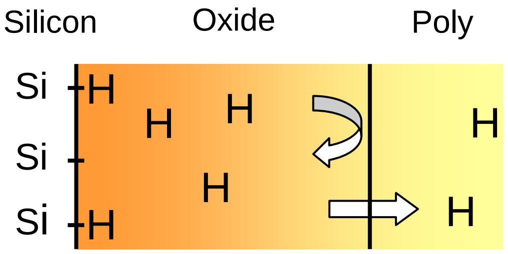

Fig. 1: Schematic diagram of NBTI degradation due to breaking of Si-H bonds at the Silicon/Oxide interface and subsequent release and diffusion of *H* away from the interface. The interface between poly-silicon and oxide is also shown. If the diffusivities of *H* in poly and in oxide are different, the trap generation rate may be modified.

Indeed, with DC stress, the shifts in threshold voltage due to NBTI in sub-130nm technologies are generally so large that without some mitigating circumstances, many ICs would have certainly failed to operate at standard voltage and temperature. In practice, however, two features of the NBTI degradation prolong IC lifetime: first, at longer times, the trap generation rate is reduced (quasi-saturation effect) [3-8] and second, at high frequencies, AC NBTI degradation is much smaller than DC values. We have discussed the AC NBTI degradation before in other publications [9]. In this paper, we focus on the quasi-saturation characteristics of NBTI degradation.

The NBTI literature until late 1990s generally characterized the interface trap (N_{IT}) generation as a function of time (t) as a power-law (i.e. $N_{IT} \sim t^n$) with n~0.25-0.30. This value of n was assumed robust, that is, independent of stress time. This time-independent exponent had been the unique signature of NBTI degradation and traditional theoretical models of NBTI strived to interpret the constancy of n satisfactorily.

However, it became increasingly clear in late 1990s, as the semiconductor industry began to struggle with NBTI issue, that the projected IC lifetime based on n~0.25-0.30 will be unacceptable. This led to an intense reexamination of the NBTI time-exponent. The essence of many measurements of n done since late 1990s is the following: n~0.25-0.30 at the early stage of degradation, however at long stress times, n gradually decreased to ~0.12-0.15. This is good news for semiconductor industry because lower exponents translate to longer extrapolated lifetime for CMOS circuits. However, this time-dependent change in n (in other words, quasi-saturation of trap generation) poses a challenge to old theories of NBTI which focused on the interpretation of *time-independent* exponents. The old NBTI theories had to be generalized to explain the new phenomena and establish appropriate voltage and temperature scaling laws.

Five different mechanisms for NBTI quasi-saturation have been proposed:

a) Reflection of H at the poly/oxide interface [3-5],
b) Breaking of all Si-H bonds at the Si/SiO$_2$ interface [6],
c) Variation of bonding strength of the precursors[10,11],
d) Transition from atomic to molecular hydrogen [8], and
e) Artifact of finite measurement delay [8].

All these models are based on various refinements of basic Reaction-Diffusion (R-D) model. The basic model has been extensively validated in the literature [3-5,9,12-14]. We discuss the elementary considerations of the R-D model in Sec. 2, and then use both analytical formulations as well as numerical solutions of the R-D model to explore the proposed mechanisms of quasi-saturation (Sec. 3). The voltage, thickness, and temperature dependence of NBTI degradation predicted by these models are different. In Sec. 4, we highlight the predicted differences among the models and propose possible methods for explicit experimental confirmation.

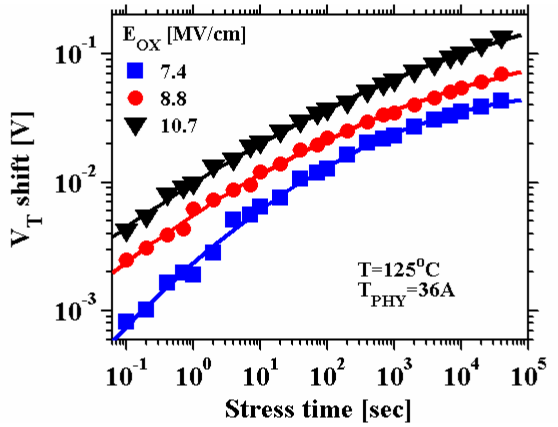

Fig. 2: Typical measurement of NBTI shows that the time exponent n, given by the slope of the log(N_{IT})–log(t) curve, is time dependent that can be interpreted as quasi-saturation of interface trap generation [18].

2. Reaction-Diffusion Model:

The R-D model has been extensively discussed in the literature [3-9, 12-13]. Here we provide an elementary discussion of the R-D model which is defined by a reaction and a drift-diffusion equation,

$$dN_{IT}/dt = k_F(N_0 - N_{IT}) - k_R N_H N_{IT} \ , \qquad (1a)$$

$$dN_{IT}/dt = D_{H2}\left[dN_{H2}/dx\right] + \mu_{H2} N_{H2} E_{ox} + {}^{\delta}\!/_{2}\,(dN_{H2}/dt)\,, \qquad (1b)$$

where N_{IT} is the number of interface-traps (broken Si-H bonds) at a given time t, k_F and k_R are forward dissociation rate and reverse annealing rate of Si-H bonds, respectively, N_0 is the total number of Si-H bonds at the beginning of the stress, N_H is the atomic Hydrogen (H) concentration at the interface, and N_{H2} is the molecular H_2 concentration in the bulk of the oxide, D_{H2} and μ_{H2} are diffusion coefficient and mobility of H_2, and the final term in (1b) accounts for time dependent H_2 build-up within the interface width δ (~2-3 A), which is negligible compared to the other terms. At the interface, the atomic and molecular concentrations are related by law of mass action, i.e. $k_1 N_H^2 = k_2 N_{H2}$.

In the standard R-D approximation, one assumes that in (1a), the trap generation rate, dN_{IT}/dt, is much slower than the dissociation and annealing rates, and $N_{IT} << N_0$, so that $k_F N_0 \sim k_R N_{IT} N_H$. Moreover, R-D model assumes neutral H_2 diffusion, therefore the second term on the right hand side of (1b) is zero and the third term is negligible. Since H_2 diffusion front moves a distance of $x \sim (D_{H2}{\cdot}t)^{0.5}$ in time t, so $dN_{IT}/dt \sim D_{H2}N_{H2} \,/\, (D_{H2})^{0.5} \sim D_{H2}N_H^2 / (D_{H2}{\cdot}t)^{0.5}$ by Eq. (1b). Substituting $k_F N_0 \sim k_R N_{IT} N_H$ in this relationship, we find,

$$\frac{N_{IT}}{N_0} \sim a\left(\frac{t}{\tau}\right)^n ; \qquad \frac{1}{\tau} \equiv \left(\frac{k_F}{k_R N_0}\right)^{1/2n} D_{H2} \qquad (2)$$

where a is a constant and n~1/6. If diffusion of hydrogen is dispersive, i.e. $D_{H2} \sim D_0 t^{-m}$ [15-16], the power-exponent of time t in (2) would change from n to $(1-m)n$ [14].

Similarly, if the diffusing species is proton (H^+), then drift term dominates diffusion in Eq. 1(b) so that $dN_{IT} / dt \sim \mu_{H2} N_{H2} E_{OX}$. Along with $k_F N_0 \sim k_R N_{IT} N_H$, we find that N_{IT} is given by the same expressions as (2), except n=1/6 is replaced by n=1/3 and D_{H2} is replaced by ($\mu_{H2}E_{ox}$). If proton drift is dispersive, i.e. $\mu_{H2} \sim \mu_0 t^{-m}$, then n is replaced by $n(1-m)$, as before [6,7,15,16].

Therefore, according to the R-D model of NBTI, the power-law in (2) arises from the interplay of generation, annealing, and (either normal or dispersive) diffusion of H_2 at the Si/SiO_2 interface. In R-D model, exponent n is intimately related to the geometry of diffusion, allowing interpretation of HCI and NBTI time exponents as a consequence of 1D and 2D diffusion, respectively [17]. And the finite value of annealing coefficients, k_R, allows interpretation of relaxation experiments [3-5]. Note that the exponent n in (2) is robust and time-independent, that is, given the assumptions, this is an exact solution. One would need to relax some of the assumptions made in deriving (2) to interpret the time-dependent exponent n that characterizes the quasi-saturation of NBTI characteristics.

3.1 Saturation Due to Poly-SiO2 interface Reflection (Reflection Model):

In the discussion above, we assumed that H_2 diffuses in oxide alone, but this assumption is unrealistic for sub-5 nm oxides. If the H_2 diffusion in poly-silicon (which tops the oxide) is slower than that in the oxide, then there will be a build-up of H_2 at the poly-oxide interface (see Fig. 1). This would increase the reverse reaction component in Eq. 1b (second term on the right) and reduce overall trap generation. As seen is Fig. 3, numerical solution of (1), that allows two different diffusion coefficients for oxides and poly, results in quasi-saturation of N_{IT} vs. time characteristics. However, as shown in Fig. 3, this saturation is *soft*, because given adequate time, the N_{IT} vs. time characteristic would regain the original exponent (except D_{H2} in (2) will have to be replaced by ($D_{H2}^{(ox)} D_{H2}^{(poly)} / (D_{H2}^{(ox)} + D_{H2}^{(poly)})$) [8].

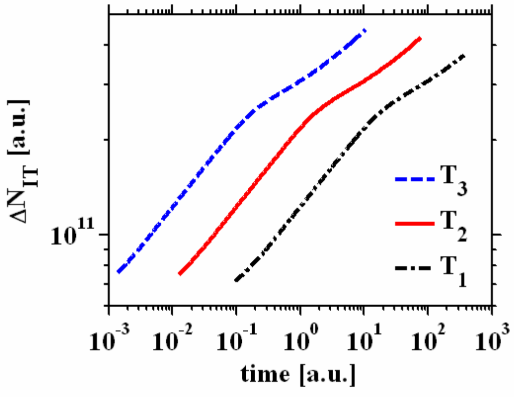

Fig. 3: Temperature(T)-dependent reduction in N_{IT} generation due to reflection at the poly-oxide interface ($D_{H2}(T_3) = 10\ D_{H2}(T_2) = 100\ D_{H2}(T_1)$).

3.2 Saturation Due to Depletion of Precursors (S-E Model):

If the poly and the oxide diffusion coefficients are equal or if diffusion in poly is faster than that in oxide $D_{H2}^{(ox)} \leq D_{H2}^{(poly)}$, there would be no build-up of H_2 at the oxide/poly interface, and reflection at that interface would not explain NBTI saturation. Assuming this is the case, the second interpretation of NBTI has been proposed which challenges the assumption that N_{IT} remains negligible compared to precursor density N_0 throughout the measurement window. The distinction of *soft* saturation due to reflection at poly/oxide interface discussed above with *hard* saturation due to depletion of all Si-H bonds as $N_{IT} \rightarrow N_0$ allows one to interpret the Stretched-Exponential (S-E) model [4,14] as an approximation to R-D model. Near the hard saturation limit, $dN_{IT}/dt \rightarrow 0$ (not merely negligible compared to other two terms in Eq. 1a), but $N_{IT} \sim N_0$, therefore, $k_F(N_0 - N_{IT}) \sim k_R N_{IT} N_H$ from (1a). Inserting this in $dN_{IT} / dt \sim D_{H2} N_{H2} / (D_{H2} t)^{0.5}$ (Eq. (1b) for neutral H diffusion), and integrating we find

$$-\frac{N_{IT}}{N_0} - \ln\left(1 - \frac{N_{IT}}{N_0 - N_{IT}(0)}\right) = b\left(\frac{t}{\tau}\right)^{2n} ; \qquad \frac{1}{\tau} \equiv \left(\frac{k_F}{K_R N_0}\right)^{1/2n} D_{H2} \qquad (3a)$$

where b is a constant (~1/2n) and n=1/6. Eq. (3a) is readily approximated as

$$\frac{N_{IT}}{N_0} \approx \left(1 - \exp\left[-\left\{\frac{t}{\tau}\right\}^n\right]\right) \qquad \text{(3b)}$$

by Taylor series expansion (correct to the 2nd order, i.e. error ~ $0.2 \cdot (N_{IT} / N_0)^2$, and absolute error bounded by a factor of 2 as N_{IT}/N_0 approaches 1). Again, if the transport is dispersive, i.e. $D_{H2} \sim D_0 t^{-m}$, the $n(1\text{-}m)$ replaces time-exponent n. In addition, if one assumes H^+ drift rather neutral H_2 diffusion, Eqs. (3a) and (3b) remain unchanged except n=1/6 is replaced by n=1/3 and D_{H2} replaced by $(\mu_{H2} E_{ox})$. Figure 4 shows that this model would also lead to quasi-saturation of the NBTI characteristics. Since the saturation point is determined by certain density of traps, it will be reached at different times at different temperatures.

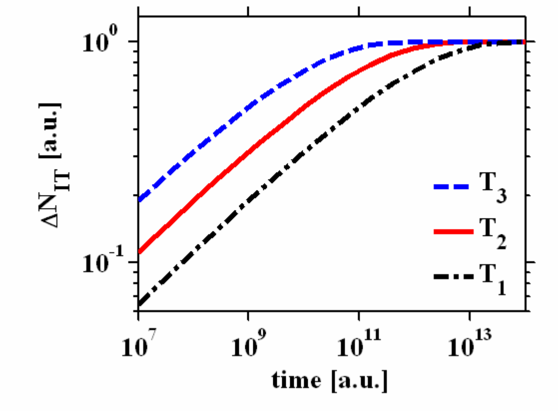

Fig. 4: Reduction in N_{IT} generation once Si-H precursors (N_0 in Eq. 1) is depleted. The saturation point increases as temperature is reduced.

3.3 Saturation Due to Dispersion in Bond Energies (B-D Model):

The third model assumes that $D_{H2}^{(ox)} \leq D_{H2}^{(poly)}$ and $N_0 >> N_{IT}$ within the experimental window, so that neither of the above models can explain quasi-saturation. In the bond dispersion model (B-D) model, quasi-saturation arises from the well known fact that Si-H bonds have a distribution of bonding energies(~0.1eV). The weaker bonds are rapidly broken as soon as the stress is applied, so that initial N_{IT} generation is high. The remaining bonds, on the average, are stronger so their dissociation rate continues to fall with time, leading to quasi-saturation of interface trap generation. The B-D models assumes H_2 diffuses away from the interface faster than N_{IT} generation ($D_H/T_{ox} >> k_F$) making (1b) unnecessary. The model also assumes that interface annealing is negligible (k_R~0) making second term of (1a) irrelevant. Since B-D model assumes that the Si-H bonding or activation energy, E_A, is not unique, so that Equation (1a) now needs to be rewritten as $dN_{IT}/dt \simeq \int_{E_A} k_F(E_A) D_0(E_A) dE_A$ with $\int_{E_A} D_0(E_A) dE_A \equiv N_0$ and $N_{IT} << N_0$, D_0 being the bond-energy distribution. Assuming that $D_0(E_A) = \dfrac{1}{\sigma}\dfrac{\exp(-\Delta\varepsilon)}{[1+\exp(-\Delta\varepsilon)]^2}$ with $\Delta\varepsilon \equiv (E_A - \langle E_A\rangle)/\sigma$ and $k_F = k_0 \cdot \exp(-E_A / k_B T)$ ($\langle E_A\rangle$ is average activation energy, σ is the standard-deviation of bonding energies, and T is temperature), trap generation is given by [11]

$$\frac{N_{IT}}{N_0} = 1 - \frac{1}{1+(t/\tau)^n}; \quad n = \left(\frac{k_B T}{\sigma}\right)^p. \qquad (4)$$

The exponent p depends on the range of fit to the experimental data. The numerical results for different temperatures are shown in Fig. 5.

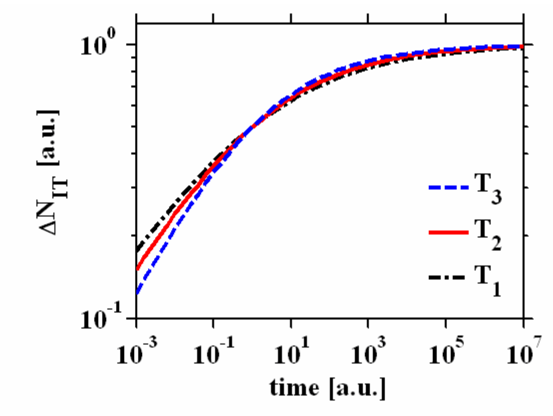

Fig. 5: Reduction in interface trap generation dispersion in Si-H bonding energies (σ=0.1eV).

While B-D model provides one possible interpretation of NBTI saturation, since annealing term in (1b) is neglected, the B-D model (in the current form) can not explain the relaxation experiments. However, like S-E and Reflection model, the T-dependence of time-exponent n arises from dispersion (bond dispersion for B-D model, diffusion or drift-dispersion in S-E and R-D models). In the absence of bond-dispersion, the other models would still predict fractional time-exponents, but the B-D model would not. Finally, since B-D model does not consider H diffusion or H^+ drift explicitly, therefore unlike R-D model, exponent n does not depend on geometry and only hard saturation of N_{IT} is possible (i.e. n approaches zero for long stresses).

3.4 Saturation Due to H to H_2 Transition (H-H_2 Model):

The H-H_2 model proposes that the quasi-saturation arises from the gradual transition from H to H_2 (rather than the instantaneous transition implied by the use of law of mass-action in Sec. 2). Since H diffuses with exponent n=1/4 and H_2 diffuses with exponent n=1/6, the continual reaction from H to H_2 is reflected in time-dependent transition from high to low trap generation exponent (in other words quasi-saturation). This model interesting because it does predict that the slope will saturate towards n=1/6, which is apparently consistent with experimental data. The details of this somewhat complicated model are will be discussed later in a separate publication. However, a representative curve from this model is shown in Fig. 6. The essential aspect of these plots are that since diffusion coefficients of both H and H_2 are temperature dependent, therefore the decay length of H into H_2 depends on temperature. This temperature dependence is the key to validate if H to H_2 transition is the determining factor of quasi-saturation characteristics of NBTI.

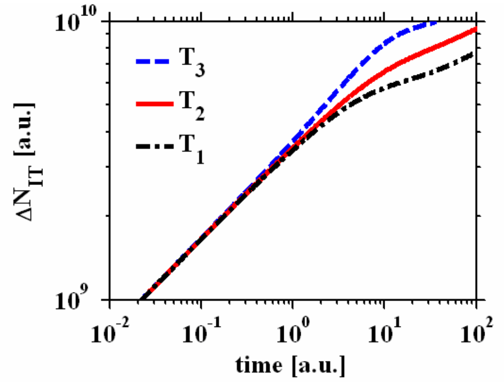

Fig. 6: Since the transition from H (which dominates early N_{IT} generation) to H_2 (dominates trap generation at later time) is temperature-dependent, the times for quasi-saturation also depends on temperature.

3.5 Saturation as an Artifact of Measurement Delay (Delay Model):

Finally, it has been argued that quasi-saturation in N_{IT} generation does not reflect any actual reduction of trap generation rate (as has been discussed above), but rather this is a consequence of limitations of N_{IT} measurement techniques. Specifically, in order to measure N_{IT} by standard techniques like charge-pumping or threshold voltage shift, the NBTI stress is temporarily interrupted, the N_{IT} measurement is done, and stress is reapplied. During this measurement interval (after the stress has been turned off and before the measurement has began, i.e. measurement delay) a significant fraction of N_{IT} is annealed because, although once the stress is removed, the field dependent k_F term in Eq. 1 becomes zero, so that no new N_{IT} generation is possible, however, the field *independent* reverse reaction term (k_R) remains nonzero and drives the reverse annealing of H and N_{IT}, with a net reduction in N_{IT}. The subsequent measure of N_{IT} reflects this annealing-induced reduction in actual trap density. The exact loss of N_{IT} due to this measurement issue depends on the relative magnitude of stress interval and measurement delay. Generally, since stress intervals are set to increase geometrically with time while the measurement delay remains constant, N_{IT}-loss is initially significant, but it continues to reduce with time so that eventually delay-interrupted N_{IT} approaches the uninterrupted N_{IT} values.

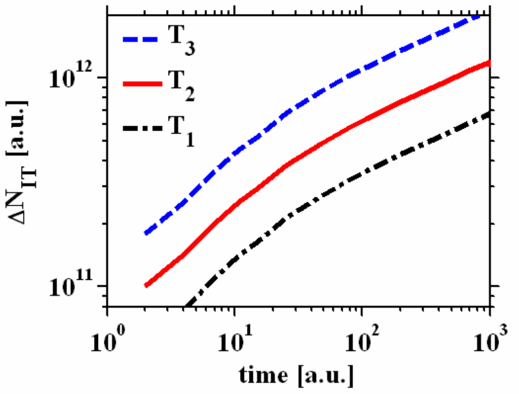

Fig. 7: Numerical simulation of measurement delay also gives rise to quasi-saturation behavior in N_{IT} vs. time characteristics (T_3> T_2> T_1 and delay=1sec).

4. Consequences and Conclusions:

Fig. 8 and Table 1 summarize the results of this paper. Fig. 8 plots the temperature dependence of saturation point (left plot) or equivalently, the time to quasi-saturation (right plot). Considering the right plot of Fig. 8, we see that strong temperature dependence of the quasi-saturation point would support the Reflection model or the S-E (Hard Saturation) model. On the other hand, if the quasi-saturation is temperature-insensitive, one would consider the Delay model or H-H_2 model (assuming k_1 and k_2 have weak temperature-dependences) as possible explanation of quasi-saturation. In addition to temperature as being of measurement variable, one can also study the predictions of the models for thickness and voltage dependence. As shown in Table 1 that the five models predict very different thickness and voltage characteristics of the quasi-saturation transition points.

Given that the delay model explores measurement artifacts, this has to be an important component of any comprehensive model for NBTI saturation. Once corrected for measurement delay, the data will then have to be analyzed for consistency by the remaining four models. Among these four models, B-D model, *in the form described above*, is an unlikely candidate because the model is not consistent with annealing data. Of the remaining three models, in analysis of temperature, thickness and stress dependence seems to support the Reflection model over the S-E model, although the issue has to be revisited once the measurement delay has been properly accounted for. Both the *H-H_2* model and the B-D model highlight important points (that H-H_2 transformation is not instantaneous and that bond-strength is not unique), but their contributions may not be the most important factors in determining NBTI quasi-saturation behavior.

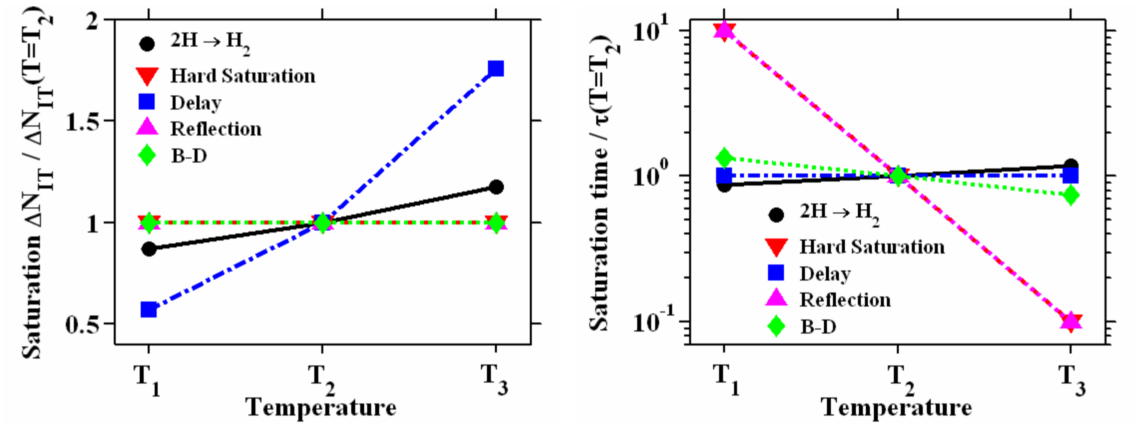

Fig. 8: Summary of results from previous figures. The interface trap density at which quasi-saturation point is reached (left), or the time (τ) at which the quasi-saturation point is reached (right) depends on the model. Each of these models predicts very different temperature dependence which can be compared against experiments.

MODEL/ DEPENDENCE	THICKNESS	TEMPERATURE	VOLTAGE	ASYMPTOTIC EXPONENT (N)
Reflection	Quadratic	Strong	None	Same as initial n
S-E Model	None	Strong	Strong	0
B-D Model	None	Weak	Strong	0
H-H_2 Model	None	Weak	None	~1/6
Delay Model	None	Weak	Weak	~1/6

Table 1: Summary of thickness, voltage, and temperature dependence (see Fig. 8, right) of quasi-saturation time-point that should be compared with experimental data to establish the mechanism of the reduction in trap generation as a function of time.

Acknowledgement:

This work is supported by NSF and SRC (Grant Number: 1238.001). We also gratefully acknowledge discussion with A. Krishnan from Texas Instruments, S. Mahapatra from IIT Bombay, and S. Zafar from IBM.

References:

[1] C. E. Blat, E.H. Nicollian, and E. H. Poindexter, “Mechanism of negative bias temperature instability,” J. Appl. Phys. 69, p. 1712, 1991.

[2] D. Schroder and J. Babcock, “Negative Bias Temperature Instability: Road to Cross in Deep Submicron Silicon Semiconductor Manufacturing”, J. of Applied Physics, 94, p. 1, 2003.

[3] M. Alam and S. Mahapatra, “A Comprehensive Model of PMOS NBTI Degradation,” Special issue of Microelectronics Reliability, 45, p. 71, 2005. This issue contains a number of excellent articles on various aspects of NBTI degradation.

[4] S. Chakravarthi *et al*., “A Comprehensive Framework for Predictive Modeling of Negative Bias Temperature Instability,” Proc. of Int. Rel. Physics Symposium, p. 273, 2004.

[5] S. Mahapatra, P. Bharath Kumar and M. A. Alam, “Investigation and Modeling of Interface and Bulk Trap Generation during Negative Bias Temperature Instability of p-MOSFETs,” IEEE Trans. Electron Devices, v.51, p.1371, 2004.

[6] S. Zafar, B. Lee, J. Stathis, A. Callegari, and T. Ning, “A Model for Negative Bias Temperature Instability in Oxide and High-k PFETs,” Proc. of VLSI Symp. p. 208, 2004.

[7] B. Kaczer *et al*., “Kinetics of Negative Bias Temperature Instability,” Proc. of Int. Rel. Physics Symposium, p. 381, 2005.

[8] H. Kufluoglu and M. Alam, unpublished results.

[9] M. Alam, “A Critical Examination of the Mechanics of Dynamic NBTI in PMOS,” Proc. of Int. Elec. Dev. Meeting, p. 346, 2003.

[10] O. Penzin, A. Haggag, W. McMahon, E. Lyumkis, and K. Hess, “MOSFET Degradation Kinetics and Its Simulation,” *IEEE Trans. on Electron Devices*, 50(6), p. 1445, 2003.

[11] K. Hess et al., “Simulation of Si-SiO2 defect generation in CMOS chips: from atomistic structure to chip failure rates,” *Proc. of IEDM*, p. 10, 2000.

[12] K. O. Jeppson, C. M. Svensson, “Negative bias stress of MOS devices at high electric fields and degradation of MOS devices,” *J. Appl. Phys.*, vol.48, p. 2004, 1977.

[13] S. Ogawa and N. Shiono, “Generalized diffusion-reaction model for the low-field charge build up instability at the Si-SiO_2 interface,” *Phys. Review B*, vol.51, no.7, p. 4218, 1995.

[14] B. Kaczer, R. Degraeve, V. Arkhipov, N. Collaert, G. Goreseneken, M. Goodwin, “The influence of recovery and temperature on the NBTI power-law exponents,” Program Book. of SISC, p. 108, 2004.

[15] J. Kakalios, R. A. Street, and W.B. Jackson, “Stretched-exponential relaxation arising from dispersive diffusion of hydrogen in amorphous silicon,” *Phys. Rev. Lett*., v.59, p. 1037, 1987.

[16] I. A. Shkrob and A.D. Trifunac, “Time-resolved EPR of spin-polarized mobile H atoms in amorphous silica: The involvement of small polarons,” *Phys. Rev. B*, v.54, p. 15073, 1996.

[17] H. Kufluoglu and M. Alam, “A Geometrical Unification of the theories of NBTI and HCI time exponents and its implications for planar and surround gate devices,” Proc. of Int. Elec. Dev. Meeting, p. 113, 2004.

[18] S. Mahapatra, unpublished results.