Origin of NBTE Variability in Deeply Scaled pFETs

B. Kaczer, T. Grasser\textsuperscript{1}, Ph. J. Roussel, J. Franco\textsuperscript{2}, R. Degraeve, L.-A. Ragnarsson, E. Simoen, G. Groeseneken\textsuperscript{2}, H. Reisinger\textsuperscript{3}

\textsuperscript{1}Christian Doppler Laboratory for TCAD, Institute for Microelectronics, TU Wien, A-1040 Vienna, Austria
\textsuperscript{2}also ESAT, KU Leuven, Leuven, Belgium
\textsuperscript{3}Infineon Technologies AG, D-81730 Munich, Germany

Abstract—The similarity between Random Telegraph Noise and Negative Bias Temperature Instability (NBTE) relaxation is further demonstrated by the observation of exponentially-distributed threshold voltage shifts corresponding to single-carrier discharges in NBTE transients in deeply scaled pFETs. A SPICE-based simplified channel percolation model is devised to confirm this behavior. The overall device-to-device $\Delta V_{th}$ distribution following NBTE stress is argued to be a convolution of exponential distributions of uncorrelated individual charged defects Poisson-distributed in number. An analytical description of the total NBTE threshold voltage shift distribution is derived, allowing, among other things, linking its first two moments with the average number of defects per device.

Keywords: pFET, Negative Bias Temperature Instability, Random Telegraph Noise, variability, Random Dopant Fluctuations

I. INTRODUCTION

The large, micrometer-sized FET devices of the past CMOS technologies were considered identical in terms of electrical performance. Similarly, the application of a given stress resulted in an identical parameter shift in all devices. With the gradual downscaling of the FET devices, the oxide dielectric was the first to reach nanometer dimensions, thus introducing the first stochastically distributed reliability mechanism—the time dependent dielectric breakdown \cite{[1]}. With the shrinking of lateral device dimensions to atomic levels, variation between devices appeared due to effects such as random dopant fluctuation and line edge roughness. Similarly, application of a fixed stress in such devices results in a distribution of the parameter shift. Understanding these distributions will be crucial for correctly predicting the reliability of future deeply downscaled technologies. The purpose of this paper is to further illuminate the causes of the variation of NBTE in deep submicron devices, already discussed in Refs. \cite{[2, 3]}.

Charging and discharging of individual oxide defects has been readily observable in sub-micron FETs in the form of random telegraph noise (RTN). Recently, threshold voltage steps due to individual defects were also observed in NBTE relaxation transients \cite{[3-6]}. We have already argued in \cite{[4]} that both effects are in fact but the two facets of the same mechanism, with RTN being the channel/gate dielectrics system in the state of dynamic equilibrium, while NBTE relaxation corresponding to the perturbed system returning to this equilibrium (Fig. 1) \cite{[7]}. Here we will show that, identically to RTN amplitude distribution, the individual NBTE relaxation steps are exponentially distributed in amplitude \cite{[8]}. The exponential distribution will be confirmed with a simplified percolation stochastic model. Combined with the assumption of the Poisson-distributed number of trapped gate oxide charges, an analytical description of the total NBTE threshold voltage shift distribution is then derived. This allows, among other things, linking its first two moments with the average number of defects per device.

Finally, we will argue that NBTE in future downscaled devices will be treated as a stochastic ensemble of individual defects, Poisson-distributed in number per device, with each defect described by its impact on the channel conduction discussed here and its capture and emission times \cite{[5, 9]}.

![Figure 1](image.png)

Figure 1. (a) At constant bias conditions, oxide defects are charged by channel carriers and subsequently discharged back into the channel with a wide range of time constants controlled by a nonradiative multiphonon emission process \cite{[9]}. The system is in dynamic equilibrium, manifested by low-frequency noise or Random Telegraph Noise (RTN) in small devices. (b) Following the perturbation by NBTE stress, excess charged oxide defects gradually discharge and the system is returning to the dynamic equilibrium of (a), resulting in long NBTE transients.
II. EXPERIMENTAL

pFETs with nominal gate length \( L = 70 \) nm (metallurgic length \( L \approx 35 \) nm), width \( W = 90 \) nm, and HfO\(_2\) dielectrics with EOT = 0.8 nm were used in this study. Lanthanum has been incorporated in the oxide to boost the complementary nFET performance [10], but this was deemed to have no effect on the pFET NBTI [11] and especially on the NBTI variation discussed here.

All measurements were done at \( T = 125 \, ^\circ\mathrm{C} \). To compensate for the variability of these aggressively scaled devices, the initial threshold voltage \( V_{\text{th0}} \) of each DUT was first automatically determined using a fixed \( I_S \) criterion. The DUT was then stressed at \( V_G = V_{\text{dso}} - 1.2\) V using the extended Measure-Stress-Measure (eMSM) sequence [12]. Specifically, the source current \( I_S \) in the linear regime \( (V_D = -0.1 \, \text{V}) \) was recorded during a series of 7 stress and relaxation phases. The relaxation phase \( I_S \) measured at \( V_G \sim V_{\text{dso}} \) was converted to \( V_{\text{th}}(\Delta V_{\text{th}}) \) using the corresponding initial \( I_S \)-\( V_G \) curve. In our aggressively-scaled devices we estimate the \( V_{\text{th}} \) measurement accuracy to be \( \sim 1 \, \text{mV} \). This is related to the relatively small current \( I_S \) at low \( V_G \) and \( V_D \) and to the conversion of \( I_S \) into \( V_{\text{th}} \) given the sub-threshold slope of our devices. Subsequently, \( \Delta V_{\text{th}} \) was calculated as \( \Delta V_{\text{th}} = V_{\text{dso}}(\Delta V_{\text{th}}) - V_{\text{dso}} \). Note that because of the large RTN in the aggressively scaled devices, \( V_{\text{dso}} \) itself is not a fixed value, but rather fluctuates with time by up to \( \sim 10 \, \text{mV} \), as shown in Fig. 1a. Consequently, it is possible to obtain small positive NBTI \( \Delta V_{\text{th}} \) in a fraction of our devices, particularly for shorter stress times.

The resulting total threshold voltage shifts were found to be uncorrelated with the initial \( V_{\text{dso}} \) (Fig. 2), which confirmed that the NBTI mechanism is decoupled from sources of the \( V_{\text{dso}} \) variation and can be studied separately.

![Figure 2](image)

Figure 2. The lack of any correlation between the pFET initial threshold voltage \( V_{\text{dso}} \) and NBTI \( \Delta V_{\text{th}} \) allows us to study the NBTI mechanisms independently of the pFET variation.

III. RESULTS

A. Observation of single discharge events

Fig. 3a shows a typical result of the MSM stress measurement. As already reported previously [3,4,6], clear steps caused by single discharge events are visible in the NBTI relaxation transients. However, the average step height is significantly larger than those reported earlier. A single discharging event in many devices routinely exceeded 15 mV, and in several devices exceeded 30 mV, the NBTI lifetime criterion presently used by some groups. For comparison, \( \Delta V_{\text{th}} \) of less than 2 mV would be expected based on a simple charge sheet approximation. As will be shown below, the large observed step height amplitude is due to the aggressively scaled dimensions of the pFETs used.

The individual down-steps were detected in relaxation traces together with the corresponding relaxation times in all measured pFETs [9]. The step-detecting algorithm was designed to work automatically even in the presence of RTN in the traces. In order to ensure this, a detection resolution of 1 mV was used. Consequently, no steps smaller than 1 mV were detected.

An example of the result of this extraction is given in Fig. 3b. As already discussed previously [4], we did not observe any obvious correlation between the step height and trap emission time—all step heights appeared to be equally likely at all measured relaxation times in our 72 high-k pFETs. We have previously argued that this property is required in order to observe the long “featureless” log(\( t_{\text{relax}} \))-like relaxation tails [4]. This, however, also implies that defects exist with emission times faster than our measurement setup and we are therefore analyzing only the visible, but representative subset of all defects.

![Figure 3](image)

Figure 3. (a) A typical result of the eMSM sequence obtained on a single device: 7 NBTI relaxation transients following stress for the indicated times. Steps of varying heights due to single discharge events are clearly visible. (b) Step heights and the corresponding relaxation (emission) times for individual defects extracted from (a) [9].

B. Single discharge \( \Delta V_{\text{th}} \) distribution

A histogram of the step heights from transients following the longest stress time is constructed from all 72 pFET devices in Fig. 4a. The figure shows that the distribution of NBTI relaxation step heights is exponential, with their probability distribution function (PDF) being
where the scaling factor \( \eta \) is the mean \( \Delta V_{th} \) value for a single charge. The cumulative distribution function (CDF) corresponding to Eq. (1) is then

\[
F_i(\Delta V_{th}, \eta) = 1 - e^{-\frac{\Delta V_{th}}{\eta}},
\]

and the variance of this distribution is \( \sigma^2 = \eta^2 \). We note that the exponential distribution has been repeatedly reported for RTN amplitudes [13-15]. This similarity further strengthens the link shown in Fig. 1. We moreover note that the large range of possible \( \Delta V_{th} \)'s gives each defect its individual signature, which e.g. allows tracing its properties under various stress conditions [5,6].

The maximum-likelihood fit to individual \( \Delta V_{th} \)'s following the longest stress, shown in the cumulative plot in Fig. 4b, confirms the exponential distribution and allows extracting the average \( \Delta V_{th} \) shift per single discharge \( \eta = 4.75\pm0.3 \text{ mV} \) in our devices. An exponential distribution is also observed for shorter stress times, including the shortest \( t_{stress} \) shown in Fig. 4a, but given the amount of collected data we only assume it has the same \( \eta \). We note that \( \eta \) varying with the stress time could indicate e.g. charging of defects at varying depths. This is not expected for the NBTI mechanism, which has been repeatedly shown to be occurring at or very close to the Si/SiO\(_2\) interface [11]. This also agrees with the non-correlation between single \( \Delta V_{th} \)'s and emission times discussed above.

\[ 
\begin{align*}
\text{Figure 5.} & \quad \text{(a) Cumulative distribution of the total } \Delta V_{th} \text{ for 72 pFETs following stress at indicated stress times. With increasing } t_{stress}, \text{ the mean } <\Delta V_{th}> \text{ increases, while the fraction of devices with negligible } \Delta V_{th} \text{ decreases. (b) Cumulative distribution of } \Delta V_{th} \text{ for the same devices during relaxation following the longest stress of 1900 s. The opposite trends are observed.}
\end{align*}
\]

\[ 
\begin{align*}
\text{C. Total } \Delta V_{th} \text{ distribution}
\end{align*}
\]

Fig. 5a shows the distribution of the total \( \Delta V_{th} \) of 72 pFETs for increasing stress times, corresponding to an increasing number of charged defects. Such total \( \Delta V_{th} \) distributions are typically reported for a particular technology [3]. In Fig. 5a we note that the mean and maximum \( \Delta V_{th} \)'s are increasing with stress, its relative deviation is decreasing (the distribution is getting relatively tighter). Perhaps surprisingly, Fig. 5a also demonstrates that a fraction of devices exists with negligible \( \Delta V_{th} \) even after the longest stress. Overall, this fraction decreases with increasing stress time, i.e., with increasing mean \( \Delta V_{th} \). The opposite trends are observed in Fig. 5b when the devices are left to relax after the longest stress.

\[ 
\begin{align*}
\text{IV. DISCUSSION}
\end{align*}
\]

As we will now show, all the trends as well as the total \( \Delta V_{th} \) distribution itself, can be fully analytically described if i) the number of defects per device is assumed to follow a Poisson distribution, while ii) the impact of each individual defect on \( \Delta V_{th} \) is exponentially distributed.

\[ 
\begin{align*}
\text{A. Single discharge } \Delta V_{th} \text{ distribution}
\end{align*}
\]

The exponential distribution of single-charge \( \Delta V_{th} \) can be understood if non-uniformities in the pFET channel due to random dopant fluctuations (RDF) are considered [13-15]. The
threshold voltage of such a device corresponds to the formation of a conduction (percolation) path in the random dopant potential between Source and Drain (Fig. 6a). To zeroth order, depending on the position of the NBTI-stress-generated oxide charge, the conduction path could be either unaffected or obstructed by the new charged defect. In the latter case, the drop in the current has to be compensated by an increase of the gate voltage, resulting in the observed $\Delta V_{th}$.

A.1. Simplified channel percolation model

An accurate reproduction of this process is typically done through computation-intensive physics-based device simulations with RDF, line edge roughness, and other realistic effects [15,16]. Here we show that the essence of the mechanism can be qualitatively captured in a very simplified channel percolation model. We emphasize that in contrast to the all-encompassing device simulations, our aim is to keep the model as simple as possible and to include only the bare minimum of assumptions. We find it very instructional that the minimalist model correctly reproduces most of the common observations.

In our simplified model (Fig. 6b), a mesh of “elementary” FETs with random $V_{th}$’s, representing variations in the local potential, is set up to represent the channel of our pFETs. For the sake of simplicity, a uniform distribution of the random “elementary” $V_{th}$’s is used and short-channel effects are not considered. A script is used to generate 400 instances of the randomized mesh, to call SPICE to solve them, and to extract the $V_{th0}$ of the simulated pFET. As is typically experimentally observed, the resulting $V_{th0}$’s are normally distributed (Fig. 7a) and their variance scales reciprocally with the FET area (Fig. 7b).

A number $n$ of “charged defects” is then inserted, each represented by an additional $V_{th}$ shift of one random “elementary” FET in the netlist. A fixed value of $V_{th}$ shift for these “single oxide charges” is assumed, representing the fact that NBTI charges are occurring very close to (i.e., at a fixed distance from) the substrate interface. Subsequently, a new $V_{th}$ is calculated, resulting in $\Delta V_{th}$ for each pFET instance. For a single additional charged defect, the simplified model shows the $\Delta V_{th}$ distribution to be Weibull-distributed with $\beta \approx 0.8$ for a range of dimensions of the channel mesh (Fig. 8). This confirms that the above described process can be responsible for the observed exponential distribution of step heights (cf. Fig. 4b).

![Figure 6](image_url)

Figure 6. (a) An illustration of a percolation path in a random potential (from [17]) such as that between FET source and drain. (b) A mesh of “elementary” FETs with random $V_{th}$’s (voltage source in series with gate) representing (a) can be readily solved with SPICE.

![Figure 7](image_url)

Figure 7. The simple percolation model correctly reproduces (a) the normal distribution of initial threshold voltages $V_{th0}$ and (b) the variance $\sigma^2$ of $V_{th0}$ scaling reciprocally with the $L \times W$ (Pelgrom’s rule) [18].

![Figure 8](image_url)

Figure 8. Cumulative $\Delta V_{th}$ distributions generated for an increasing number of gate oxide defects $n$ by the simplified channel percolation model. For a single charged defect ($n = 1$), the model well reproduces the observed exponential distribution in Fig. 4b.

![Figure 9](image_url)

Figure 9. The step heights due to individual discharge events will increase with both decreasing $L$ and $W$, explaining why large single-charge steps can be expected in very small devices.
The simplified model predicts that \( \eta \) scales inversely with both \( W \) and \( L \) (Fig. 9), i.e., the smaller the device the larger the steps, thus explaining the large observed value of \( \eta \). This can be understood as the impact of a single charge relatively increasing as the device becomes smaller. We note here that a more thorough discussion of the dependence of \( \eta \) on \( W \), \( L \), EOT, and channel doping in the framework of RTN is given in Ref. [15], which infers \( \eta \sim L^{-1/2} \) for short devices.

The simplified model also predicts the total \( \Delta V_{th} \) distribution for the number of defects \( n > 1 \) (Fig. 8). Since the subsequent charge lateral locations are uncorrelated, the overall \( \Delta V_{th} \) distribution can be readily expressed as a convolution of individual exponential distributions (Eq. 1), and the PDF and the CDF are respectively described by

\[
 f_n(\Delta V_{th}, \eta) = \frac{e^{-\frac{\Delta V_{th}}{\eta}} N_{th}^n}{n!} \tag{3}
\]

and

\[
 F_n(\Delta V_{th}, \eta) = 1 - \frac{n!}{(n-1)!} e^{-\frac{\Delta V_{th}}{\eta}} \Gamma(n, \frac{\Delta V_{th}}{\eta}) \tag{4}
\]

The fraction in Eq. 4 can be also recognized as the regularized gamma function. CDF in Eq. 4 well describes the distribution can be therefore obtained by summing distributions \( F_n \) weighted by the Poisson probability

\[
 P_n(n) = \frac{e^{-N} N^n}{n!} \tag{5}
\]

In Eq. 5, \( N \) is the mean number of defects in the FET gate oxide and is related to the oxide trap (surface) density \( N_{ox} \) as \( N = W L N_{ox} \) (note that \( N \) is not an integer).

This reasoning then results in the total \( \Delta V_{th} \) CDF given by

\[
 F_n(\Delta V_{th}, \eta) = \sum_{n=1}^{\infty} e^{-N} N^n F_n(\Delta V_{th}, \eta) \tag{6}
\]

The corresponding PDF is

\[
 f_n(\Delta V_{th}, \eta) = e^{-N} \left[ \delta(\Delta V_{th}) + N e^{-\frac{\Delta V_{th}}{\eta}} \frac{d}{d\Delta V_{th}} \Gamma(2; N \frac{\Delta V_{th}}{\eta}) \right] \]  

where the hypergeometric function \( \Gamma(2; X) \) can be also written in terms of the modified Bessel function \( I_1 \) as \( e^{-\Delta V_{th}/\eta} F_1(\Delta V_{th}, \eta/\eta) \). The Dirac \( \delta(\Delta V_{th}) \) term represents the fraction of devices with 0 V shift [14], which decreases with increasing \( N \).

The advantages of describing the total \( \Delta V_{th} \) distribution in terms of Eqs. 6 and 7 are their relative simplicity and tangibility of the variables, while the analytical description allows further statistical treatment. The mean of the above-derived distribution is

\[
 \langle \Delta V_{th} \rangle = N \eta \tag{8}
\]

i.e., it should be independent of FET gate area \( L \times W \) provided \( N \) and \( \eta \) are respectively directly and inversely proportional to \( L \times W \). The variance of the distribution is then

![Figure 10. Eq. 6 for several values of the average number of defects N (lines). (a) Weibull plot emphasizes the fraction of devices with \( \Delta V_{th} = 0 \) V, cf. Fig. 5. Monte Carlo calculation with 1000 samples is included for comparison (symbols). (b) Eq. 6 in a probit plot rescaled to fit experimental distributions from Fig. 10 of Ref. [3], with the corresponding values of \( N \) and \( \eta \) readily extracted.](image-url)
\[ \sigma_{\Delta V_{th}}^2 = 2N\eta^2, \]  
(9)
i.e., it increases with decreasing gate area. The relative deviation \[ \sigma_{\Delta V_{th}}/\langle \Delta V_{th} \rangle = (2N)^{-1/2} \] is therefore decreasing with increasing \( N \), as observed in Fig. 5.

With the value of \( \eta \) extracted from the single discharge step height histogram in Fig. 2 earlier, we can use Eq. 8 to convert \( <\Delta V_{th}> \) to the average number of trapped defects \( N \). In our devices, \( N \) increases from 2.6 \( (t_{stress} = 0.24 \text{ s}, \text{Fig. 1}) \) to 6.9 \( (t_{stress} = 1900 \text{ s}) \), and then decreases to 3.4 \( (t_{relax} = 10 \text{ s}) \). These values correspond to effective trap densities of \( 1-2 \times 10^{11} \text{ cm}^{-2} \), typically observed for NBTI in large devices.

We acknowledge that obtaining the value of \( \eta \) as in Fig. 2 could be rather laborious. Eqs. 8 and 9, however, allow us expressing both \( N \) and \( \eta \) in terms of \( <\Delta V_{th}> \) and \( \sigma_{\Delta V_{th}}^2 \) as

\[ N = \frac{2\langle \Delta V_{th} \rangle^2}{\sigma_{\Delta V_{th}}^2}, \]  
(10)
and

\[ \eta = \frac{\sigma_{\Delta V_{th}}^2}{2\langle \Delta V_{th} \rangle}. \]  
(11)

This means that both \( N \) and \( \eta \) can be extracted from the first two moments of a measured total NBTI distribution, \textit{without having to characterize individual step heights}. This way we \textit{independently} obtain \( N \) increasing from 1.9 to 4.6 with stress and \( \eta \) varying between 7 and 9 mV. For the limited population of devices measured, these values are very close to those obtained directly by counting individual \( \Delta V_{th} \) step heights in Fig. 4. Note that other effects potentially increasing \( \sigma_{\Delta V_{th}} \), such as the variance of the initial \( V_{th0} \) (Fig. 1a) and FET variability (line-edge roughness, work-function fluctuations, etc.) have not been considered here.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure11}
\caption{(a) All large devices behave identically upon stress and are expected to fail when reaching the projected “hard” degradation criterion. (b) NBTI degradation in deeply scaled devices can be described in terms of the total number of defects in each device, their (voltage and temperature dependent) capture and emission times, and their impact on the device (demarcated by the size of the point). (c) Schematic showing the progress of stress in the three devices described in (b). The origin of the NBTI variability is apparent.}
\end{figure}

\begin{itemize}
    \item \textit{D. Reliability projection of future downscaled devices}

Finally, we review the described concepts to illustrate the paradigm shift in understanding NBTI reliability in aggressively scaled devices. Fig. 11a illustrates the reliability projection in large devices of the past. Since all devices are expected to behave identically, there is no device-to-device variation. Measuring and extrapolating one stress condition per device is therefore sufficient. Additionally, there is also the application-independent hard failure criterion.

In contrast to that, future failure criteria will be application-dependent and will depend specifically on the circuit immediately surrounding the device in question. Each decananometer device will be described by the Poisson-distributed total number of defects, each of which is characterized by i) voltage and temperature dependent capture time \( \tau_c \) and ii) emission time \( \tau_e \), and iii) its impact on the FET current (e.g. via \( \Delta V_{th} \)). This is schematically illustrated in Fig. 11b. The PDFs of all three parameters are known: i) and ii) appear to be uniform on the log scale (at least within the measured \( 10^{-6} - 10^6 \text{ s range so far} \) [12]. In this description, “permanent” defects can be seen as those with \( \tau_e \sim \infty \). Significant progress in describing the voltage and temperature dependence of \( \tau_e \) and \( \tau_c \) using nonradiative multiphonon theory has been made [9]. The exponential PDF of iii) has been justified here. Fig. 11c then schematically summarizes the progress of NBTI in the three hypothetical devices shown in Fig. 11b. For simplicity, only continuous stress is illustrated, although the specified parameters for each device (Fig. 11b) allow evaluating the degradation following an arbitrary waveform [5]. The origin of the NBTI variability at a given time or a given \( \Delta V_{th} \) is apparent.
V. CONCLUSIONS

The correspondence between RTN and NBTI relaxation was further strengthened by our observation of exponentially distributed step heights in NBTI transients in deeply scaled pFETs. A simplified channel percolation model was devised to illustrate and to confirm this behavior. The overall \( \Delta V_{th} \) distribution was argued to be a convolution of exponential distributions of uncorrelated individual charged defects. The analytical description derived for this distribution should prove useful for both reliability data analysis and simulations of deeply-scaled CMOS circuitry. The proposed picture allows us to predict that the reliability of future deca-nanometer devices will be treated as a stochastic ensemble of the Poisson-distributed total number of defects, each characterized by capture and emission times and its impact on the FET current.

REFERENCES


