Oxide Traps in MOS Transistors: Semi-Automatic Extraction of Trap Parameters from Time Dependent Defect Spectroscopy

P.-J. Wagner*, T. Grasser*, H. Reisinger†, and B. Kaczer‡
*Christian Doppler Laboratory for TCAD in Microelectronics at the Institute for Microelectronics
TU Wien, Gußhausstraße 27–29/E360, A-1040 Vienna, Austria
Phone: +43-1-58801/36022, Fax: +43-1-58801/36099, Email: {grasser|pjwagner}@iue.tuwien.ac.at
†Infineon Technologies AG, D-81730 Munich, Germany
Email: hans.reisinger@infineon.com
‡IMEC, Kapeldreef 75, B-3001 Leuven, Belgium
Email: kaczer@imec.be

***Abstract*—An algorithm to extract the statistical properties of oxide traps in ultra-small metal oxide silicon field effect transistors (MOSFETs) is presented. The algorithm uses data from time dependent defect spectroscopy (TDDS) measurements. It works reliable in a great range of circumstances, and automatically detects and correctly processes variations in the measurement data due to channel percolation path modulations. The algorithm is designed to require minimal user interaction, making parameter extraction from a large base of experimental data feasible.**

I. Introduction

Being a charge-driven device, the metal oxide silicon field effect transistor (MOSFET) suffers from a number of effects that build up parasitic charges in the oxide or at the silicon/oxide interface. Particularly annoying in this respect are charges that vary with time and/or bias of the device. While the charge traps at the silicon/oxide interface can follow the bias very quickly, oxide traps have a wide range of capture and emission times. Exactly this property made them ideal candidates to explain flicker noise ($1/f$ noise) [4, 6, 10]. Recently, we have been investigating the role of oxide traps in the negative bias temperature instability [2]. Negative bias temperature instability (NBTI) is one of the most critical degradation mechanisms in p-channel MOSFETs today. It is observed when a pMOSFET is subjected to negative bias at the gate with the other terminals grounded. The degradation is considerably accelerated at elevated temperatures. In our study [2] we employed a new technique termed 'time dependent defect spectroscopy' (TDDS). Within this technique, a large amount of measurement data has to be collected and processed, calling for an automatic and sophisticated algorithm. Such an algorithm will be presented in this paper.

II. Time Dependent Defect Spectroscopy (TDDS)

In large-area devices, where a large ensemble of defects acts simultaneously, the impacts of individual defects are averaged out. Consequently, unambiguous identification of the individual defects' physics is impossible. In small-area devices, only a handful of defects are present. In these devices, observation of the effects of individual defects and extraction of their characteristic properties is possible. The differences between large and small devices can also be observed in their low frequency noise: In large devices the defects create flicker noise, which has a uniform $1/f$-spectrum and therefore allows only to draw conclusions about the whole ensemble. In small devices, on top of the ubiquitous flicker noise, random telegraph noise is observed. From this random telegraph noise, characteristic parameters of the individual defects can be extracted.

The duality between large and small devices can also be seen with NBTI, where the recovery of small devices proceeds in discrete steps. Contrary to the prediction of the MOSFET charge sheet approximation, these steps are not equal in height. This fact is explained by the non-uniform current distribution in the channel [7]: A defect in a current percolation path has a much higher influence on the current than a defect in a region where for some reason the current density is lower. Hence, each defect produces current steps of individual magnitude, allowing to relate the observed steps to the individual defects.

The TDDS proceeds as follows: The MOSFET is subjected to NBTI stress at some temperature T with a gate voltage of V_s for some time $t_{\mathrm{s},1}$. The following recovery phase at a gate voltage V_r is recorded from $t_{\mathrm{r,min}}$ to $t_{\mathrm{r,max}}$, where $t_{\mathrm{r,min}}$ is restricted by the measurement equipment, and $t_{\mathrm{r,max}}$ is set to a value that permits all defects which were charged within $t_{\mathrm{s},1}$ to emit with sufficient probability. This procedure is repeated N times in order to gain a statistically significant number of samples. After that, the whole procedure of recording N transients is repeated $M-1$ times with different stress times $t_{\mathrm{s},2}, \ldots, t_{\mathrm{s},M}$. As such, the TDDS is similar in spirit to the successful deep-level transient spectroscopy (DLTS) technique [5], which has also been applied to small devices [3]. However, rather than assuming that after application of a charging pulse for a certain amount of time all defects are fully charged [3], we exploit the fact that the capture time constants are widely

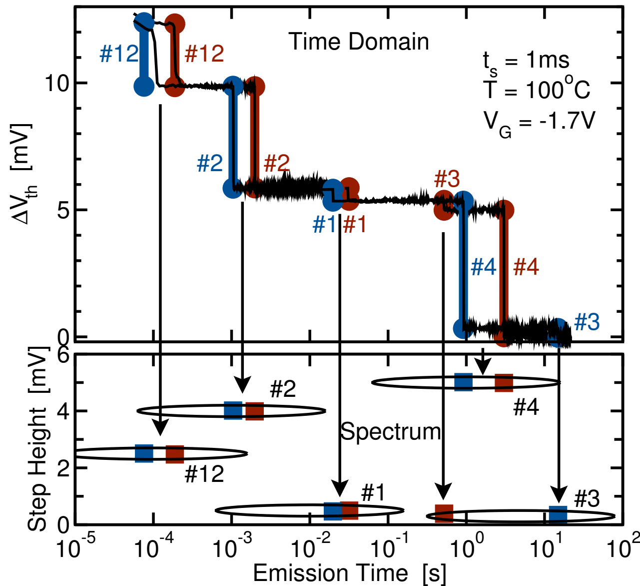

Fig. 1. Two typical recovery transients of a previously stressed pMOSFET. The measured data are given by the (slightly noisy) thin black lines in the top part of the figure. The thick blue and red lines together with the symbols mark the automatically extracted emission times and step heights (bottom).

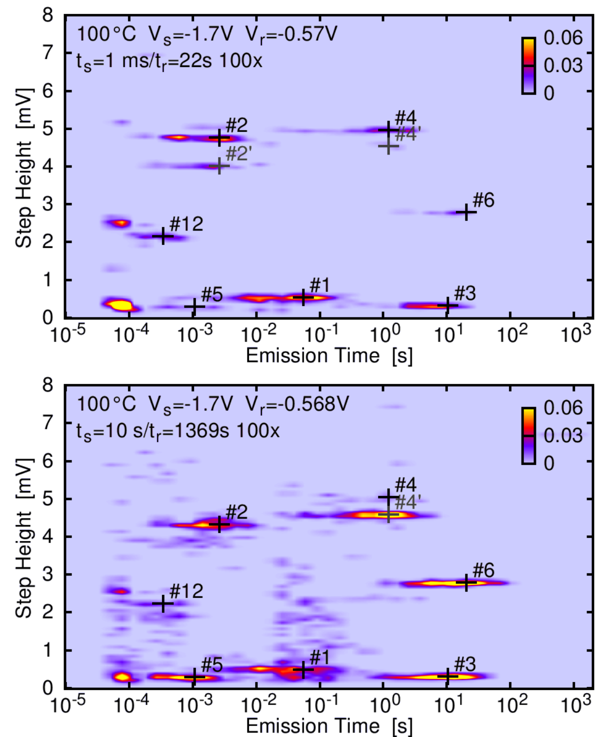

Fig. 2. Spectral maps obtained after two different stress times, $t_s = 1\,\mathrm{ms}$ (top) and $t_s = 10\,\mathrm{s}$ (bottom). With increasing stress time, the number of defects contributing to the map increases as defects with $\bar{\tau}_c \lesssim t_s$ have a significant probability of being charged after stress.

distributed, an issue also observed in DLTS literature as a non-saturating behavior in the DLTS spectra [9]. It is particularly this time-dependence of the spectral maps obtained from the TDDS that allows for a detailed assessment of the physical phenomenon.

From the $N \times M$ recovery transients the detrapping events and their associated step heights are detected. The events, represented by tuples (τ_e, d) of their emission time and step height, are binned into 2D-histograms, one histogram for each row of N experiments with the same stress time. The resulting M histograms, which have the axis of emission times in logarithmic scale, are termed 'spectral maps'. Figure 1 illustrates the creation of the spectral maps. Two examples of such spectral maps obtained from a production quality pMOSFET (cf. Section III) are shown in Figure 2. Clearly, marked clusters of (τ_e, d) pairs evolve, the intensity of which typically follows $P_c = 1 - \exp(-t_s/\bar{\tau}_c)$, where $\bar{\tau}_c$ is the average capture time. The position on the emission time axis remains fixed with stress time, while the step height is subject to slight variations, depending on the most dominant current conduction path [1], which can change with the occurrence of additional charged defects. Each cluster corresponds to an individual defect and is labeled accordingly in the maps.

The described procedure can be repeated for different stress voltages and temperatures. Moreover, additional information can be obtained by varying the readout voltage V_r, which alters the channel percolation paths. This method also proves successful if some defects are not well-separated in their step heights.

III. Experimental Setup

We used pMOSFETs with dimensions of $W = L = 0.1\,\mu\mathrm{m}$ and a 2.2 nm thick plasma nitrided gate oxide. The charge-sheet approximation gives $\Delta V_{\mathrm{th}} \simeq 1\,\mathrm{mV}$ for these devices. Thus the gate area is small enough to conveniently resolve the trapping/detrapping of a single carrier in the channel, and large enough to have at least a handful of active defects in the device. Stress pulses ranging from $t_{s,1} = 1\,\mu\mathrm{s}$ to $t_{s,M} = 100\,\mathrm{s}$ were applied to the gate. The shift in the threshold voltage corresponding to a preset drain current (typically $1\,\mu\mathrm{A} \times W/L$) was directly recorded after the pulse using the fast feedback loop described in [8].

IV. Semi-Automatic Defect Parameter Extraction

A considerable number of recovery traces must be recorded and analyzed in order to exhaustively characterize the defects of a particular device. It is therefore desirable to devise a method to process these transients with an as high as possible degree of automatization. Currently, our algorithm works as follows: After measuring N recovery traces to each of the M different stress times $t_{s,1}, \ldots, t_{s,M}$, the detrapping events and their associated step heights are detected. The events, represented by tuples (τ_e, d) of their emission time and step height, are binned and displayed in their corresponding M spectral maps. By inspection, clusters of events are tagged. Clusters in different spectral maps located around the same emission times and step heights belong to the same defect.

Note that the clusters' step heights may be somewhat fuzzy because of channel percolation path modulation.

In the next step, to each set of M clusters belonging to a particular defect the M theoretical probability distributions are fit by a least-squares method. The theoretical probability distributions are derived under the following assumptions: (i) The distributions in step height are Gaussian. Their mean values are different for each spectral map, allowing for a different channel percolation path after experiments of different stress time. The events' deviations from the mean values are attributed to measurement noise, and are of no particular interest. Therefore, the Gaussian distributions' standard deviations were empirically set to a single, fixed value for all defects and all spectral maps. (ii) The distributions in emission time are exponential. Since we assume the capture and emission processes independent, the mean emission time will not depend on the stress time; hence, the same exponential distribution is used for a defect's events in all of the M spectral maps. (iii) The emission times and step heights are statistically independent.

Local variations in the electrostatic configuration in the vicinity of a defect due to percolation path modulation actually have an impact on the defect's capture and emission behavior. Since we consider the effects of these variations in the distribution in step height, one could argue that these variations should also be taken into account in the distribution in emission time. This would right away invalidate assumption (ii), and by the now existing correlation between step heights and emission times also assumption (iii). In the vast majority of emission events encountered in our experiments, however, we could not observe significant modulation of the mean emission times.

The histograms are logarithmically spaced with regard to the emission times. Therefore, it is not possible to directly fit the exponential probability density function to the histogram data. Instead, the probability density function integrated over intervals $[\tau_j, \tau_{j+1}]$ with $\tau_{j+1} = \lambda\tau_j$ must be used; the τ_j are the histogram bins' boundaries and λ is the logarithmic increment. Regarding the step heights, the (empirically set) standard deviation is usually smaller or in the order of magnitude of the bin width. Therefore, also the Gaussian probability density has to be integrated over intervals $[d_k, d_{k+1}]$ with $d_{k+1} = d_k + \kappa$, where d_k are the bins' boundaries and κ is the bin width. Altogether, the theoretical probability distributions read

$$a_i\left[\exp\left(-\frac{\tau_j}{\bar{\tau}_\mathrm{e}}\right) - \exp\left(-\lambda\frac{\tau_j}{\bar{\tau}_\mathrm{e}}\right)\right] \times \tfrac{1}{2}\left[\operatorname{erf}\left(\frac{d_k+\kappa-\bar{d}_i}{\sqrt{2}\sigma_d}\right) - \operatorname{erf}\left(\frac{d_k-\bar{d}_i}{\sqrt{2}\sigma_d}\right)\right] , \quad (1)$$

where $i = 1, 2, \ldots, M$ runs through all spectral maps. The amplitudes a_i account for the fact that after the corresponding stress times $t_{\mathrm{s},i}$ not all defects will have captured a charge; they are parameters of the fit. The statistically independent probability distributions are parametrized by the mean step heights $\bar{d}_i$ and the mean emission time $\bar{\tau}_\mathrm{e}$ common to all spectral maps, as discussed above. In our work, the standard deviation of the step height was set to $\sigma_d = 100\,\mu\mathrm{V}$, which worked reasonably well. Smaller values resulted in a more narrow distribution in the direction of the step height, making it impossible to capture all events constituting a cluster. This in turn makes the fit more unstable by arbitrarily increasing the overall fit error. Significantly larger values for σ_d degraded the precision of the fitted mean values $\bar{d}_i$, ultimately leading to the fits of adjacent clusters 'running into another'.

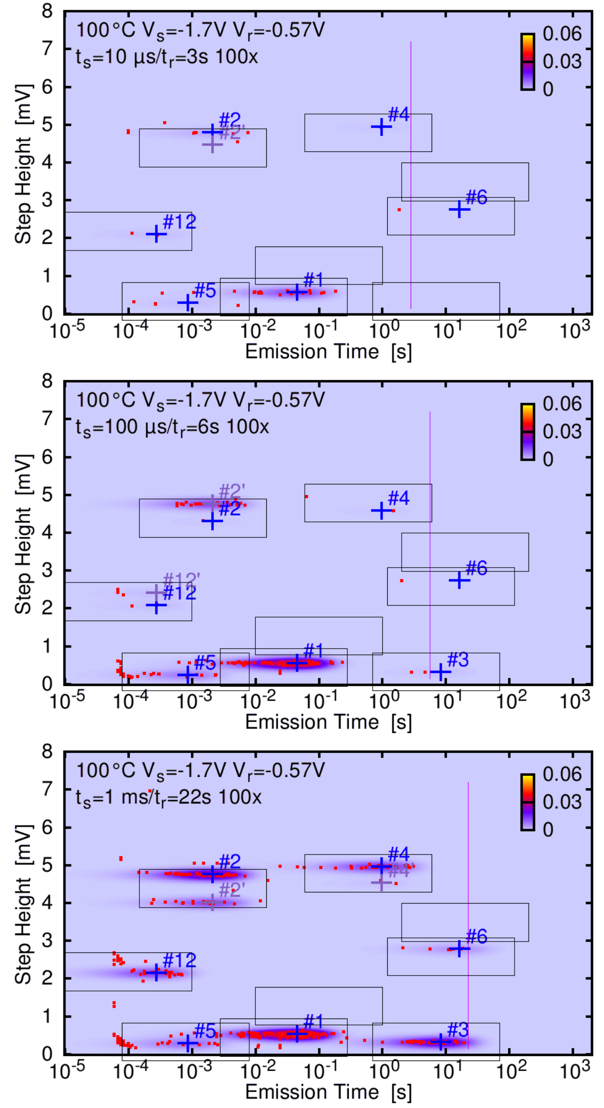

Fig. 3. Example of the fitting procedure. The groups are indicated by the rectangles. Empty rectangles belong to defects with larger capture times which are not yet visible. The fitted spectral map is given by the color gradient while the actual emission events which contribute to the histograms are shown as small red squares. The crosshairs show the fitted tuples $(\bar{\tau}_\mathrm{e}, \bar{d}_i)$, and are labeled by the number of the defect. If split-clusters were detected, the second sub-cluster is labeled with a prime.

The centers $(\bar{\tau}_\mathrm{e}^\circ, \bar{d}^\circ)$ of the previously designated clusters are used as initial guesses for the nonlinear and hence iterative least-squares algorithm. To sensibly limit the amount of data, only events in the subsets $\{(\tau_\mathrm{e}, d) : |d - \bar{d}^\circ| < 5\sigma_d \wedge |\log_{10}(\tau_\mathrm{e}/\bar{\tau}_\mathrm{e}^\circ)| < 1\}$ of the spectral maps are used for fitting this particular defect's clusters. The sum of probabilities in each subset is used as initial guess for the corresponding

cluster's a_i. If this sum is zero for a particular subset i', the respective spectral map is excluded from this fit; $a_{i'} = 0$ is left as result, and $\bar{d}_{i'}$ is set to some dummy value, which will be ignored in the postprocessing.

In order to aid convergence to physical solutions of the least-squares algorithm, the fit parameters are incorporated indirectly through mappings $a_i = \alpha(A_i)$ with $\alpha\colon \mathbb{R} \to]0,2[$ and $\bar{d}_i = \delta(\bar{D}_i)$ with $\delta\colon \mathbb{R} \to]\bar{d}^\circ - 5\sigma_d, \bar{d}^\circ + 5\sigma_d[$. Note that although the a_i in theory should be bounded to the interval $[0,1]$, spurious events sometimes cause some $a_i > 1$. This somewhat unphysical result should not be suppressed by using a mapping $\alpha\colon \mathbb{R} \to]0,1[$. Doing so would just impair convergence of the fit procedure and arbitrarily increase the final fit error. In a refined approach, clusters having $a_i > 1$ could be used to trigger a more sophisticated algorithm of selecting which events are discarded when fitting these clusters. The mappings themselves should be sufficiently smooth, and preferably invertible; we used $\alpha(x) = 2/(1+\exp(x))$ and $\delta(x) = \bar{d}^\circ + 5\sigma_d(1-\exp(x))/(1+\exp(x))$.

A. Split Clusters

As discussed previously, changes in the channel's percolation paths can alter a defect's step height. In fact, this regularly happens when comparing recovery traces from rows of experiments with different stress times; hence the introduction of M distinct mean step heights $\bar{d}_1, \ldots, \bar{d}_M$. But this effect is also observed within a row of experiments with the same stress time, especially if one defect strongly modulates another defect's percolation path and the two defects have similar mean emission times. In this case, the cluster of events for that particular defect in that particular spectral map splits into two (or rarely even more than two, although we will just deal with the case of two) sub-clusters, cf. Figure 4. Since the sub-clusters are often well separated by a multiple of the adopted standard deviation σ_d, the fit algorithm arbitrarily converges to either of the sub-clusters. This raises two issues: First, convergence is poor, especially if the number of events in the two sub-clusters is similar, and the fit results are unstable in the sense that minor alterations of the initial values or other fit parameters may lead to the fit 'snapping' to the other sub-cluster. And second, a considerable amount of events is not covered, thus giving wrong amplitudes a_i.

To rectify these issues, an attempt to detect split clusters is made after fitting all clusters with (1). If in some spectral maps split clusters are detected, the fit procedure is repeated, with (1) replaced by

$$a_i\Big[\exp\Big(-\frac{\tau_j}{\bar{\tau}_e}\Big) - \exp\Big(-\lambda\frac{\tau_j}{\bar{\tau}_e}\Big)\Big] \times \tfrac{1}{2}\Big[b_i\Big(\operatorname{erf}\Big(\frac{d_k+\kappa-\bar{d}_i}{\sqrt{2}\sigma_d}\Big) - \operatorname{erf}\Big(\frac{d_k-\bar{d}_i}{\sqrt{2}\sigma_d}\Big)\Big) + (1-b_i)\Big(\operatorname{erf}\Big(\frac{d_k+\kappa-\bar{d}'_i}{\sqrt{2}\sigma_d}\Big) - \operatorname{erf}\Big(\frac{d_k-\bar{d}'_i}{\sqrt{2}\sigma_d}\Big)\Big)\Big] \tag{2}$$

for those spectral maps where clusters are split. The newly introduced fit parameter b_i represents the fraction of events

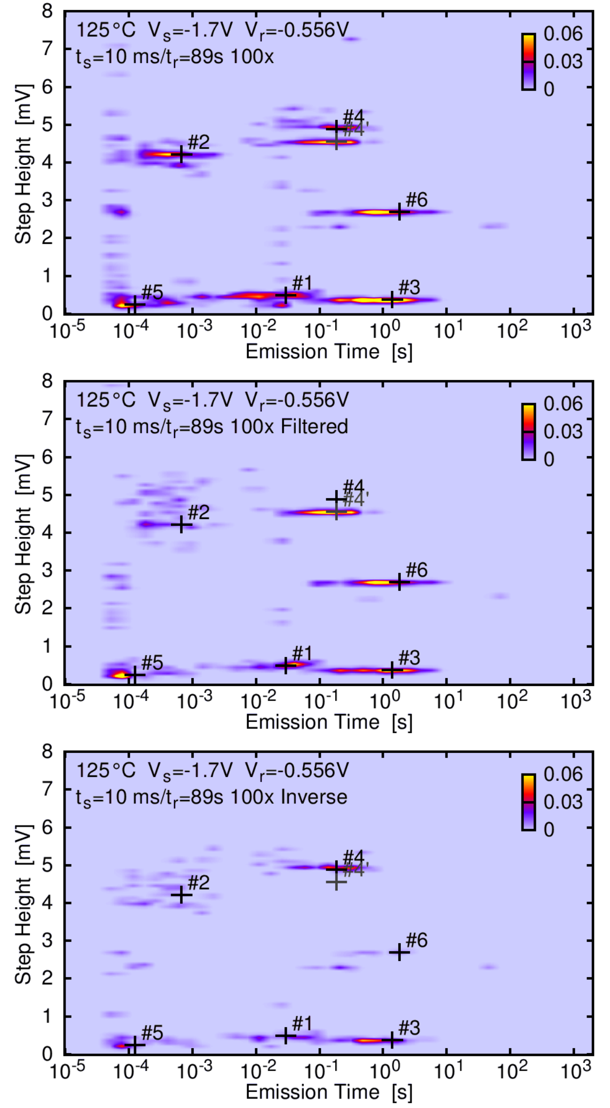

Fig. 4. Quite regularly, a single defect produces peaks of different height in the maps. For instance, #4 appears as #4 and #4′ (top). Filtering out all traces that produce an event in #4′ (top) reveals that as soon as #6 is occupied, #4 produces an event in the #4 cluster. Otherwise, for an unoccupied #6, an event in the #4′ cluster is obtained (bottom).

covered by the Gaussian peak at $\bar{d}_i$. The remainder of events are covered by the peak at $\bar{d}'_i$, which is also an additional fit parameter. The parameters b_i and $\bar{d}'_i$ are also mapped via $b_i = \alpha(B_i)$ and $\bar{d}'_i = \delta(\bar{D}'_i)$.

The detection algorithm for split clusters works as follows: (The spectral map index i is omitted in this paragraph for the sake of brevity.) After the first fit, in each spectral map the sum of probabilities $\tilde{p}$ represented by 'outliers', i.e. points of the cluster with $2\sigma_d < |d-\bar{d}| < 5\sigma_d$ and $|\log_{10}(\tau_e/\bar{\tau}_e)| < 0.8$ is calculated for the defect in question. Furthermore, by evaluation of the first and second moments in direction of d, the center of gravity of the outliers, $\tilde{d}$, and the deviation from this center of gravity, $\tilde{\sigma}_d$, is calculated. A split-cluster situation is assumed if the following three conditions are fulfilled: (i)

The uncovered probability $\tilde{p}$ is larger than 3 % of the total probability represented by this group of events. (ii) The center of gravity of the outliers is spaced more than σ_d apart from $\bar{d}$, i.e. $|\bar{d}-\tilde{d}| > \sigma_d$. (iii) The normalized standard deviation from this second center of gravity is less than 5 %, i.e. $\tilde{\sigma}_d/\tilde{d} < 0.05$. Requirement (ii) ensures that the second sub-cluster is well separated from the first cluster, otherwise the fit parameter b will not be well-defined, resulting in slow convergence. (If the sub-clusters are close, they have been captured by the fit of (1) anyway.) Requirement (iii) ensures that the outliers really form a secondary, separated cluster and are not spread randomly. The empirical quantity $\tilde{p}$ is used to get an initial guess for the fit parameter b through $b^\circ = 1-\tilde{p}$; the initial guess for the fit parameter $\bar{d}'$ as directly obtained as $\bar{d}'^\circ = \tilde{d}$.

B. Fit Error Estimation

Fit errors are calculated in a standard fashion, e.g.

$$\Delta a_i = \sqrt{\frac{\chi^2}{\eta}}\sqrt{C_{\mathrm{vv}}}\,\alpha'(A_i)\ , \qquad (3)$$

where χ^2 is the sum of the fit residuals' squares, η is the number of degrees of freedom, i.e. the number of overall data points in the M clusters minus the number of fit parameters, and C_{vv} is the main-diagonal element of the least-squares covariance matrix corresponding to A_i. The last term accounts for the fact that the fit is actually carried out through the map α, of which α' is its derivative. The errors of $\bar{\tau}_{\mathrm{e}}$, $\bar{d}_i$, $\bar{d}'_i$, and b_i are calculated analogously.

C. Extraction of Capture Times

Using the intensities $a_1,\ldots,a_M$ determined from the set of spectral maps, the capture time constant can be determined. Provided charge trapping is given by the first-order differential equations obtained from standard capture and emission models, $a_i = P_{\mathrm{c}}(t_{\mathrm{s},i}) = 1-\exp(-t_{\mathrm{s},i}/\bar{\tau}_{\mathrm{c}})$ provided that $\tau_{\mathrm{e}} \gg \tau_{\mathrm{c}}$. Similarly to the previous fits, the capture time constant is evaluated using a standard least-squares algorithm. Figure 5 shows an extraction of capture time constants.

V. Summary and Conclusions

We developed an algorithm accompanying the time dependent defect spectroscopy to analyze oxide traps. The algorithm works reliably also in more complicated situations, namely when modulations of the defects' characteristic step heights occur, while requiring a minimum of user attention. It is ideally suited to process large amounts of TDDS measurements. In a future version, the still required manual definition of clusters may be replaced by an automatic cluster search algorithm, making the procedure fully automatic.

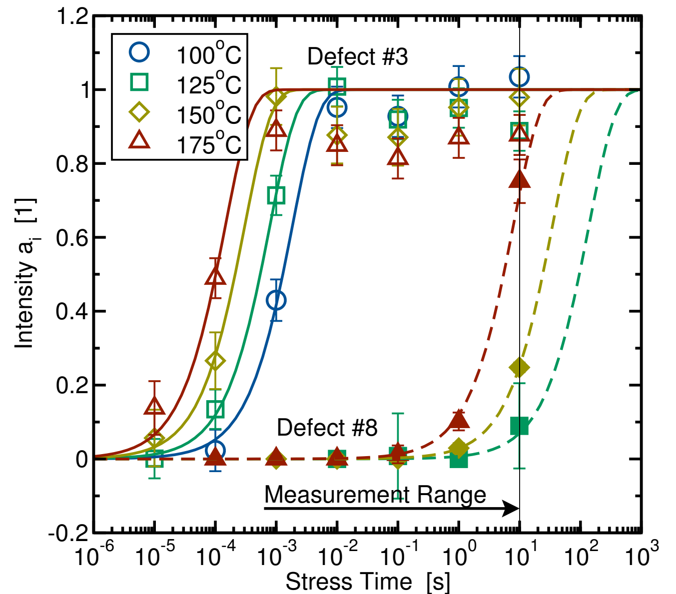

Fig. 5. Example for the extraction of the capture time constant for four different temperatures. With increasing stress time and temperature the number of defects contributing to the map increases. This makes the identification of the discrete steps more difficult and the noise level in the maps increases. Consequently, the clusters become wider, resulting in a spurious decrease of the intensities a_i, which may show a visible deviation from 1 even for $t_{\mathrm{s}} > \bar{\tau}_{\mathrm{c}}$.

References

[1] A. Asenov, R. Balasubramaniam, A. Brown, and J. Davies, "RTS Amplitudes in Decananometer MOSFETs: 3-D Simulation Study," *IEEE Trans.Electron Devices*, vol. 50, no. 3, pp. 839–845, 2003.

[2] T. Grasser, H. Reisinger, P.-J. Wagner, W. Gös, F. Schanovsky, and B. Kaczer, "The Time Dependent Defect Spectroscopy (TDDS) Technique for the Bias Temperature Instability," in *Proc. Intl.Rel.Phys.Symp.*, 2010, (in press).

[3] A. Karwath and M. Schulz, "Deep Level Transient Spectroscopy on Single, Isolated Interface Traps in Field-Effect Transistors," *Appl.Phys.Lett.*, vol. 52, no. 8, pp. 634–636, 1988.

[4] M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures: A new perspective on individual defects, interface states and low-frequency $(1/f)$ noise," *Advances in Physics*, pp. 367–468.

[5] D. Lang, "Deep-Level Transient Spectroscopy: A New Method to Characterize Traps in Semiconductors," *J.Appl.Phys.*, vol. 45, no. 7, pp. 3023–3032, 1974.

[6] A. McWhorter, "$1/f$ Noise and Germanium Surface Properties," *Sem.Surf.Phys*, pp. 207–228, 1957.

[7] H. Mueller and M. Schulz, "Conductance Modulation of Submicrometer Metal-Oxide-Semiconductor Field-Effect Transistors by Single-Electron Trapping," *J.Appl.Phys.*, vol. 79, no. 8, pp. 4178–4186, 1996.

[8] H. Reisinger, U. Brunner, W. Heinrigs, W. Gustin, and C. Schlunder, "A comparison of fast methods for measuring NBTI degradation," *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 531–539, 2007.

[9] D. Vuillaume, J. Bourgoin, and M. Lannoo, "Oxide Traps in Si-SiO_2 Structures Characterized by Tunnel Emission with Deep-Level Transient Spectroscopy," *Physical Review B*, vol. 34, no. 2, pp. 1171–1183, 1986.

[10] M. Weissman, "$1/f$ noise and other slow, nonexponential kinetics in condensed matter," *Rev.Mod.Phys*, vol. 60, no. 2, pp. 537–571, 1988.