Microelectronics Reliability 54 (2014) 22–29

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/microrel

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

Filippo Alagi *

STMicroelectronics, Via Tolomeo, 1, 20010 Cornaredo, MI, Italy

ARTICLE INFO

Article history:
Received 22 March 2013
Received in revised form 12 July 2013
Accepted 12 July 2013
Available online 21 August 2013

ABSTRACT

Experiments on a silicon p-channel MOS FET with heated gate electrode have provided evidence that the increase of the interface-trap density observed upon homogeneous negative bias temperature stress can be modelled as the unbalance of two reaction-limited activation and deactivation processes. The variation of the measured trap density with the stress time shows saturating trends, with bias-dependent saturation level. This feature is explained as the dynamic balance of a field-dependent forward reaction and a field-independent backward reaction. The rate constants of both reactions appear to be Arrhenius-like, with distributed activation energy. The activation energy of the forward reaction, which is identified as the depassivation of native Si/SiO_2 (near-) interface defects, turns out to be a nearly square root function of the electric field in the oxide. By analogy with the early Frenkel's theory of electron transport in dielectric materials we discuss whether this field dependence can be considered as the first evidence that the defect depassivation would proceed through a stage where a positively charged hydrogen specie escapes from negatively charged defect.

1. Introduction

One interpretation of the Negative Bias Temperature Instability (NBTI) in silicon MOS FET devices admits the coexistence of two distinct components in the NBT-induced threshold voltage instability. One component would be quickly relaxing upon releasing the 'stressing' condition and the other would relax quite more slowly [1,2]. Even if a complete agreement has not yet been reached about, it seems plausible that the quickly relaxing component should be associated with the capture-emission of semiconductor charge carriers on localized electronic states in the gate dielectrics [3]. On the other hand, it has been pointed out [4,5] that the more slowly relaxing component arises from the variation of the number of localized interface states (traps), which can be related to the activation of pre-existing dangling bond defects at the Si/SiO_2 interface. It has been observed that the carrier capture-emission may exhibit unexpectedly long and widely spread characteristic times, and it is known since long time [2,6–8] that the NBT-activated interface traps can be annealed-out even at moderate temperature. Hence it appears that a simple recoverable/permanent classification of the two components is misleading, and that NBTI should be better described in terms of (at least) two microscopic mechanisms, both reversible and both characterized by dispersive kinetics.

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

The focus of this work is on the interface-state formation kinetics in p-channel type silicon MOS FET upon NBT stress (NBTS) in the range of low electric field, which is of practical interest for the reliable design of analogic circuits and which has been scarcely explored to date. We will present the experimental results drawn from a device designed on the purpose, consisting of a modified p-channel MOS FET with salicided polycrystalline silicon gate which can be joule-heated at wafer level. The key advantage of localized heating is that the device temperature can be quickly ramped down while the gate-to-source voltage is kept at the desired (stress) value, allowing a drastic reduction of the NBTI recovery effects. We found that the build-up kinetics of the integrated interface states can be modelled as a reaction-limited process, with Arrhenius-like rate constants and distributed activation energy. We will show that the kinetic details of the interface-state build-up can be explained as the unbalance of forward and backward reactions, and that the activation energy of the forward reaction is a nearly square root function of the electric field in the oxide. We believe that the latter result is the first experimental evidence that the depassivation of native Si/SiO_2 interface defects would proceed through a stage where a nearly free positively

* Fax: +39 02 93519 355.
E-mail address: filippo.alagi@st.com

0026-2714/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.microrel.2013.07.045

charged hydrogen specie escapes from the coulombic attraction of the remaining negatively charged defect.

2. Materials and methods

The use of localized joule heating for studying device reliability at wafer level was been reported by Muth and Walter [9], who designed a MOS FET featured with a polycrystalline silicon heater located close to the transistor and capable of rising the device temperature up to 255 °C, starting from room temperature, and by Hamada and Roesch [10], who used a similar approach but with a thin-film resistor heater capable of achieving up to about 420 °C on the device under test. With the device used in the present work, the same gate electrode of a MOS FET serves also as the heating element, allowing reaching a 'junction' temperature of more than 320 °C, starting from room temperature. The device is derived from a p-channel MOS FET formed in phosphorus doped, (100) oriented crystalline silicon, with SiO_2 gate dielectrics (oxide) of thickness t_{ox} = 13 nm and salicided p^+-doped polycrystalline silicon gate. The gate oxide is grown in two steps, steam at 900 °C and wet at 825 °C, and then thermally (NO) nitrided. The device process comprises a final annealing at 400 °C in N_2/H_2, atmospheric pressure, and 6% H_2 partial flow. The nominal total channel width W is 12.5 μm and the nominal channel length L is 5 μm. A controllable heating current can be forced through the salicided p-doped polycrystalline silicon layer, which serves both as the gate electrode and as the heating element. The typical threshold voltage extrapolated from the condition of peak transconductance at V_{DS} = −0.1 V is −0.74 V.

2.1. Device connections

With the connection scheme shown in Fig. 1 and under a PC control, the full measure–stress–measure sequence iteration, including charge-pumping measurement, was completely automated, i.e. no manual action was needed during the test of each device. The experimental setup comprises one High Power Source Measurement Unit (HP SMU), capable of supplying up to ±300 mA heating current at up to ±10 V DC, one Pulse Generator Unit (PGU) for the charge-pumping measurement, two switches (SW1 and SW2), and four medium power Source Generator Units (SMU). SW1 and SW2 are both set in the low position, during both the stress and the measurement of the transfer characteristics, and are both set in the high position, during the charge-pumping measurement. Since the electrical connection to the wafer substrate is granted via the wafer holder (chuck), a minimum number of five bench probes were used.

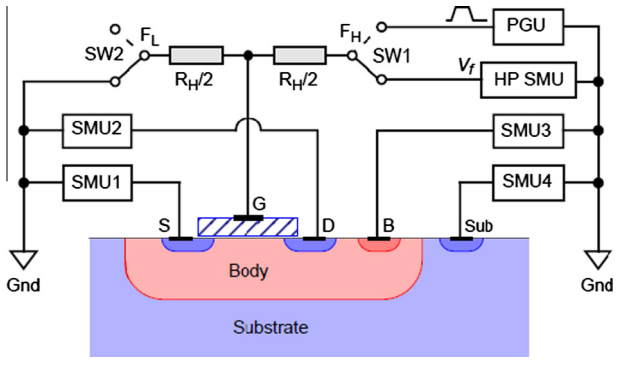

Fig. 1. Schematic electric connections of the heated gate device. The heater/gate current supplied by the high-power Source Measure Unit HP SMU flows also through the gate actually. SW1 and SW2 are remote-controlled switches.

2.2. Stress-measure sequence

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

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

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

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

2.3. Temperature sensing

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