

Chapter 2

Application of On-Chip Device Heating for BTI Investigations

Thomas Aichinger, Gregor Pobegen, and Michael Nelhiebel

Abstract This chapter introduces a new experimental approach allowing to switch the temperature of a device in a very fast and defined way. The new hardware tool, which we will herein refer to as polycrystalline silicon heater or simply poly-heater, allows overcoming previously strict experimental limitations regarding the speed of temperature variation and the accessibility of temperature range. Having broadened one's mind to the possibility of switching the temperature very fast at arbitrary points in time, the poly-heater technique opens up unprecedented experimental capabilities for bias temperature instability (BTI) characterization. For instance, one can achieve decoupling of stress and characterization temperature by making use of degradation quenching. Such or similar experiments can probe our understanding of the BTI physics in a novel manner.

2.1 In Situ Device Heating and Cooling Using On-Chip Poly(-crystalline Silicon) Heaters

In this paragraph we are going to address the calibration procedure of the poly-heater tool, investigate its capabilities, and demonstrate dedicated experimental setups which become feasible once having a well-calibrated and stable temperature switching tool at hand.

Having embedded the poly-heater in the measurement software environment, the technique allows to run stress-recovery experiments with variable temperature. From these experiments new fundamental features of [Negative bias temperature](#)

T. Aichinger (✉)

Infineon Technologies Austria AG, Siemensstraße 2, Villach, Austria

e-mail: thomas.aichinger@infineon.com

G. Pobegen • M. Nelhiebel

KAI Kompetenzzentrum für Automobil- und Industrielektronik, Europastraße 8, Villach, Austria

e-mail: gregor.pobegen@k-ai.at; michael.nelhiebel@k-ai.at

instability (NBTI) arise, helping to clarify conflicting issues in literature and allowing to draw new conclusions which lead to a more consistent microscopic picture of the degradation and the recovery mechanism.

2.1.1 Hardware Assembly and Poly-heater Design

Polycrystalline silicon resistors surrounding an active device can be used to perform fast in situ heating on a single device on wafer level which is commonly applied in time-critical **Fast wafer level reliability (fWLR)** monitoring [1–5]. The main advantage of the poly-heater is that it provides an elevated stress temperature without the use and the limitations associated with a conventional heating system like a thermo chuck. Although this alone is already a valuable feature, it does not even begin to explore the multitude of possibilities one finds in a more scientific use of the tool. By correct calibration and automation of the poly-heater–device system, the temperature becomes a quasi arbitrary experimental parameter which can be switched easily by more than ± 200 K within a couple of seconds. For such an application conventional thermo chuck systems are unsuitable because they are slow and their heating/cooling durations depend strongly on the difference to the target temperature. Also, when attempting to keep the junction biases applied during the temperature switch, it is a mandatory requirement not to lose probe needle contact. Since heating and cooling of a wafer on a thermo chuck involves considerable thermal expansion of wafer, needles, and pads, this request cannot be fulfilled over a wide temperature range without continuous manual readjustments. As opposed to heating the whole wafer on the thermo chuck, in situ heating by poly-heater wires is very local. Hence, there is no thermal expansion of needles and pads, which saves us from losing mechanical contact during the temperature switch.

Once calibrated and implemented in the software, the poly-heater feature opens up unprecedented possibilities for device characterization and reliability testing. Its application potential is thereby way beyond the scope of **NBTI** characterization. The concept of in situ heating by poly-heater wires has been taken up also for instance for thermo cycling [6] or for performing on chip high temperature annealing of irradiated **p-channel metal oxide semiconductor (PMOS)** dosimeters [7].

A common challenge of in situ heating can be found in the fact that the device which we want to heat is in most cases a distance away from the actual heating source. Consequently, the poly-heater wires are always hotter than the tested structure itself, which results in a temperature gradient between heater, device, and ambient. In our following experiments, the ambient temperature is always controlled by the programmable temperature of a conventional thermo chuck. From a design point of view, the distance between the poly-heater wires and the active areas of the device must be large enough and well isolated to prevent leakage current flow between the individual components. However, a larger distance between heater and

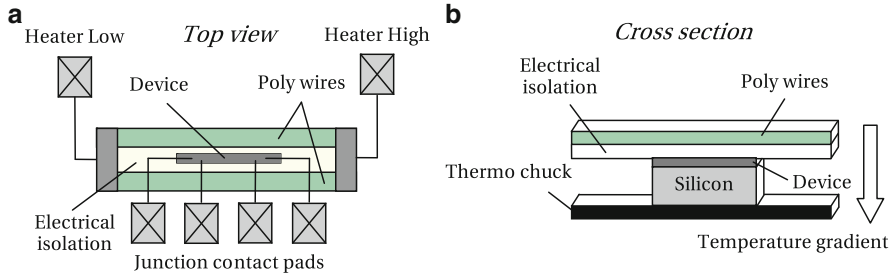


Fig. 2.1 (a) A schematic illustration of a poly-heater–device system placed on a thermo chuck. Electrically isolated poly-heater wires surround the silicon device. When a voltage is applied to the wires, a current flows and the dissipated heat elevates the temperature within the subjacent electrically active device regions. A schematic cross section of the poly-heater–device system is given in (b). Due to the lower chuck temperature, a temperature gradient arises between heater and wafer bottom

device reduces the accessible temperature range because more power has to be applied to the wires in order to adjust a certain elevated device temperature. Last but not least, the time delay for restoring thermal equilibrium must also be taken into account as we switch the heater on or off abruptly. The farther the heater is away from the device, the longer it takes to restore thermal equilibrium. Thus, in order to find an optimal design of the heating source and the active device for a particular application, the above-specified issues have to be considered carefully.

Figure 2.1 schematically illustrates a simple design proposal of an active transistor surrounded by a poly-heater. To adjust a well-defined ambient temperature, the wafer is placed on a thermo chuck which is held at a fixed programmed temperature. Depending on the thermo chuck temperature, a certain power has to be applied to the heater in order to reach a certain elevated device temperature. From a layout point of view the heater should overlap the device considerably in order to guarantee a homogeneous temperature distribution over the whole active transistor area. A calculation of the poly-heater and the device temperature as a function of the applied heater power needs some extra work, which will be treated in more detail in Sect. 2.1.3, but is not needed if the target device temperature is within the temperature range which is accessible by the thermo chuck system. As such, we present in the following subsections a procedure allowing to experimentally determine the power supply which is necessary to bring an active transistor to a certain device temperature. We will evaluate both the heater and the device temperature which enables us to estimate also the temperature gradient between the actively heated poly-heater wires and the active device. Furthermore, we will extract the transient heating and cooling characteristics for different target temperatures and discuss effects that may delay restoration of thermal equilibrium and destabilize the adjusted target temperature.

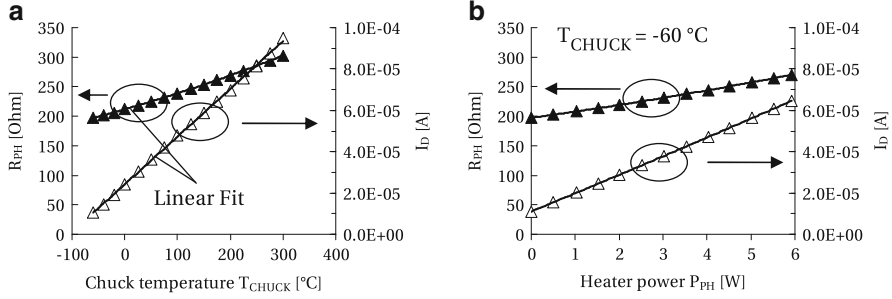


Fig. 2.2 (a) The poly-heater resistance R_{PH} (full triangles) and the drain current I_D (open triangles) as a function of the chuck temperature T_{CHUCK} . R_{PH} is characteristic for the poly temperature while I_D reflects the interface temperature of the active device. (b) The poly-heater resistance R_{PH} (full triangles) and the drain current I_D (open triangles) as a function of the heater power P_{PH} ($T_{CHUCK} = -60$ °C). The increase in the heater power causes a linear increase of R_{PH} and I_D

2.1.2 Calibration of Heater and Device Temperature

In order to extract the poly-heater and device temperature as a function of the applied power, we call on an appropriate temperature-dependent parameter of the material. In the case of the highly doped poly-heater wire, the temperature dependent electric resistance (R_{PH}) would be such a parameter. In the case of the **Metal oxide semiconductor (MOS)** transistor the forward current of the source/bulk diode or the source/drain current (around the threshold voltage of the device) can be used as an appropriate thermometer. For reliability issues, the source/drain current (I_D) is the preferred reference since it directly reflects the temperature of the interface between the silicon substrate and the gate oxide. This interface is of major interest because most studies suggest this region to be the location of concern for **NBTI**.

In the following, we demonstrate the temperature calibration using a lateral **PMOS** transistor embedded into two poly-heater wires similar as illustrated in Fig. 2.1. To record reference values for the poly-heater resistance (R_{PH}) and the drain current (I_D), we heat the wafer on the thermo chuck from -60 to 300 °C and measure R_{PH} and I_D at different temperatures, cf. Fig. 2.2a. The sense currents and voltages (which represent certain temperatures) must be chosen carefully in order to prevent self-heating during the measurement. Within the scanned temperature range (-60 to 300 °C) the increase of the poly-heater resistance and the drain current can be fitted very well by a polynomial of first (linear) or second order. From a physical point of view, the increase in the poly resistance is due to a reduction of the carrier mobility due to enhanced lattice scattering at higher temperatures, while the increase in the drain current is due to an enhancement of the concentration of inversion carriers at higher temperatures. From the coefficients of the polynomial fit, we can interpolate the poly-heater and device temperature for arbitrary heating powers. Figure 2.2b illustrates the poly-heater resistance and the drain current as a

function of the power supply (P_{PH}). The chuck temperature was -60°C . Note that the relation between the measured poly-heater resistance and the applied heating power is not perfectly linear. The reasons will be elaborated in more detail in the next section. When applying 0 W ($P_{PH} = 0\text{ W}$), the poly resistance and the drain current correspond to their values extracted for -60°C in Fig. 2.2a. As we increase the power supply from 0 W toward 6 W, the poly-heater resistance and the drain current grow simultaneously.

2.1.3 Maximum Accessible Temperature Range

Another remarkable application of the poly-heater technique can be found in the ability of reaching device temperatures far beyond the scope of conventional thermo chuck systems. This allows, for instance, probing a much wider temperature range to study Arrhenius-type processes. Extremely high temperatures can be reached when providing additional heating power at the maximum temperature range of the thermo chuck. The calibration procedure of the poly-heater–device system in such high temperature regimes requires, however, somewhat more effort because the above-described method for device temperature determination is only applicable for target temperatures within the maximum temperature range of the thermo chuck system. To determine the drain current for the experimentally not accessible range one would need to rely on TCAD device simulations. But the simulation itself would be based on material parameters in the high temperature regime which are rather difficult to characterize. Additionally, at high temperatures also the low biasing required to measure the drain current could already lead to degradation of the device, which makes an accurate calibration impossible. To avoid these problems we developed a method which extrapolates the functional dependence of the drain current *on the power supplied to the poly-heater* rather than on the temperature.

To illustrate this, we measured the increase of the device temperature T_{DV} (determined from the drain current increase) with poly-heater power P_{PH} at several different chuck temperatures and depicted the result in Fig. 2.3. It becomes evident from Fig. 2.3 that the device temperature does not depend linearly on the poly-heater power. A linear dependence is suggested for simple Joule heating where dissipated power is directly converted into a proportional temperature increase. The reason for the stronger than linear increase of the device temperature with power is the simultaneous increase of the thermal resistivity R^{th} of the materials which surround the heater and the device. This effect allows the device to become even hotter than what is expected from simple Joule heating. The functional dependence of the device temperature increase is, however, not a simple low-ordered polynomial and can therefore not be straightforwardly extrapolated.

In order to find an appropriate analytical expression for the observed increase of the device temperature, we exploit the definition of the thermal resistance and use Joule's first law that the electrical power is transferred into an equivalent amount of heat flow $\dot{Q}$, to find

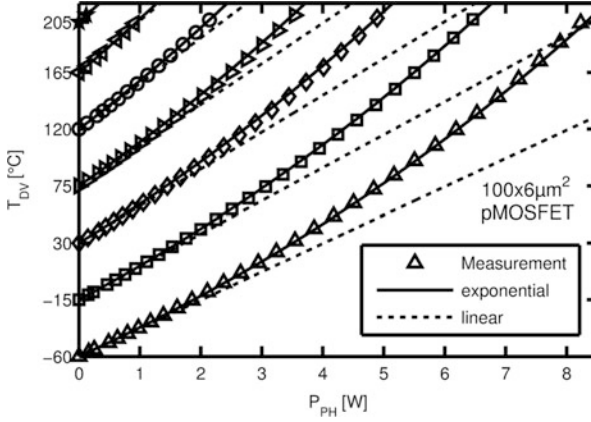


Fig. 2.3 Dependence of the device temperature on the power supplied to the poly-heater within the maximum accessible temperature range. The different marker symbols indicate different chuck temperatures. The increase of the device temperature is not linear

$$R^{\text{th}}(T) = \frac{dT}{d\dot{Q}} \equiv \frac{dT}{dP} = T'(P). \quad (2.1)$$

This formula is a differential equation for the temperature as a function of power, which is in our case $T_{\text{DV}}(P_{\text{PH}})$, and opens a way to calculate the temperature directly when the temperature-dependent thermal resistance $R_{\text{sub}}^{\text{th}}(T)$ of the substrate is known. In particular, for silicon the thermal resistance depends on temperature as $R^{\text{th}}(T) \propto T^{1.324}$ which leads to a rather complicated solution to (2.1) (not shown). However, a linear approximation of this power law introduces an error which is below the resolution limit for the measurement of the thermal resistance. Consequently, the thermal resistance can be approximated as

$$R_{\text{sub}}^{\text{th}}(T) = R_{\text{sub},0}^{\text{th}} (1 + \alpha(T - T_0)) \quad (2.2)$$

with three constants $R_{\text{sub},0}^{\text{th}}$, α , and T_0 . By combining (2.1) and (2.2), with the requirement that the device temperature equals the chuck temperature T_{chuck} when there is no heater power supplied, we obtain

$$T(P) = T_0 - \frac{1}{\alpha} + \left(\frac{1}{\alpha} + T_{\text{chuck}} - T_0 \right) \exp\left(\alpha R_{\text{sub},0}^{\text{th}} P\right). \quad (2.3)$$

This means the functional dependence of the device temperature on the power supplied to the heater is an exponential function.

With (2.3) it is now possible to calculate the device temperature directly from the power supplied to the heater. However, the coefficients for the thermal resistance must be known. Using literature values can make the extrapolation erroneous because the thermal resistance can vary largely with the doping level [8] and only limited information exists about the thermal resistance of the interface between the

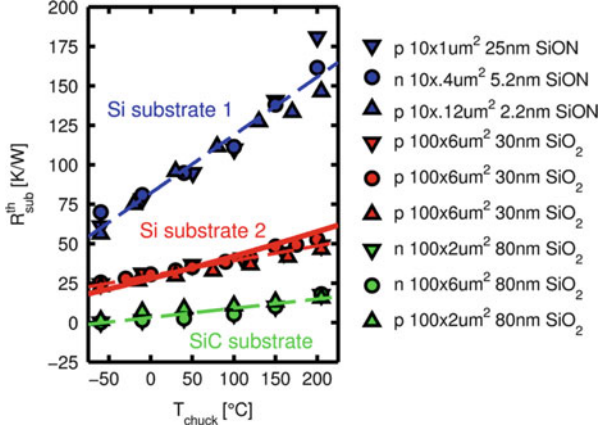


Fig. 2.4 Experimentally determined thermal resistances of several different devices from different technologies. In the legend, n and p refer to the MOSFET channel type, respectively, the micrometer size is the width times the length of the device, and the nanometer size is the thickness of the gate oxide. The *thin dashed lines* are linear fits of the data points of a group, while the *thick solid line* displays the theoretical dependence of $R^{\text{th}} \propto T^{1.324}$

wafer and the chuck [9, 10]. As such, it appears beneficial to *measure* the thermal resistance for the given wafer/chuck system. This can be achieved by applying only little power to the poly-heater and measuring the change of the device temperature for different chuck temperatures

$$R_{\text{sub}}^{\text{th}}(T_{\text{chuck}}) = \frac{T_{\text{DV}}(P_{\text{PH}}) - T_{\text{chuck}}}{P_{\text{PH}}} \quad (2.4)$$

to acquire $R_{\text{sub},0}^{\text{th}}$, α , and T_0 . We measured several different devices of different technology and different substrate types as depicted in Fig. 2.4. The type and thickness of the substrate have the largest impact on the thermal resistance. For all investigated technologies the data can be reasonably well approximated by a linear relationship, further supporting the previously stated assumptions. The application of the extrapolation method (2.3) can now be compared within the temperature range of the thermo chuck to measurement data, as depicted in Fig. 2.5. The exponential equation (2.3) estimates the device temperature with only a few percent relative error.

Since previous investigations have shown that the thermal resistance of common semiconductors keep their functional dependence of the change of the thermal resistance with temperature until the melting or sublimation point [11–13], it is a safe assumption that (2.3) will hold also for the high temperature regime. As such it can be assumed that the small relative error of the method will also apply for temperature ranges much above the highest temperature of the chuck system.

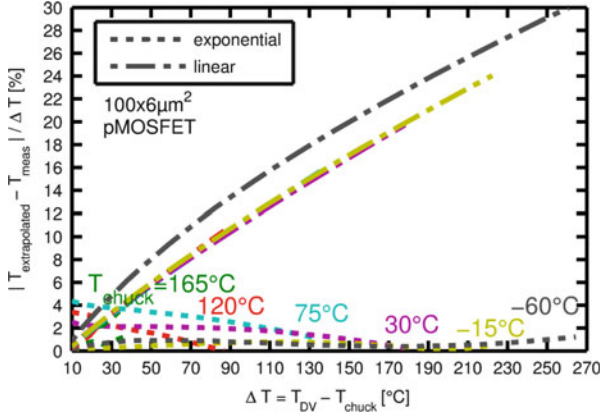


Fig. 2.5 Relative error of a linear extrapolation and the exponential extrapolation (2.3) from the first 10 °C at the respective chuck temperature. The underlying data is that of Fig. 2.3. The relative error for the linear extrapolation increases with increasing power supplied to the poly-heater. In contrast, for the method (2.3) the relative error stays well below a few percent

2.1.4 Heating and Cooling Dynamics

In this subsection we elaborate on the time-dependent heating and cooling dynamics of the device as a heating voltage/power is applied to or removed from the heater, respectively. The chuck (ambient) temperature was -60°C during the following experiments. By applying a certain heating voltage to the poly-heater wires, the device temperature quickly elevates and stabilizes after a couple of seconds. On the other hand, when removing the heating voltage, the device cools down immediately. In order to determine the exact heating voltage necessary to reach a certain device temperature, the heater–device system was calibrated in the way discussed in the previous paragraphs. The output of this initial calibration were eight different heating voltages appropriate to heat the device from -60 to -40 , -20 , 0 , 25 , 50 , 75 , 100 , and 125°C . In the experiments illustrated in Fig. 2.6, different heating voltages were applied and later removed abruptly while recording in parallel the heater current for 100 s. From the heating voltage and current characteristics, the time-dependent power dissipation of the heater was calculated.

As can be seen in Fig. 2.6a, when turning the heater on, it takes up to 1 ms until the maximum power dissipation is reached. This delay time is mainly limited by the finite speed of the voltage source. Using our particular heater design, poly-heater supply voltages up to 34 V are required in order to overcome a temperature range of 185 K ($-60^{\circ}\text{C} \rightarrow 125^{\circ}\text{C}$). After the voltage source has stabilized (>1 ms), the heater power tends to decrease slightly for a couple of seconds. This is because some time is needed to restore thermal equilibrium between heater, wafer, and thermo chuck. The decrease in power within the very first moments after turning on the heater is the greater the larger the temperature difference between heater and chuck. When turning the heater off, the heater power vanishes within approximately 1 ms. Again, this 1 ms is originated in the finite speed of the voltage source.

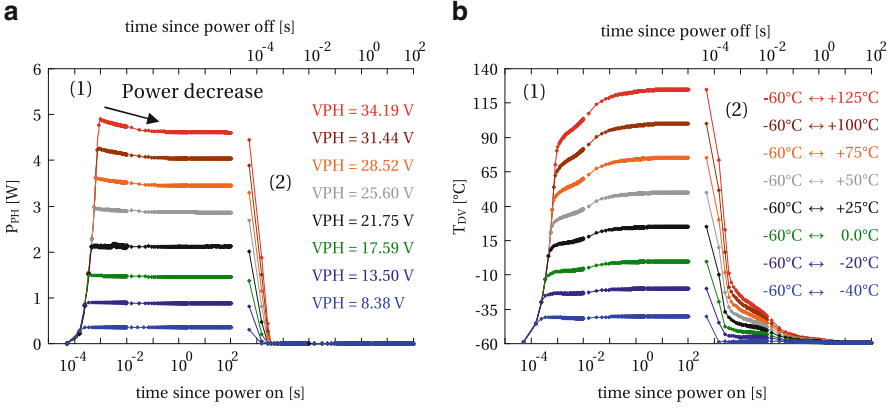


Fig. 2.6 (a) The heating power when turning the heater voltage abruptly on (1) and off (2). The chuck temperature was -60°C . During the heater calibration specific voltages were determined to reach certain device temperatures. At the moment the heating voltage is turned on (1) or off (2), we record the heating current in parallel and calculate P_H . (b) The heating and cooling characteristics of the device when turning the heater power abruptly on (1) or off (2). The heater/device/chuck system needs a couple of seconds to restore thermal equilibrium causing a shoulder in the device temperature at the very beginning of the heating and cooling procedure

In Fig. 2.6b the same experimental sequence as in Fig. 2.6a was performed but this time the drain current of the device was recorded as a representative for the Si/SiO₂ interface temperature. By using the results of Fig. 2.2a one can calculate the evolution of the device temperature T_{DV} from I_D .

As can be seen in Fig. 2.6b, when turning the heater on abruptly, it takes up to 10 s until the device has stabilized at its calibrated target temperature. The larger the temperature difference, the longer it takes to reach the target temperature. The larger time delay is due to the finite time interval necessary to restore thermal equilibrium a distance away from the actual heating source. The shoulder visible in the evolution of T_{DV} during heating is due to the power decrease illustrated in Fig. 2.6a. We remark that although heater power and heater temperature reach a maximum 1 ms after turning the heating voltage on, the device temperature does never exceed its target value due to the delayed thermal coupling between the poly-heater and the device. This is an important aspect since we do not want to subject the device to an elevated pre-stress at the moment the heater is turned on.

When turning the heater off, the situation is similar as during turn on. It takes a couple of seconds until the excess heat generated by the poly-heater can be removed by the thermo chuck. From Fig. 2.6b, we conclude that at an ambient temperature of -60°C any temperature switch up to $\pm 200\text{K}$ can be executed with high accuracy within a time interval of maximal 10 s. In fact it takes approximately 0.1 s to reach the target temperature by 3%, 1 s to obtain a 1% accuracy and after 10 s the target temperature is adjusted by 0.1% which corresponds to the maximum resolution of the measurement.

2.1.5 Summary of the Poly-heater Features

In the previous subsections the features and performance of the in situ poly-heater measurement technique were discussed. The temperature calibration procedure for determining the poly temperature and device temperature was elaborated in detail for different ambient temperatures and power supplies. The thermal resistances of the poly-heater and the device were found to depend on the ambient temperature which is consistent with the nonlinear thermal conductivity reported for silicon. It was shown that a temperature range of more than 200 K can be bridged by additional power supply provided by the poly-heater. In particular, the ability of reaching device temperatures far beyond conventional thermo chuck ranges was pointed out. A thorough study on the heating and cooling dynamics of the device has revealed that a maximum time of 10 s is needed to switch the temperature within an interval of ± 200 K with a maximum precision of 0.1 K. Thereby, the heating and cooling procedure was found to be nearly independent of the difference between ambient temperature (chuck temperature) and target temperature. Equipped with these features the poly-heater tool exhibits a remarkable and unique tool for device reliability testing and characterization purposes which will be applied in the following sections for **NBTI** investigations.

2.2 The Principle of Degradation Quenching and Its Application for BTI Investigations

As demonstrated in the previous section, poly-heaters can be used to perform fast and reliable in situ heating on a single device on wafer level. The following section explains how such a feature can be used to perform **NBTI** stress at a certain stress temperature, which generates a certain degradation level, while the recovery itself can be studied at arbitrary recovery temperatures. By turning the heater on during stress and switching it off during recovery, the tool enables us (a) on the one hand to bring identically processed devices to the same degradation level and (b) on the other hand to fix a different temperature or vary the temperature in a defined way during recovery. By using this technique, our understanding of the recovery physics can be probed in a novel manner. Until now degradation and recovery mechanisms as well as the knowledge about relative contributions to the total threshold voltage shift and recovery are still controversial points in literature which require an unambiguous clarification in order to probe and formulate reliable degradation and recovery models [14, 15].

2.2.1 Why Temperature Quenching?

A trivial problem encountered in the observation of temperature effects in **NBTI stress** is the need to dispose of a set of (i) comparable devices that are brought

to (ii) different degradation levels by different stress temperatures, but which are then (iii) characterized directly post stress at the same unique characterization temperature. In contrast, when temperature effects in NBTI *recovery* are studied, the following problems are encountered: one has to dispose a set of (i) comparable devices that are brought to (ii) the same degradation level by a unique stress temperature, but which are then (iii) characterized directly post stress at different characterization/recovery temperatures.

The first condition (comparable devices) may be solved by careful sample selection, involving thorough characterization before stress. Because stress and recovery dynamics are strongly temperature dependent, the second condition cannot be solved by classically available stress and characterization methods, if the third condition has to be maintained. Classical methods are either based on performing stress and characterization at the same temperature, or strictly separate stress and recovery by long, scarcely observable and basically undefined transition periods. While the first approach can provide the observation of recovery at very good time resolution [16], it obviously always violates condition (ii) if condition (iii) is fulfilled and vice versa.

To fulfill all three conditions at once, it is necessary to conserve the degradation level during cooling. Therefore, the temperature switch has to be fast, well controlled, and practically independent of the difference between stress temperature and recovery/characterization temperature. This demand cannot be fulfilled by a conventional thermo chuck system since the cooling duration of such systems is typically very long (>30 min) and strongly dependent on the target temperature. As a consequence, one has to deal either with *additional degradation* when maintaining the stress bias applied during cooling or with *uncontrolled recovery* when leaving the device floating during cooling. Also, contact difficulties arise due to thermal expansion of probe needles and metal pads which make it hard to maintain device biases during the temperature switch. In short, using the thermo chuck for temperature switches suffers from several systematic errors and drawbacks.

Our approach to harmonize all conditions and to get rid of the above-described technical difficulties is to make use of the poly-heater technique, cf. Fig. 2.7a. During stress a certain stress bias (V_{GS}) is applied to the gate and the (previously calibrated) heater generates an elevated device temperature (T_S) for a defined stress time t_S . Before initiating the recovery/characterization cycle, the heater is switched off, the device reaching ambient (chuck) temperature within a couple of seconds (t_D). In order to prevent any relaxation during the temperature switch, the stress bias remains applied within the delay time t_D . During t_D stress continues in an undefined way, but this additional degradation is negligible compared to the degradation occurring within the main stress period typically performed at a much higher stress temperature ($T_S \gg T_R$). This was verified in Fig. 2.7b where we have stressed different PMOS devices for 1,000 s at $T_S = 125^\circ\text{C}$ and $E_{OX} = 6.0$ MV/cm and afterwards let them recover at $T_R = -60^\circ\text{C}$ and V_{TH} . The cooling delay time t_D between turning off the heater and switching the gate bias from V_{GS} to V_{TH} was varied between 0 and 1,000 s. When using a delay time between 0 and 1 s, the device has not reached the target temperature T_R at the moment the stress bias is removed, cf. Fig. 2.6. Consequently, due to the larger temperature the measured

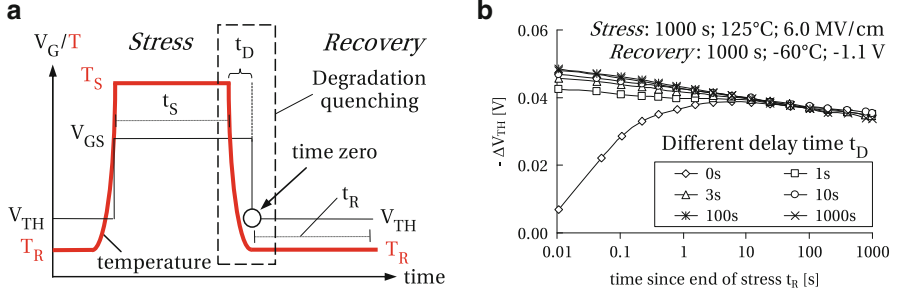


Fig. 2.7 (a) The principle of degradation quenching. Subsequently to the initial characterization phase at the analyzing temperature T_R and the gate bias V_{TH} , the heater is turned on, elevating the device temperature quickly toward the stress temperature T_S . Once at T_S , the stress phase is initiated by switching the gate bias from V_{TH} to V_{GS} . After the stress time t_S has elapsed, the heater is turned off. The stress bias remains applied for a delay time t_D until the device is at T_R (degradation quenching). The recovery cycle (t_R) is then initiated by switching the gate bias from V_{GS} to V_{TH} . (b) Threshold voltage shift as a function of the delay time t_D . Different PMOS devices were stressed for 1,000 s at $T_S = 125^\circ\text{C}$ and $E_{OX} = 6.0 \text{ MV/cm}$. After degradation quenching, the V_{TH} shift was monitored at $T_R = -60^\circ\text{C}$ and V_{TH} using different delay times t_D

V_{TH} shift is afflicted with an error affecting predominantly the first couple of seconds after removal of the stress bias. However, when using a delay time $\geq 3 \text{ s}$, we obtain similar recovery characteristics for arbitrary delay times suggesting (i) that 3 s is sufficiently enough to reach the target temperature and (ii) that we can safely neglect additional degradation or recovery during t_D provided the stress temperature exceeds the recovery temperature by far. Since the cooling time of the poly-heater–device system is nearly independent of the temperature difference ($T_S - T_R$), statement (ii) and (iii) are fulfilled simultaneously independent of the stress or recovery/characterization temperature provided T_R is significantly lower than T_S .

In the following the conservation of the degradation level during the temperature switch will be called “degradation quenching.” Degradation quenching after **Negative bias temperature stress (NBTS)** allows to switch the device temperature *first* from its stress level to its recovery level and *then* triggers V_{TH} recovery by switching the gate bias from the stress level to the threshold voltage of the device. Having demonstrated that degradation quenching can be achieved by using the poly-heater technique, we use the method to investigate the role of temperature in **NBTI** recovery.

2.2.2 The Temperature Dependence of BTI Recovery

To investigate the temperature dependence of **NBTI** recovery, the following experimental procedure was performed on different PMOS devices by making use of the previously described degradation quenching method. During stress, the heater generates a defined interface temperature of 125°C and a certain stress bias is

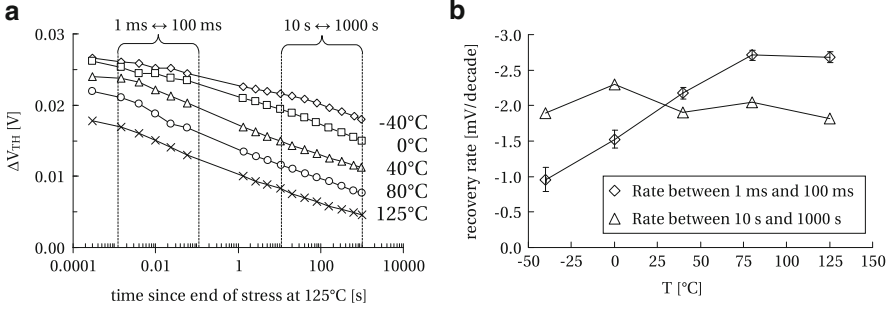


Fig. 2.8 (a) Recovery of the threshold voltage shift recorded at different temperatures after stressing all samples at $E_{OX} = 5.5 \text{ MV/cm}$ and $T_S = 125^\circ\text{C}$. Recovery conditions: $V_{GR} = -1.1 \text{ V}$; $V_{DR} = -2.5 \text{ V}$; $T_R = -40, 0, 40, 80$, and 125°C . (b) Temperature-dependent recovery rate between 1 and 100 ms (diamonds) and between 10 and 1,000 s (triangles)

applied to the gate, subjecting different devices to an oxide field of approximately $E_{OX} = 5.5 \text{ MV/cm}$. During stress, source and drain were at 0 V. Stress field, time, and temperature were identical for all samples, creating a unique degradation level of each device at the end of the stress time t_S which was 1,000 s.

After the 1,000 s had elapsed, the heating current was taken away and the device cooled down rapidly toward the individual ambient temperature which was defined by the temperature of the underlying thermo chuck. When intending to study recovery at -40°C , the chuck has to be at that temperature already before stress. Naturally, the lower the base temperature of the thermo chuck, the greater the required power supply for the poly-heater to reach the unique stress temperature of 125°C , the lower the base temperature of the thermo chuck. One second after the heater was turned off, the gate bias was switched to the threshold voltage ($V_{GR} = -1.1 \text{ V}$) of the device. While the switch of the device temperature terminates the stress, the switch of the gate bias initiates the recovery cycle (t_R). In parallel to the gate bias switch, the drain bias was set to its read-out value ($V_{DR} = -2.5 \text{ V}$) in order to measure the recovery of the saturation drain current. The transition from stress to read-out bias conditions required approximately $200 \mu\text{s}$ and was limited solely by the speed of the voltage unit. An additional $100 \mu\text{s}$ was needed for the measurement. Thus, the first current value at the individual recovery temperature was recorded about $300 \mu\text{s}$ after removal of the stress voltage. The time-dependent evolution of the saturation drain current was later converted into a stress/recovery induced threshold voltage shift, cf. [17].

The result of this temperature quenched recovery measurement is illustrated in Fig. 2.8a. The unique stress temperature, supplied by the poly-heater, was 125°C . The individual recovery temperatures were $-40, 0, 40, 80$, and 125°C . Stress and recovery durations were 1,000 s respectively. As can be seen in Fig. 2.8a, the recovery curves look quite similar except for a temperature-dependent offset which was already present at the first measurement point recorded $300 \mu\text{s}$ after removal of the stress field. Although the recovery temperature varies by more than 160 K

with respect to the individual analyzing temperatures, there is no significant long-term temperature dependence visible in the recovery slopes. The recovery traces are nearly parallel.

In Fig. 2.8b the recovery rate per decade was evaluated more precisely for the first 100 ms and for the last two decades of the recovery traces. On closer inspection, there is a temperature dependence visible within the first 100 ms right after stress. We observe an increasing slope of the recovery curves with increasing temperatures. While at -40°C the amount of recovery is only 1 mV per decade, it is about three times larger for temperatures above 80°C . A reason for the initial temperature dependence might be the fact that the device was probably not exactly at the target temperature due to a slightly too short cooling delay time (t_D) of only 1 s, cf. Fig. 2.7b. A few seconds after the termination of stress all traces become nearly parallel independent of T_R . After 10 s the decrease of the threshold voltage shift has leveled off to about 2 mV/dec for all samples. This holds at least for two or three decades in time.

The offset can be explained qualitatively by assuming an inelastic tunneling process and a homogeneous distribution of trap energy levels which are responsible for the observed log-like recovery traces. In such a model, lowering the analyzing temperature would increase all recovery time constants simultaneously, thereby shifting the entire recovery curve to larger times [18].

It has to be mentioned that a possibly remaining small offset could be also explained by the temperature-dependent position of the Fermi level at the read out gate bias $V_{GR} = -1.1\text{ V}$. At low temperatures, the Fermi level is pinned close to the valence band edge. When increasing the temperature (at constant gate bias), the Fermi level moves a little bit closer toward midgap. Considering creation of interface states within the silicon bandgap as a result of NBTS, their charge state (occupation probability) would be governed by the position of the Fermi level. Consequently, at lower temperatures more of them tend to be positively charged causing a slightly larger threshold voltage shift. This temperature-dependent variation of the Fermi level is a systematic error which is, however, believed to be much too small to explain the full offset.

In summary, the obtained recovery characteristics are quite surprising. A crucial point here is the temperature independence of the recovery rate (slope of the recovery curves in a semi-logarithmic plot) which challenges recovery models based on hydrogen diffusion. For instance, in dispersive diffusion models hydrogen is believed to be stored in traps within the oxide, at the interface or somewhere else close to the interface. In order to re-passivate stress-induced damage during recovery, the H atoms or H_2 molecules have to be released from there and overcome a thermodynamic barrier. At lower temperatures the probability of release as well as the diffusion rate of hydrogen is much lower. If such a mechanism would be the controlling process, one would expect freezing of recovery at -40°C . However, time-dependent recovery is still observed at similar rates even at temperatures as low as -40°C . The same argument also holds for the switching of hydrogen from a bonding to an antibonding Si-H configuration, as proposed by Tuttle [19]. Since the transfer from a bonding to an antibonding configuration is a thermodynamical process, it should be highly temperature activated.

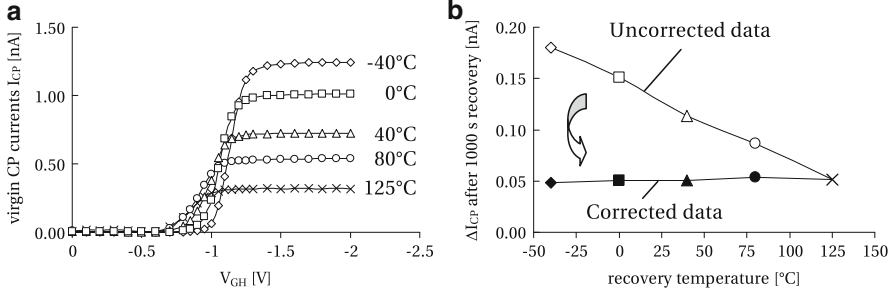


Fig. 2.9 (a) Virgin CP current characteristics recorded at the individual recovery temperatures. The lower the temperature, the higher the CP signal due to a larger profiled active energy range (ΔE_{CP}). (b) The remaining CP current degradation at the end of the constant bias recovery phases performed at different temperatures. CP setup: $V_{GB} = 1.0$ V; $V_{GH} = -2.0$ V; $f = 500$ kHz; $t_r = t_f = 375$ ns. The remaining CP signal is the larger the lower the temperature, however, when accounting for the temperature-dependent active energy interval (ΔE_{CP}), the obtained differences in the uncorrected ΔI_{CP} data (*open symbols*) is removed. The corrected data (*full symbols*) reveals a similar remaining degradation level of the interface after 1,000 s constant bias recovery at different temperatures

In addition to ΔV_{TH} shifts, we have investigated the role of interface states in the recovery process. Therefore, we have performed CP measurements at the end of the 1,000 s lasting constant bias recovery phases. Considering that the obtained ΔV_{TH} shifts are considerably different at the end of each recovery phase, one would expect a similar difference in the remaining CP current if interface state re-passivation would be the dominating recovery mechanism. Such a difference is actually obtained at the end of the recovery, cf. uncorrected data in Fig. 2.9b. The -40°C data shows a considerably larger ΔI_{CP} than for example the 125°C data. However, when taking the temperature dependence of CP into account, the differences in the maximum CP currents at the end of the recovery vanish, cf. corrected data in Fig. 2.9b. We have corrected the original data for different analyzing temperatures by referencing to the initial offsets in the CP currents of the unstressed devices, cf. Fig. 2.9a. We remark that all samples showed a similar CP current before stress, when recorded at the same temperature. This result indicates that either no interface state recovery takes place at all, or interface state recovery is independent of temperature.

2.2.3 Identically Stressed Devices Subjected to Abrupt Temperature Switches

In Fig. 2.10a the V_{TH} recovery is illustrated for four different PMOS devices. All devices were stressed at an oxide field of 5.5 MV/cm and at a temperature of 125°C for 1,000 s. Three reference devices recovered at a constant temperature of -40 , 40 , and 125°C , respectively. The fourth device was subjected to two abrupt

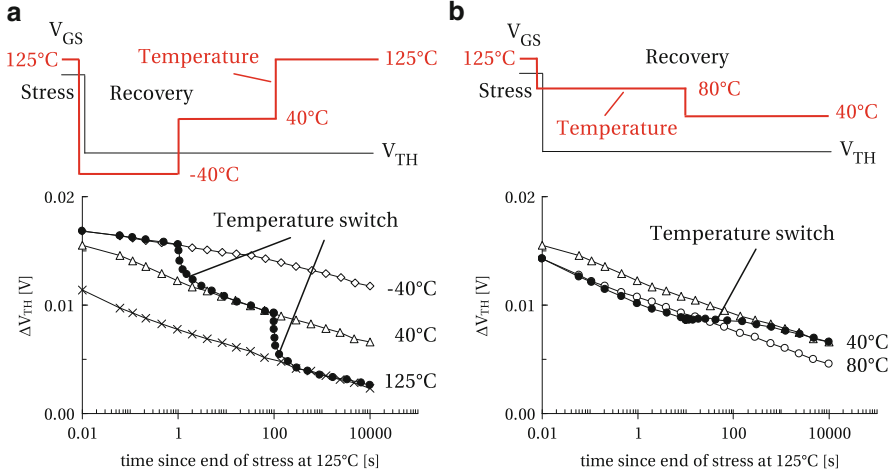


Fig. 2.10 (a) Two step heating performed during constant bias recovery. Reference measurements at -40 , 40 , and 125°C are illustrated by *open diamonds/triangles/crosses*. Recovery can be accelerated twice as we heat the device abruptly from -40 to 40°C (after 3 s) and from 40 to 125°C (after 100 s), cf. *full symbols*. (b) Cooling performed during constant bias recovery. Reference measurements at $40/80^\circ\text{C}$ are illustrated by *open triangles/circles*. Lowering the recovery temperature ($80^\circ\text{C} \rightarrow 40^\circ\text{C}$) leads to frozen recovery until the cooled (*full symbols*) reaches the reference curve at 40°C

temperature switches by making use of the poly-heater technique. Right after stress it recovered for 1 s at -40°C , then for 100 s at 40°C , and finally for another 10,000 s at 125°C providing two decades of inspection at each temperature. As can be seen in Fig. 2.10a, when elevating the device temperature abruptly during recovery, ΔV_{TH} relaxation is immediately accelerated approaching the reference curves after a couple of seconds. Such temperature accelerated recovery at constant gate bias conditions can definitely not be ascribed to elastic tunneling. In Fig. 2.10b we have performed the complementary experiment to Fig. 2.10a: at first, the device recovered at 80°C for 10 s. Afterwards, the heater power was lowered so that the device cooled down to 40°C . While heating accelerates recovery, it can be seen that cooling leads to frozen recovery for a certain interval of time. Indeed, recovery does not proceed before the cooled measurement curve reaches the 40°C reference curve. Again, frozen recovery at constant gate bias conditions cannot be ascribed by an elastic hole trapping model.

2.2.4 Conclusions

Based on the upper key experiments performed on identically stressed PMOS devices, one may draw the following conclusions on the temperature dependence of ΔV_{TH} and CP current recovery:

1. Threshold voltage recovery is accelerated considerably by elevating the temperature. On the other hand, when decreasing the temperature during recovery, the degradation level remains frozen for a certain interval of time.
2. The mechanism causing ΔV_{TH} recovery at elevated temperature is a true chemical relaxation process which is not reversible by subsequent device cooling.
3. CP current recovery is only marginally influenced by either heating or cooling, suggesting interface state repassivation to play only a minor role in the recovery as long as the gate bias is maintained constant around the V_{TH} .
4. The recovery rates of the V_{TH} shift and the CP current (recorded under continuous gate pulsing conditions) are widely independent of temperature. This holds at least for long-term recovery measurements recorded between 1 and 1,000s after the termination of stress.

Since NBTI shows both bias and temperature dependence, but our measurements support neither elastic tunneling nor interface state repassivation, a different mechanism has to be responsible for the observed recovery characteristics. Because bias dependence is totally incompatible with a diffusion process of neutral hydrogen species, we take the dependence of the V_{TH} recovery on the read-out voltage as an indication for a trapping/detrapping phenomenon and attempt to expand the idea of elastic carrier exchange to a temperature sensitive model. As opposed to elastic tunneling, inelastic phonon-assisted tunneling is temperature dependent [20]. Oxide defects and valence band electrons having different energetic positions cannot exchange carriers elastically. However, if they gain energy from lattice vibration (through phonons), they may pass the thermodynamical tunneling barrier ΔE_B with a certain temperature-dependent probability [21].

The lifetime of a single trap can be expressed by an Arrhenius law:

$$\tau(\Delta E_B, T_R) = \tau_0 \exp\left(\frac{\Delta E_B}{k_B T_R}\right), \quad (2.5)$$

where $\tau(\Delta E_B, T_R)$ is the inelastic tunneling lifetime of a single trap, τ_0 is the pseudo-elastic tunneling exchange time between a trap and a substrate carrier at a barrier height zero, ΔE_B is the thermodynamical tunneling barrier, k_B is the Boltzmann constant, and T_R is the analyzing temperature. At a constant temperature the time constants of different traps are solely determined by their individual barrier heights ΔE_B .

When assuming NBTI recovery to be mainly determined by the neutralization of positive oxide defects via electron capture from the silicon substrate (respectively hole emission into the silicon substrate), the observed threshold voltage recovery can be interpreted as a continuous decay of traps with different barrier heights ΔE_B .

Based on this idea, we can schematically illustrate the ΔV_{TH} recovery curve for three different traps having different thermodynamical barrier heights, cf. Fig. 2.11: According to their individual barrier heights, each trap has a certain characteristic

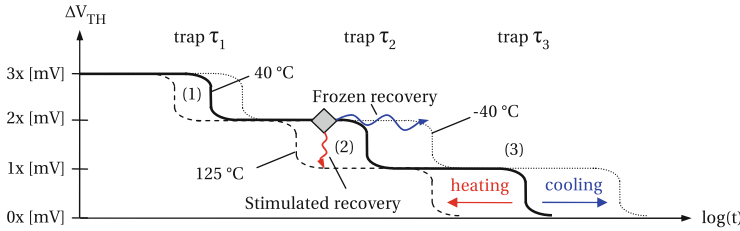


Fig. 2.11 First order model of temperature-dependent recovery effects using a simple picture of three different traps having three different time constants, barrier heights, respectively. Each trap is assigned to an arbitrary threshold voltage shift of $1 \times \text{mV}$. Heating or cooling shifts the recovery curve to the left or the right (shorter or longer time constants) resulting in stimulated or frozen recovery. The *diamond* indicates a hypothetical temperature switching event

time constant at which it recovers with maximum probability. A variation of temperature ($T_{R1} \rightarrow T_{R2}$) impacts all time constants in parallel thereby shifting the plateaus along the time axis in log scale.

The respective shift in time for a trap with barrier height ΔE_B can be calculated as

$$\log(\tau(\Delta E_B, T_{R1})) - \log(\tau(\Delta E_B, T_{R2})) = \frac{\Delta E_B}{k_B T_{R1}} - \frac{\Delta E_B}{k_B T_{R2}}. \quad (2.6)$$

For $T_{R2} > T_{R1}$, the plateaus will shift to the left since the time constants of all traps will decrease, for $T_{R2} < T_{R1}$ the time constants increase leading to a shift to the right. Note that the widths of the plateaus are proportional to ΔE_B . The trap level which recovers first (τ_1) has the lowest barrier (ΔE_{B1}) and is therefore least temperature dependent. On the other hand, trap levels with higher barriers (ΔE_{B2} and ΔE_{B3}) depend stronger on temperature which results in a more significant temperature impact on the plateau broadness.

In this first order model, heating or cooling the device during recovery leads to stimulated recovery (at the diamond, stepping from the solid to the dashed line in Fig. 2.11, cf. Fig. 2.10a) or frozen recovery (at the diamond, stepping from the solid to the dotted line in Fig. 2.11, cf. Fig. 2.10b) compatible with our measurement results. When further assuming that the barrier ΔE_B itself can be lowered by a bias change, the model covers also bias change experiments and includes “mathematically” elastic tunneling in the limit $\Delta E_B = 0$. Furthermore, homogeneously distributed thermodynamical barriers would lead to a large variety of time constants resulting in a large number of small steps like the ones described in Fig. 2.11. In a realistic experiment (large device), one would therefore expect large amounts of small steps to be smeared out as a straight line in a $\log(t)$ diagram consistent with our recovery experiments.

Acknowledgements This work was jointly funded by the Austrian Research Promotion Agency (FFG, Project No. 831163) and the Carinthian Economic Promotion Fund (KWF, contract KWF-1521|22741|34186).

References

1. W. Muth, W. Walter, in *Proc.ESSDERC* (2007), pp. 1251–1262
2. C. Schluender, R.P. Vollertsen, W. Gustin, H. Reisinger, in *Proc.ESSDERC* (2007), pp. 131–134
3. T.K. Kang, C.S. Wang, K.C. Su, *Jpn.J.Appl.Phys.* **46**, 7639 (2007)
4. C.S. Wang, W.C. Chang, W.S. Ke, C.T. Chiang, C.F. Lee, K.C. Su, in *Proc.SSDM* (2005), pp. 580–581
5. C.S. Wang, W.C. Chang, W.S. Ke, K.C. Su, in *Proc.IIRW* (2006), pp. 136–138
6. H. Köck, V. Košel, C. Djelassi, M. Glavanovics, D. Pogany, *Microelectron.Reliab.* **49**, 1132 (2009)
7. A. Kelleha, W. Lane, *IEEE Trans.Nucl.Sci.* **43**, 997 (1996)
8. W. Liu, M. Asheghi, *J. Appl. Phys.* **98**, 123523 (2005)
9. A. Cardoso, A.K. Srivastava, *J. Vac. Sci. Tech. B* **19**, 397 (2001)
10. H. Ibele, K. Reitingner, in *IEEE Semiconductor Wafer Test Workshop* (2005)
11. P. Leturcq, J.M. Dorkel, A. Napieralski, E. Lachiver, *Trans. Elec. Dev.* **34**, 1147 (1987)
12. C.J. Glassbrenner, G.A. Slack, *Phys. Rev.* **134**, A1058 (1964)
13. G.A. Slack, *J. Appl. Phys.* **35**, 3460 (1964)
14. T. Aichinger, M. Nelhiebel, T. Grasser, in *Proc.ESREF* (2008), pp. 1178–1184
15. T. Aichinger, M. Nelhiebel, T. Grasser, in *Proc.IRPS* (2009), pp. 2–7
16. H. Reisinger, O. Blank, W. Heinrigs, A. Mühlhoff, W. Gustin, C. Schlünder, in *Proc.IRPS* (2006), pp. 448–453
17. B. Kaczer, V. Arkhipov, R. Degraeve, N. Collaert, G. Groeseneken, M. Goodwin, in *Proc.IRPS* (2005), pp. 381–387
18. G. Pobegen, T. Aichinger, M. Nelhiebel, T. Grasser, in *IEDM Tech. Dig.* (2011), pp. 27.3.1–27.3.4
19. B. Tuttle, *Phys.Rev.B* **59**, 12884 (1999)
20. S. Ganichev, E. Ziemann, W. Prettl, I. Yassievich, A. Istratov, E. Weber, *Phys.Rev.B* **61**, 61 (2000)
21. W. Göss, M. Karner, S. Tyaginov, P. Hehenberger, T. Grasser, in *Proc.SISPAD* (2008), pp. 69–72

Bias Temperature Instability for Devices and Circuits

Grasser, T. (Ed.)

2014, XI, 810 p. 601 illus., 318 illus. in color., Hardcover

ISBN: 978-1-4614-7908-6