A study of the interface-trap activation kinetics in the Negative Bias Temperature Instability

Filippo Alagi

STMicroelectronics, 20010 Cornaredo, MI - Italy, Via Tolomeo, 1
E-mail address: filippo.alagi@st.com

Abstract: Experiments on a silicon p-channel MOS FET with heated gate electrode have provided evidence that the increase of the interface-trap density observed upon homogeneous negative bias temperature stress can be modelled as the unbalance of two reaction-limited activation and deactivation processes. The variation of the measured trap density with the stress time shows saturating trends, with bias-dependent saturation level. This feature is explained as the dynamic balance of a field-dependent forward reaction and a field-independent backward reaction. The rate constants of both reactions appear to be Arrhenius-like, with distributed activation energy. The activation energy of the forward reaction, which is identified as the depassivation of native Si/SiO₂ (near-) interface defects, turns out to be a nearly square root function of the electric field in the oxide. By analogy with the early Frenkel’s theory of electron transport in dielectric materials we discuss whether this field dependence can be considered as the first evidence that the defect depassivation would proceed through a stage where a positively charged hydrogen specie escapes from negatively charged defect.

Keywords: NBTI; heated gate; dispersive kinetics; dynamic balance
1. Introduction

One interpretation of the Negative Bias Temperature Instability (NBTI) in silicon MOS FET devices admits the coexistence of two distinct components in the NBT-induced threshold voltage instability. One component would be quickly relaxing upon releasing the ‘stressing’ condition and the other would relax quite more slowly [1] [2]. Even if a complete agreement has not yet been reached about, it seems plausible that the quickly relaxing component should be associated with the capture-emission of semiconductor charge carriers on localized electronic states in the gate dielectrics [3]. On the other hand, it has been pointed out [4] [5] that the more slowly relaxing component arises from the variation of the number of localized interface states (traps), which can be related to the activation of pre-existing dangling bond defects at the Si/SiO₂ interface. It has been observed that the carrier capture-emission may exhibit unexpectedly long and widely spread characteristic times, and it is known since long time [2] [6] [7] [8] that NBT-activated interface traps can be annealed-out even at moderate temperature. Hence it appears that a simple recoverable/permanent classification of the two components is misleading, and that NBTI should be better described in terms of (at least) two microscopic mechanisms, both reversible and both characterized by dispersive kinetics.

Although the different aspects of NBTI are still the subject of restless investigation, it appears that this instability is more strictly electric-field dependent rather than gate-voltage dependent [5]. Discriminating the correct NBTI dependence on the stress conditions is important for the purpose of modelling the aging of real devices and is crucial for identifying the microscopic mechanisms behind.

The focus of this work is on the interface-state formation kinetics in p-channel type silicon MOS FET upon NBT stress (NBTS) in the range of low electric field, which is of practical interest for the reliable design of analogic circuits and which has been scarcely explored to date. We will present the experimental results drawn from a device designed on the purpose, consisting of a modified p-channel MOS FET with salicided polycrystalline silicon gate which can be joule-heated at wafer level. The key advantage of localized heating is that the device temperature can be quickly ramped down while the gate-to-source voltage is kept at the desired (stress) value, allowing a drastic reduction of the NBTI recovery effects. We found that the build-up kinetics of the integrated interface states can be modelled as a reaction-limited process, with Arrhenius-like rate constants and distributed activation energy. We will show that the kinetic details of the interface state build-up can be explained as the unbalance of forward and backward reactions, and that the activation energy of the forward reaction is a nearly square root function of the electric field in the oxide. We believe that the latter result is the first experimental evidence that the depassivation of native Si/SiO₂ interface defect would proceed through a stage where a nearly free positively charged hydrogen specie escapes from the coulombic attraction of the remaining negatively charged defect.

2. Materials and methods

The use of localized joule heating for studying device reliability at wafer level was been reported by Muth and Walter [9], who designed a MOS FET featured with a polycrystalline silicon heater located close to the transistor and capable of rising the device temperature up to 255
°C, starting from room temperature, and by Hamada and Roesch [10], who used a similar approach but with a thin-film resistor heater capable of achieving up to about 420 °C on the device under test. With the device used in the present work, the same gate electrode of a MOS FET serves also as the heating element, allowing reaching a ‘junction’ temperature of more than 320 °C, starting from room temperature. The device is derived from a p-channel MOS FET formed in phosphor doped, (100) oriented crystalline silicon, with SiO$_2$ gate dielectrics (oxide) of thickness $t_{ox} = 13$ nm and salicided p$^+$-doped polycrystalline silicon gate. The gate oxide is grown in two steps, steam at 900 °C and wet at 825 °C, and then thermally (NO) nitrided. The device process comprises a final annealing at 400 °C in N$_2$/H$_2$, atmospheric pressure, and 6% H$_2$ partial flow. The nominal total channel width $W$ is 12.5 µm and the nominal channel length $L$ is 5 µm. A controllable heating current can be forced through the salicided p-doped polycrystalline silicon layer, which serves both as the gate electrode and as the heating element. The typical threshold voltage extrapolated from the condition of peak transconductance at $V_{DS} = -0.1$ V is $-0.74$ V.

2.1 Device connections

With the connection scheme shown in Fig. 1 and under a PC control, the full measure-stress-measure sequence iteration, including charge-pumping measurement, was completely automated, i.e. no manual action was needed during the test of each device.

![Fig. 1 Schematic electric connections of the heated gate device. The heater/gate current supplied by the high-power Source Measure Unit HP SMU flows also through the gate actually. SW1 and SW2 are remote-controlled switches.](image)

The experimental setup comprises one High Power Source Measurement Unit (HP SMU), capable of supplying up to ±300 mA heating current at up to ±10 V DC, one Pulse Generator Unit (PGU) for the charge-pumping measurement, two switches (SW1 and SW2), and four medium power Source Generator Units (SMU). SW1 and SW2 are both set in the low position, during both the stress and the measurement of the transfer characteristics, and are both set in the high position, during the charge-pumping measurement. Since the electrical connection to the wafer substrate (Sub) is granted via the wafer holder (chuck), a minimum number of five bench
probes were used.

2.2 Stress-measure sequence

The sequence for starting a stress iteration is first to sweep the gate-to-source \( V_{GS} \) to the desired negative stress bias value, by ramping actually the source, drain, body and substrate terminals while keeping the heater/gate at ground, and then ramping up the heater/gate force voltage \( V_f \) up to the value needed for achieving the desired junction temperature, while keeping the \( V_{GS} \) value. The \( V_{GS} \) sweep is exploited to record an ‘upward’ transfer characteristic, at \( V_{DS} = -0.1 \) V. The stress stage is considered to start since the completion of the warming \( V_f \) ramp-up.

The sequence for ending a stress iteration is first to ramp the heater/gate force voltage down to zero while keeping the gate stress bias applied, wait a fixed time (3 s) for a coarse temperature stabilization, and then sweeping \( V_{GS} \) to zero. The \( V_{GS} \) sweep is exploited to record a ‘downward’ transfer characteristic, at \( V_{DS} = -0.1 \) V. The stress stage is considered complete at the beginning of the \( V_f \) ramp down (cool down). Since the independent ramping of the different device terminals requires the full control of the SMU units, a low-level programming language is used for implementing the stress/measure sequences.

The gate potential \( V_G \) is never actually probed during the stress/measure trials with this connection setup (two more bench probes would have been required for precise gate potential sensing). Due to the symmetry of the device, we assumed that \( V_G \) would be equal to \( V_f/2 \), on a first approximation. This assumption introduces some uncertainty in the actual value of \( V_G \), since the uncontrolled parasitic series contact resistance (probes, wires) is a non-negligible fraction of the overall heater/gate impedance. Another source of vagueness of the \( V_{GS} \) bias during the stress stage is inherent to the concept of the device itself since the heating current produces a non-negligible ohmic voltage drop in the gate electrode, across the MOS FET active area. We managed to minimize the systematic effects related to both these issues by splitting each partial stress stage into two sub-stages of same duration, and by reversing the heater/gate voltage (current) sign in the second sub-stage.

At junction temperature of 300 °C, the endurance of this device is limited to about six hours of cumulative stress, independently of \( V_{GS} \). After such amount of stress the resistance of the heating element increases without limit, eventually leading to an ‘open’ circuit failure, likely due to the electromigration of the salicided polycrystalline silicon.

2.3 Temperature sensing

Sensing the local temperature and estimating its uniformity is perhaps among the most challenging tasks in achieving the controlled heating of a microscopic device. In order to accomplish the highest level of confidence about the estimation of the stressing temperature we preliminary defined and measured the temperatures of three representative device regions: junction, channel and gate. The junction temperature \( T_j \) was estimated by measuring the forward bias of the diode formed by the MOS FET source and drain as the anode and the MOS FET body as the cathode, at \( V_{GS} = 0 \), forward current of 1 \( \mu \)A, and floating substrate. The forward bias is sensed at the beginning of each partial stress stage, for determining the current to be forced through the heater/gate in order to achieve the desired junction temperature. No wafer holder temperature control is used
in these experiments in order to avoid any possible disturbance from the thermal controller driving pulses on the device under test. No feedback loop was provided for ensuring stable device temperature during each partial stress stage: in order to limit the small device temperature increase, the duration of each partial stress was limited to never exceed two thousand seconds. The ambient temperature, ranging from 20 °C to 24 °C, was measured at the beginning of each trial to the purpose of determining the diode forward bias at the reference temperature of 25 °C, which turned out to be about 684 mV.

In order to increase the confidence level about the temperature reached by the active region of the MOS FET we preliminary estimated also the channel temperature $T_c$ through the characterization of the sub-threshold source current with the temperature. A drain-to-source bias $V_{DS}$ of −0.2 V was used for that characterization in order to ensure that the condition $|V_{DS}| > k_B T / q$ would be fulfilled even at the highest envisaged stress temperature. We focused on the source current $I_S$ since it behaves more ideally than the drain current at high temperature. Both the sub-threshold voltage at $I_S = 100$ nA and the maximum sub-threshold slope were characterized as a function of the wafer holder temperature. By repeating the sub-threshold measurements with the wafer holder at room temperature we found that a fair estimate of $T_c$ upon localized heating in this device is $T_c = T_a + (T_j - T_a) \alpha_T$ where $T_a$ is the wafer holder temperature and $\alpha_T = 1.0785 \pm 0.0075$.

Finally, an estimate of the heater/gate temperature $T_g$ was obtained by measuring the voltage drop at two kelvin sensing points located as close as possible to the MOS FET active area. In summary, we show the gate, junction and channel temperatures compared in Fig. 2 as a function of the heater/gate force voltage (notice that $T_c$ is computed from $T_j$ through the coefficient $\alpha_T$, not measured, in Fig. 2). The temperature of the heating element (the gate) $T_g$ turns out to be the highest one, and $T_c$ is bounded to lie between $T_j$ and $T_g$, in the whole device operating temperature range.

![Fig. 2](image)

**Fig. 2** Comparison of the estimated gate, junction and channel temperatures of this device upon localized joule heating.
2.4 Charge-pumping measurement setting

The charge-pumping (CP) current [11] is recorded before starting each partial stress stage, after the measurement of the downward transfer characteristic of the previous stress stage, and before the measurement of the upward transfer characteristic. The schematic connections of the CP setup correspond to Fig. 1 with both switches SW1 and SW2 in the upper position. The charge pumping current $I_{cp}$ is measured on the n-type bulk (body) of the transistor, which is hosted in a p-type substrate. The electrical connection to the wafer substrate is provided, allowing minimizing the spurious ‘geometric current’ component [12]. The two-level (trapezoidal) CP gate waveform is used. The default CP conditions are: (maximum) pulse amplitude $V_a = 5$ V, pulse frequency $f = 500$ kHz, pulse width $= 1$ µs, pulse leading time $t_l = 200$ ns, pulse trailing time $t_t = 200$ ns, drain and source voltage $V_D = 0$, substrate voltage $V_{sub} = -0.2$ V, and room temperature.

A variation with respect to the standard constant-amplitude-base-sweep method has been implemented here. In order to avoid excessive electric stress to the device under test, a limitation to the gate voltage during the CP base level sweep is introduced by keeping the gate voltage bounded to range between −3 V and +3.5 V. With this limitation, the pulse amplitude is not constant upon sweeping: when the voltage sweep starts the amplitude is at a low value, then it increases, flattening to $V_a$ around the middle part of the sweep, where the maximum $I_{cp}$ is typically detected, and finally it decreases to a low value until the base sweep ends. An example of charge-pumping current measured after different cumulative stress times, for stress at $V_{GS} = -9$ V and $T_j = 200$ °C, is shown in Fig. 3 as a function of the sum of the pulse base $V_{base}$ and pulse peak $V_{peak}$ voltage.

![Fig. 3](image-url) An example of charge-pumping current evolution upon NBTS at $V_{GS} = -9$ V and $T_j = 200$ °C. The dashed lines are the CP pulse base voltage $V_{base}$ and pulse peak voltage $V_{peak}$, as read on the right y-axis.

The integrated interface-trap density $N_{it}$ is estimated, as usual, according to:
where $\overline{D_{it}}$ is the average interface trap density per unit area and energy in the silicon energy gap window sensed by the CP measurement, $\Delta E$ the width of the energy window, $q$ the elementary charge and $A$ the device active area. The estimated silicon energy gap range sensed with this setup is estimated to be about ±0.36 eV around the intrinsic level.

\[
N_{it} = \overline{D_{it}} \Delta E = \frac{\max[-I_{cp}]}{qfA}
\]  

(1)

Fig. 4 Frequency charge-pumping characterization of a sample stressed at $V_{GS} = -9$ V and $T_j = 200$ °C for 18.5 h and measured at room temperature. The leading and trailing pulse times $t_l$ an $t_t$ are kept at 200 ns.

The charge-pumping measurement capability is proven by the room temperature frequency characterization of a sample stressed at $V_{GS} = -9$ V and $T_j = 200$ °C for 18.5 h shown in Fig. 4. To be noticed the very low CP frequency allowed with this setup (ten hertz) with $N_{it}$ in the mid $10^{11}$ cm$^{-2}$ range, and the substantially flat $N_{it}$ outcome even at very low frequency. The observed small slope might be the track of a background of bulk oxide traps [13]. Since some contribution of bulk traps cannot be excluded even at the highest (default) CP frequency of 500 kHz, the $N_{it}$ values reported in this work might be somehow overestimated.

3. Theory

According to a common interpretation, the NBT-related interface traps would result from the depassivation of native Si dangling bonds at the Si/SiO$_2$ interface, which were mostly passivated by hydrogen during the device fabrication process. We will assume that the build-up of Si/SiO$_2$ interface traps upon NBT stress would result from the unbalance of depassivation and passivation reactions [14].

\footnote{We also checked that repeating the customized CP measurement does not produce any spurious degradation in a virgin device.}
where $\text{Si}_3\equiv\text{Si}–\text{H}$ represents the hydrogen passivated Si/SiO$_2$ interface defect, $\text{Si}_3\equiv\text{Si}•$ the depassivated Si/SiO$_2$ interface defect, which is known to be the active amphoteric P$_b$-type center [15] [16] [17] [18] forming localized interface states in the silicon energy gap, H a hydrogen atom and H$_2$ a hydrogen molecule. We will call the depassivation reaction Eq. (2) the forward reaction, and we will assign it the reaction rate constant $k_f$. Furthermore, we will call the passivation reaction Eq. (3) the backward reaction, and we will assign it the reaction rate constant $k_r$. A questionable point about this reaction scheme is that both the activation and the deactivation of interface states is triggered very efficiently by atomic hydrogen [19], which is known to produce also additional (non-native) unstable interface states [20] [21] [22]. We shall assume that no significant concentration of atomic hydrogen is present in the device at time-zero, and that the atomic hydrogen produced by the two reactions Eq. (2) and Eq. (3) would readily move away from the interface without triggering significant depassivation.

3.1 Pseudo first-order kinetic model

Some simplifying assumptions may help to the aim of modelling the $N_{it}$ kinetics according to the reactions Eq. (2) and Eq. (3). We will basically assume, as a rough approximation, that the concentration of molecular hydrogen [H$_2$] is uniform in the gate oxide and that it does not change with time. This assumption can be justified if the hydrogen flow rate within the oxide is never rate-limiting [19], and if a non-zero H$_2$ concentration is present at time-zero into the gate oxide as a product of the N$_2$/H$_2$ annealing. It is also understood that only ‘native’ Si/SiO$_2$ interface defects are considered to be involved, i.e. that no extra interface defects are generated upon stressing. Under these assumptions, the depassivation and passivation reaction rates would be solely proportional to the concentration of passivated and depassivated Si/SiO$_2$ interface defects, respectively, and (pseudo) first-order reaction kinetics would result. Stated in symbols, the probability $p$ that an interface defect is found depassivated at the stress time $t$ would be a solution of the equation:

$$\frac{dp}{dt} = k_f(1-p) - k_r p$$  \hspace{1cm} (4)

where $k_f$ and $k_r$ may depend only on the individual defect properties ($k_r$ on [H$_2$] too), and on the parameters (voltage, temperature) of the external stimulus. If, starting from $t = 0$, the stimulus condition is changed and kept at a constant value, with the values of $k_f$ and $k_r$ associated to the new condition, then the solution of the kinetic equation Eq. (4) for the activation probability of a given defect is:

$$p(t) = p_0 + \left(\frac{k_f}{k_f + k_r} - p_0\right)\left[1 - \exp\left\{-(k_f + k_r)t\right\}\right]$$  \hspace{1cm} (5)

where $p_0$ is the probability that the defect is already activated at $t = 0$. 
3.2 Implementation of dispersive kinetics

Dynamic processes in which many timescales coexist are called dispersive. As a possible explanation of the dispersive character of the NBTI we will assume that the value of the relevant physical defect properties (the depassivation and passivation activation energies, typically) may not be the same for all defects, but may be spread, owing to the disorder of the amorphous SiO₂ layer [4] [23]. If the individual depassivation (passivation) events are independent, i.e. if the probability that a specific defect is found depassivated (passivated) at any time does not depend on the state of the other defects, then the implementation of a dispersive kinetics is simply achieved by the superimposition (integral) of elementary kinetic time-laws. Hence, since the concentration of depassivated interface defects may be assumed to be proportional to the concentration of interface traps \( N_{it} \) measured by the charge-pumping technique, we will state that the variation of \( N_{it} \) with time under static NBTS is given by the integral:

\[
N_{it}(t) = N_{it0} + n_0 \cdot \int_0^\infty dE_{af0} \int_0^\infty dE_{ar} D(E_{af0}, E_{ar}) [p(E_{af0}, E_{ar}, t) - p_0(E_{af0}, E_{ar})]
\]

where \( D \) is a distribution function to be determined, \( E_{af0} \) and \( E_{ar} \) are the defect depassivation and passivation energy, \( N_{it0} = N_{it}(0) \) is the time-zero \( N_{it} \) value, and \( n_0 \) the concentration of the interface traps at the condition that all the defects are depassivated. The distribution \( D \) is intended to be normalized such that its integral over the full range of \( E_{af0} \) and \( E_{ar} \) is equal to unity.

3.3 Physical aspects

According to the observations reported in [4] [24], a plausible form of the depassivation reaction rate constant \( k_f \) would be Arrhenius-like, with the activation energy \( E_{af} \) decreasing linearly with the magnitude of the electric field in the oxide, i.e. \( E_{af} = E_{af0} - \alpha F_{ox} \). A point with such model is that the experimental value of \( \alpha \) comes out to be unexpectedly high for the Si–H bond, which is just weakly polar. According to a theoretical model [25], the typical electric dipole moment \( p_{Si-H} \) of the Si–H bond (in amorphous silicon) is about 0.075 qÅ for a bond length \( l \) of about 1.5 Å, and increases with \( l \) at a constant rate \( e_d = dp_{Si-H}/dl \) of 0.21 q. By assuming the microscopic electric field at the Si–H bond location \( F_{loc} \) to be equal to the Lorentz field, and by thinking of the Si–H interaction potential as a short-range one, a rough estimate of the critical bond elongation at the bond-breaking condition \( \Delta l^* \) is:

\[
\Delta l^* = \frac{3\alpha}{(\epsilon_r + 2)e_d}.
\]

With the effective relative permittivity \( \epsilon_r \) ranging between 3.9 (SiO₂ static) and 7.8 (average of SiO₂ static and Si) and the published [4] \( \alpha \) (slope) value of 0.056 eVcm/MV = 5.6 qÅ, \( \Delta l^* \) would roughly range between 8 Å and 13 Å. Such large values would contrast with the assumption of short range Si–H interaction potential, and would raise a question about the correctness of using \( F_{loc} \) in the place of the macroscopic average \( F_{ox} \). It is interesting to notice that a strong electric field dependence of the apparent activation energy of the degradation rate is common also to other dielectric aging phenomena such as, for instance, the Time Dependent Dielectric Breakdown (TDDB) of most silica based materials [26] [27], which may be modelled with a linear dependence of the activation energy on the electric field as well.
The strong dependence of the activation energy on the electric field might suggest that the (long-range) electrostatic interaction potential is involved in the microscopic reaction responsible for the degradation [27] [28]. A consequence coming with the electrostatic interaction, however, is that the dependence of the activation energy on the electric field strength would be all but linear. In both cases of pure coulombic [29] and image-charge [30] potentials, for instance, a square root dependence on $F_{ox}$ should be observed. One may also check that with an electric dipole potential, the dependence would turn to $F_{ox}^{2/3}$. By following these considerations, we tried an Arrhenius-like depassivation rate constant $k_f$, with the activation energy varying according to a power of the oxide electric field strength:

$$k_f(E_{af}, F_{ox}, T) = \nu \exp\left(-\frac{E_{af} - \delta F_{ox}^{\beta}}{k_B T}\right)$$  (8)

where $E_{af}$ is the activation energy of the depassivation reaction at null electric field, $\nu$ a constant pre-exponential factor (prefactor) and $\delta$, $\beta$ two constants, which values have to be determined, $k_B$ the Boltzmann constant and $T$ the channel absolute temperature. As the passivation rate constant $k_r$ we tentatively adopted an Arrhenius form [14] with the same prefactor of Eq. (8), and a bias-independent activation energy $E_{ar}$:

$$k_r(E_{ar}, T) = \nu \exp\left(-\frac{E_{ar}}{k_B T}\right)$$  (9)

The fit of such dispersive kinetic model to the experimental data is made easier if the shape of the distribution $D$ of Eq. (6) is assumed a priori, so that just the values of the distribution parameters are left to be best fitted. We checked that, out of the most common ones, lognormal distribution of $E_{af}$ and $E_{ar}$ were best suitable to fit our $N_{it}$ drift data. Hence the statistical variables $E_{af}$ and $E_{ar}$ are understood to follow lognormal distributions, with mean $<E_{af}>$ and $<E_{ar}>$ and standard deviation $s_f$ and $s_r$, which values have to be determined. In order to limit the degree of freedom of the model, $E_{af}$ and $E_{ar}$ were forced to be strictly correlated, i.e. that their linear correlation coefficient $\rho(E_{af}, E_{ar})$ was forced to be equal to either $+1$ or $-1$.

In summary, our dispersive kinetic model comprises the integral Eq. (6) of the solution Eq. (5), with $p_0 = 0$, $k_f$ and $k_r$ defined according to Eq. (8) and Eq. (9), and with the values of the eight free parameters $n_0$, $\nu$, $\delta$, $\beta$, $<E_{af}>$, $s_f$, $<E_{ar}>$ and $s_r$ to be determined.

With reference to the generalized Reaction-Diffusion (GR-D) model [5], the present one relies on the plausible assumption that the diffusion of the mobile species is never rate-limiting. Hence any dependence on the spatial coordinate $x$ disappears, and only one ordinary differential (rate) equation in the variable time is kept. As further basic differences with the GR-D model we also highlight that: 1) the defect re-passivation is triggered by H$_2$, not by H; 2) non-zero constant H$_2$ concentration is assumed to be present in the device dielectrics at any stress time; 3) both the passivation and depassivation activation energies are distributed, accounting for the dispersive kinetic character, and 4) the activation energy of the defect depassivation is assumed to depend on the magnitude of the oxide electric field raised to an exponent, which is determined by fitting.
4. Results

We ran several static NBTI trials with $T_j$ between 100 °C and 300 °C, corresponding to $T_c$ ranging from about 106 °C to 322 °C, and $V_{GS}$ between $-1.2 \text{ V}$ and $-9 \text{ V}$, corresponding to $F_{ox}$ values ranging from 1 MV/cm to 6.6 MV/cm. Each trial covered typically more than four decades of cumulative stress time. An example of measured parametric drifts obtained upon stressing at $V_{GS} = -5 \text{ V}$ and $T_j = 200 \text{ °C}$ is shown in Fig. 5. The change of the threshold voltage $\Delta V_{thd}$ and $\Delta V_{thu}$ is extracted from the downward and upward transfer characteristic respectively, at $I_S = 6 \text{ nA}$. The change of the integrated interface trap density $\Delta N_{it} = N_{it} - N_{it0}$ is derived from the CP measurement.

![Fig. 5 Drift of the two parametric threshold voltages from downward and upward transfer characteristics, and of the integrated interface trap density upon NBTS at $V_{GS} = -5 \text{ V}$ and $T_j = 200 \text{ °C}$.

In Fig. 5, the dominant slope (dashed line) of the $\Delta V_{thd}$ trend is comparable with that reported with ultra-fast $V_{th}$ measurements [31] and is typical of ‘on-the-fly’ [32] measurements. The complete stress-measure-stress sequence is: stress $\rightarrow$ cool-down $\rightarrow$ downward $V_{GS}$ sweep $\rightarrow$ CP measurement $\rightarrow$ upward $V_{GS}$ sweep $\rightarrow$ warm-up $\rightarrow$ stress. Hence, the downward measurement yielding $\Delta V_{thd}$ might be assimilated to a relatively fast measurement, with a small amount of instability recovery, whereas the post-CP, upward measurement yielding $\Delta V_{thu}$ might be assimilated to a slow measurement, performed after a larger amount of recovery has already taken place.

The near ‘coincidence’ of the $\Delta V_{thu}$ and $\Delta N_{it}$ trends in Fig. 5 derives from setting a fixed ratio of the left $y$-axis scale ($-\Delta V_{th}$) to the right $y$-axis scale ($\Delta N_{it}$) equal to $q/C_{ox}$, where $q$ is the elementary charge and $C_{ox} = \varepsilon_{ox}/t_{ox}$ the specific gate oxide capacitance [33]. This coincidence be suggestive of a 1:1 charge to interface trap density ratio, unexpected for amphoteric defects. The early experimental evidence of a 1:1 ratio on NBTI experiments can be found in [6] [7] [34]. Its occurrence was recently pointed out for ultra-thin oxides too [2].
4.1 $\Delta N_{it}$ data fitting

The aim of this work is studying the kinetics of the interface trap density variation $\Delta N_{it}$. Some experimental outcomes are shown in Fig. 6, Fig. 7 and Fig. 8 (symbols), for $V_{GS} = -9 \text{ V}$, $-5 \text{ V}$ and $-2 \text{ V}$, respectively, and different temperatures, together with the model fit (lines). Other stress trials (not shown) were run at $V_{GS}$ of $-7 \text{ V}$, $-4 \text{ V}$, $-1.5 \text{ V}$ and $-1.2 \text{ V}$. The non-linear least squares technique was used for fitting the whole experimental data set to the model at once, to the aim of achieving the best accuracy on the free model parameter values. The data fitting routine was implemented as a MATLAB® code. Since the experimental values of $\Delta N_{it}$ turned out to span more than three decades a logarithmic transformation was applied, i.e. $\log_{10}(\Delta N_{it})$ was considered as the output instead of $\Delta N_{it}$. The best fit parameter values are shown in Table 1.

**Fig. 6** $\Delta N_{it}$ drift with stress time. $T_j = 100$, 150, 200, 250 and 300 °C. $V_{GS} = -9 \text{ V}$. CP measurements (symbols) and model predictions (lines).

**Fig. 7** $\Delta N_{it}$ drift with stress time. $T_j = 100$, 150, 200, 250 and 300 °C. $V_{GS} = -5 \text{ V}$. CP measurements (symbols) and model predictions (lines).
The model achieves a small overall root mean square error (rmse) of 0.062 decade over a response range covering about 3.6 decades, and a total of 702 observations, proving to be globally adequate. In particular, the apparent dependence of the drift ‘saturation’ level on $V_{GS}$ is caught as the effect of the dynamic balance of the forward and backward reactions, through to the ratio $k_f/(k_f + k_r)$ in Eq. (5). All the parameters of the model are statistically significant. The fitted values are estimated with small relative standard error, ranging from 0.9% for $\log_{10}(\nu/1\text{Hz})$ to about 10% for $s_r$.

Table 1: The parameters of $\Delta N_{it}$ fitted to the pseudo first-order kinetic model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$&lt;x&gt;$</th>
<th>~ std. err.</th>
<th>m.u.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_0$</td>
<td>$1.9 \times 10^{12}$</td>
<td>$0.2 \times 10^{12}$</td>
<td>cm$^{-2}$</td>
</tr>
<tr>
<td>$E_{af0}$</td>
<td>3.11</td>
<td>0.05</td>
<td>eV</td>
</tr>
<tr>
<td>$s_f$</td>
<td>0.51</td>
<td>0.01</td>
<td>eV</td>
</tr>
<tr>
<td>$\log_{10}(\nu/1\text{Hz})$</td>
<td>11.0</td>
<td>0.1</td>
<td>--</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.50</td>
<td>0.03</td>
<td>eV/(MV/cm)$^\beta$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.48</td>
<td>0.02</td>
<td>--</td>
</tr>
<tr>
<td>$E_{ar}$</td>
<td>1.52</td>
<td>0.03</td>
<td>eV</td>
</tr>
<tr>
<td>$s_r$</td>
<td>0.10</td>
<td>0.01</td>
<td>eV</td>
</tr>
</tbody>
</table>

The average forward activation energy $<E_{af}> = <E_{af0} > - \delta|F_{ox}|^\beta$ at any electric field can be computed according to the best fit values of Table 1. For an oxide electric field of 4 MV/cm, for instance, one finds $<E_{af}> = 2.14$ eV. The regression analysis points out that $E_{af0}$ and $E_{ar}$ are anti-correlated, i.e. a defect with $E_{af0}$ greater than average would likely have $E_{ar}$ less than average.
5. Discussion

The apparent decreasing of the activation energy of the interface-state build-up with the increasing of the gate voltage upon NBTS is not a novelty. In [4] the slowly relaxing part of the threshold voltage shift in NBTI was associated to the depassivation of native Si/SiO₂ interface defects (dangling bonds). By assuming a reaction-limited trap activation process with distributed activation energy, a linear relationship between the median Si–H dissociation energy of passivated dangling bonds and the oxide electric field was drawn therein for CMOS FETs of different technology nodes. A field dependence (slope) of $-0.056 \text{ eV} / \text{cm/MV}$ and a zero-field (flat-band) median dissociation energy of 1.5 eV were reported for a transistor with 2.1 nm thick gate oxide. In the present work a lower range of stress electric field is explored and the dynamic balance of forward and backward reactions is considered. Our model yields definitely higher average zero-field activation energy $\langle E_{af0} \rangle$ of about 3.11 eV. We believe that $\langle E_{af0} \rangle$ should be compared to the values drawn from high-temperature vacuum annealing experiments [35] [36]. The average values reported by Stathis [36] for (100) oriented silicon range between 2.86 eV and 3.13 eV, depending on different interface defect types ($P_{b0}$ and $P_{b1}$) and post-oxidation treatments. Those values were estimated by assuming first-order reaction kinetics with a prefactor of $2 \times 10^{13} \text{ Hz}$. By forcing the same prefactor value we would obtain $\langle E_{af0} \rangle = 3.53 \text{ eV}$, above the range of Stathis’ results.

The average activation energy of the forward reaction keeps relatively high at any practical oxide electric field value. At 4 MV/cm, for instance, $\langle E_{af} \rangle$ is still 2.14 eV. Also the average activation energy of the backward reaction $\langle E_{ar} \rangle$ is of comparably high value. Since the activation energy of the diffusion of neutral hydrogen species is known to range roughly between 0.18 eV and 0.45 eV (neutral species) and between 0.4 eV and 0.8 eV (positively charged specie), these results help justifying the basic assumption of this model that the $N_{it}$ formation kinetics would be reaction-limited.

A theoretical prefactor $\nu_H$ of about $2 \times 10^{13} \text{ Hz}$ is expected if the Si–H dissociation through a bending bond-breaking path is involved [37]. Stathis [36] reported that the least squares fit of his experimental outcomes for (100) interface defects would have yielded a much lower prefactor value of about $2 \times 10^6 \text{ Hz}$. Huard et al. [4] found a frequency ($1/\tau_0$) of $7.5 \times 10^7 \text{ Hz}$, independent of the oxide field, for 2.1 nm thick nitrided oxide. Our best fit prefactor value of $1.0 \times 10^{11} \text{ Hz}$ lies between those experimental values and $\nu_H$. In summary, the Arrhenius prefactor drawn from different kinds of experiments related to the interface defect depassivation tends to be lower, in orders of magnitude, than the expected theoretical value. According to Stesmans [38] such inconsistency should be ascribed to the limited experimental range but this seems not the case of this work, since the explored temperature range is rather wide. By admitting that the activation energy $E_{af0}$ would be not truly constant, but would increase linearly with the temperature, one may check that a lower effective prefactor would be measured. This appears not to be our case, however, since the discrepancy of the prefactor value is too high. Hence we believe that the discrepancy about our fitted prefactor value would arise from both the approximations adopted in our model and the possible systematic effects inherent to the experimental setup.

The maximum concentration of $P_{b0}$ and $P_{b1}$ defects after vacuum annealing is estimated [36] in $3 \times 10^{11} \text{ cm}^{-2}$ and $6 \times 10^{11} \text{ cm}^{-2}$ respectively. The value of $n_0$ extracted from our data is actually lower than two times (the factor two is due for the amphoteric character) the sum of those two concentrations. The $n_0$ value agrees well also with the $D_{it}$ data reported by Thoan et al. [18] for (100) Si, which vary in the low $10^{12} \text{ cm}^{-2} \text{eV}^{-1}$ range over the middle part of the energy
gap sensed by the CP technique.

We used dispersive first-order kinetics with not unique depassivation activation energy \cite{4} \cite{23} for fitting our drift data. The same choice was tried in \cite{36} and was confirmed to be appropriate by Stesmans \cite{38} for the depassivation of $P_b$ ((111)Si/SiO$_2$ interface) defects upon vacuum annealing. The standard deviation of the activation energy $s_f$ of 0.51 eV found here appears however higher than that mentioned in \cite{36} for (100)-orientation (0.17 eV, at most, by taking the defects $P_{b0}$ and $P_{b1}$ together). As a further indirect comparison term, standard deviation values of about 0.14 eV and 0.15 eV were reported \cite{14} for the passivation activation energy of $P_{b0}$ and $P_{b1}$ defects of (100)Si/SiO$_2$ by molecular hydrogen. Even if the latter result makes plausible that a high relative activation energy spread would be plausible for the depassivation reaction too, the rather high value of $s_f$ remains basically unexplained.

Our interface-trap measurements show a rather evident saturating trend with stress time, with bias-dependent saturation level. The latter feature, which is scarcely documented \cite{39} \cite{40}, is caught by this kinetic model as the effect of the dynamic balance of forward (depassivation) and backward (passivation) reactions. Compared with the results of hydrogen passivation experiments, our value of the average activation energy $<E_{ar}>$ lies well between the two average values reported by Stesmans \cite{14} for both $P_{b0}$ and $P_{b1}$, and agrees with the result of Ragnarsson and Lundgren \cite{41} too. In the Stesmans’ work, the prefactor was (correctly) put proportional to the concentration of molecular hydrogen, which was assumed to be equal to its solubility in vitreous silica at the relevant temperature and H$_2$ pressure. Since the average value of the prefactor frequency there ($8.3 \times 10^{11}$ Hz) is higher than ours the passivation reaction rate in our experiments appears to be lower than in Stesmans’ ones. This discrepancy should be accepted by considering that, since no external hydrogen source is present during the NBTS trials, the molecular hydrogen present into the device would be just the traces of that previously introduced during the N$_2$/H$_2$ annealing. In addition, the experiments by Stesmans were carried on at about atmospheric H$_2$ pressure, while a relatively low H$_2$ atmospheric partial pressure (6%) is used for the N$_2$/H$_2$ annealing treatment of our material. We also assumed $E_{ar}$ to be independent of $V_{GS}$. This assumption can be justified by considering that, in the NBTS $V_{GS}$ range, the relative position of the Fermi level is varied from the inversion level towards the valence band edge, in a range where the dependence of $E_{ar}$ on the Fermi level position would be expected to be weak \cite{41}.

We assumed a lognormal distribution of $E_{ar}$ and we found a distribution standard deviation $s_r$ of 0.10 ± 0.01 eV. This value is compatible with that reported in \cite{14} (0.15 ± 0.03 eV) for both $P_{b0}$ and $P_{b1}$, even if a different (Gaussian) distribution is found there.

Finally, according to the early Frenkel’s theory of electron transport in dielectric materials \cite{29}, the variation $\Delta E_d$ of the relevant activation energy with the electric field magnitude $F$ would follow a square root law:

$$\Delta E_d = q \sqrt{\frac{qF}{\pi\varepsilon}} \equiv K_F \sqrt{F}$$

(10)

where $\varepsilon$ is the optical dielectric permittivity of the material. Eq. (10) was simply derived by considering the coulombic attraction between the electron and the remaining positive ion, modified by the effect of the applied electric field, in the classical mechanics approximation. Since we obtained evidence of nearly square root dependence of the depassivation activation energy on the electric field in the oxide, we speculate about the possibility that the coulombic potential would
be involved here as well. In the case of the interface defect depassivation depicted here the escaping species would be a massive ion instead than an electron. Hence one should consider whether using the (oxide) permittivity at the optical frequencies would be still appropriate. The oscillation frequency of the Si–H bond bending mode $\nu_{\text{IH}}$, corresponding to a wavenumber $\nu_{\text{IH}}/c$ of 667 cm$^{-1}$, falls in the low wavenumber range of the infrared absorption spectrum of amorphous silicon dioxide [42]. This circumstance complicates the determination of the permittivity value, due to possible interaction (resonance) with the silicon dioxide optical phonons. Roughly, the suitable value of $\varepsilon$ for the breakage of the Si–H bond should should be comparable with both the optical ($n^2\varepsilon_0 = 2.13 \varepsilon_0$) and the static (3.9 $\varepsilon_0$) ones. For singly charged ions in SiO$_2$, the static and optical values of the constant $K_F$ in Eq. (10) would be equal to 0.384 eV(MV/cm)$^{-1/2}$ and 0.520 eV(MV/cm)$^{-1/2}$ respectively. We checked on our data that, by keeping the value of the exponent $\beta$ bounded to 1/2, the best fit value of $\delta$ would turn to 0.47 eV(MV/cm)$^{-1/2}$, i.e. between the ‘static’ and ‘optical’ $K_F$ values drawn from the Frenkel’s theory. This agreement may suggest that the Si–H dissociation at the Si/SiO$_2$ interface under NBT would run through a coulombic interaction energy path. More in details, both the near square-root dependence of the depassivation energy on the oxide electric field and the value of the coefficient $\delta$ suggest that the depassivation of Si/SiO$_2$ (near-) interface defects would proceed through a stage where a positively charged hydrogen specie attempts to escape from a fixed, negatively charged, interface defect. This picture should be meditated since one would expect that an amphoteric (interface) defect would quickly charge up positively under the NBTS condition. It may be reasonable however to admit that a freshly depassivated defect can stay negatively charged for a while, before capturing a hole from the inversion layer.

6. Conclusions

Experiments on a MOS FET that could be locally heated at wafer level suggest that the increase of the integrated interface-trap density in silicon p-channel MOS FET with thermally nitrided SiO$_2$ gate oxide upon homogeneous NBT stress can be modelled as the unbalance of two reaction-limited processes. From the values of the extracted activation energies, the two processes are identified as the spontaneous depassivation (dissociation) of native, hydrogen saturated, interface silicon dangling bonds and the (re-)passivation by molecular hydrogen, respectively. Accurate fitting of the experimental data is obtained by assuming first-order kinetics with distributed activation energy of both the depassivation and the passivation reactions. The observed dependence of the saturation level of the trap density trend on the gate bias is explained as the dynamic balance of the two reactions. Each of the parameters of the kinetic model can be associated to a measurable physical property of the Si/SiO$_2$ interface defects and most of their best-fit values are in reasonable agreement with published results. The experimental evidence of a nearly square root dependence of the depassivation activation energy on the electric field in the gate oxide is provided too. Our results suggest that the depassivation of Si/SiO$_2$ interface defects would proceed through a stage where a positively charged hydrogen specie escapes from the remaining negatively charged interface defect.
Acknowledgements

The author wishes to gratefully acknowledge G. Marchesi, who cared the layout engineering of the dedicated device, R. Stella for explanations about the temperature dependence of the MOS FET parameters, A. Balzarotti and F. Borella for training him on the high-temperature characterization bench, F. Belletti, for providing the device processing information, M. Rossetti for contributions in the data fitting, and S. Manzini for carefully reviewing the draft. The final version of the manuscript got substantially improved thanks to the reviewers’ extensive comments and suggestions.

References


[29] Frenkel J, On Pre-Breakdown Phenomena in Insulators and Electronic Semi-conductors


