Numerical Confirmation of Inelastic Trap-Assisted Tunneling (ITAT) as SILC Mechanism

Ting-Kuo Kang, Ming-Jer Chen, Senior Member, IEEE, Chuan-Hsi (Jason) Liu, Yih J. Chang, and Shou-Kong Fan

Abstract—This paper presents a quite comprehensive procedure covering both the stress-induced leakage current (SILC) and oxide breakdown, achieved by balancing systematically the modeling and experimental works. The underlying model as quoted in the literature features three key parameters: the tunneling relaxation time \( \tau \), the neutral electron trap density \( N_f \), and the trap energy level \( E_t \). First of all, 7-nm thick oxide MOS devices with wide range oxide areas are thoroughly characterized in terms of the optically induced trap filling, the charge-to-breakdown statistics, the gate voltage developments with the time, and the SILC \( I-V \). The former three are involved together with a percolation oxide breakdown model to build \( N_f \) explicitly as function of the stress electron fluence. Then the overall tunneling probability is calculated, with which a best fitting to SILC \( I-V \) furnishes \( \tau \) of \( 4.0 \times 10^{-13} \) s and \( E_t \) of 3.4 eV. The extracted \( \tau \) is found to match exactly that extrapolated from existing data. Such striking consistencies thereby provide evidence that inelastic trap-assisted tunneling (ITAT) is indeed the SILC mechanism. Differences and similarities of the involved physical parameters between different studies are compared as well.

Index Terms—Flash, gated-diode, inelastic tunneling, MOSFET, oxide breakdown, percolation, SILC, stress-induced leakage current, trap-assisted tunneling.

I. INTRODUCTION

STRESS-induced leakage current (SILC) is one of the biggest reliability issues in MOS devices, especially the nonvolatile flash memory [1]. In nonvolatile flash memory, high-field or Fowler–Nordheim (F–N) tunneling during the erase cycle can produce a variety of defects within tunnel oxides, among which the most concerned for SILC are the neutral electron traps. The principal reasons are that these generated traps can serve as a stepping stone to effectively shorten the tunneling distance, causing more electrons leaking out of the floating gate. Such knowledge of the impact of SILC stems from an early series of studies [2]–[5]. Recent experimental demonstrations in terms of carrier separation technique [6] and oxide thickness dependent measurement [7] both witness the inelastic behavior of trap-assisted tunneling. That is, the tunneling electrons from the cathode side lose part of their energy via inelastic scattering and fall down to the underlying trap site from which they are instantly de-trapped out to the anode side. Despite few literature efforts made to address the scattering/capturing process microscopically, inelastic trap-assisted tunneling (ITAT) proposed as the origin of SILC is currently largely accepted. By analogy to the most efficient generation/recombination center in the mid-gap of a semiconductor p-n junction, there should exist a certain trap position featuring that a local tunneling probability \( P_1 \) from cathode to that site equals another probability \( P_2 \) from it to anode. Under such situation, a maximum overall tunneling probability \( P \) is created through \( P = P_1 P_2/(P_1 + P_2) \) [8]. Such a picture of the maximum likelihood had led to the following analytic model for the SILC current–voltage (\( I-V \)) [8]:

\[
J_{\text{SILC}} = \frac{q t_{\text{ox}} N_f}{2\tau} \exp\left\{ \left( \frac{4}{3} \frac{(2m_{\text{ox}})^{1/2}}{\hbar} \frac{1}{q E_{\text{ox}}} \right) \times \left( E_t^{3/2} - [E_t - q E_{\text{ox}}(t_{\text{ox}} - x_t)]^{3/2} \right) \right\}
\]

where
- \( t_{\text{ox}} \) oxide thickness;
- \( E_{\text{ox}} \) oxide field;
- \( m_{\text{ox}} \) (0.42\( m_0 \) [9]) effective electron mass in the oxide;
- \( \tau \) tunneling relaxation time, \( x_t \approx 3.6 \) nm calculated [8] for 7-nm thick oxide, for example) lies at the most favorable position, and the trap properties at this specific location are usually described by the neutral electron trap density \( N_f \) and the trap energy level \( E_t \).

This unique model had exhibited the comparable ability of tackling SILC \( I-V \) [8] as the complicated version [6], [7]; strictly speaking, however, reported agreement with singly SILC \( I-V \) unnecessarily means that the current understandings of underlying physical aspects (i.e., the maximum likelihood of tunneling to and from the traps located at the most favorable position, etc.) would have gotten clarification in the way. Part of the reasons is that the tunneling relaxation time \( \tau \) was not fully explored in [8]. Here \( \tau \) is better viewed a fundamental parameter and can be defined the tunneling time extrapolated at a zero oxide thickness [9].

This paper presents a quite comprehensive procedure of balancing systematically both the modeling and experimental works from SILC to oxide breakdown. Equation (1) is favored here due to its uniqueness as mentioned above. In contrast to the pioneering work [8] where \( N_f \) was treated as a fitting parameter, however, this work manages it explicitly in advance, achieved by extending to experimental Weibull distributions of charge-to-breakdown as well as with the aid of some percolation oxide breakdown models [10]–[15]. As a result,
the number of the unknown parameters in (1) reduces to two, namely, $\tau$ and $E_t$, whose values can further be accessed by fitting SILC $I-V$ while simultaneously accounting for the maximum overall tunneling probability. Fortunately, there appears in the open literature [9] the experimental tunnel time versus oxide thickness, making possible direct verification on the extracted $\tau$. Should an expected coincidence turn out, it can be acknowledged that current understandings involved in SILC mechanism indeed stand on the ground. Also addressed are differences and similarities of physical parameters as compared with the quotation [5] and [8].

II. EXPERIMENTAL PROCEDURE

A variety of MOS devices were fabricated in the same process. The oxide film was thermally grown in dry oxygen ambient to 7-nm thick. The oxide areas were drawn in a wide range of four decade with aim to judge the present SILC theory that, whatever the areas used are, the traps generated during high field stress are spatially randomly allocated within the whole oxide space. This is valid until a percolation path for breakdown is formed locally. The first sample was n-channel MOSFETs having gate width-to-length ratio of 20 $\mu$m/0.3 $\mu$m (oxide area of $6 \times 10^{-8}$ cm$^2$) to build a linkage between the trap generation density and the stress electron fluence. The stress condition was gate voltage $V_g$ with source, drain, and substrate tied to ground, which was followed by the optically induced trap filling. Fig. 1 schematically shows this optical injection method involving two distinct processes: the photo-generation process via a tungsten lamp to supply electron seed in the substrate; and the carrier heating process via a negative substrate bias of $-3$ V to raise these electrons up to the higher energy level enough to surmount the Si/SiO$_2$ barrier and fill the traps in oxide. Then operating in gated-diode forward mode [16] (i.e., a forward bias of 0.2 V was applied to the drain with source open and substrate grounded) can sensitively detect the filled traps.

The secondary sample was $n^+$-poly/p-substrate MOS capacitors having two very large oxide areas of $1.0 \times 10^{-4}$ and $6.25 \times 10^{-4}$ cm$^2$ for charge-to-breakdown test. The stress condition was constant current of 400 mA/cm$^2$ with positive gate voltage. The third sample was n-channel MOSFETs with oxide area of $5.2 \times 10^{-6}$ cm$^2$ for monitoring the time evolution of gate voltage. The stress condition was constant current of 38.5 mA/cm$^2$ with positive gate voltage and with source, drain, and substrate connected to ground. The stressing was periodically interrupted to measure SILC $I-V$ characteristics until oxide wear-out occurred.

III. QUANTIFYING NEUTRAL TRAPS

From the measured drain current in forward gated-diode mode versus gate voltage for the first sample, it is observed that the gate-voltage shift $\Delta V_G$ associated with the current peak in the depletion region increases with the filling or illumination time and gradually tends to saturate as depicted in Fig. 2. Assuming that the occupied traps are spatially distributed uniformly within the oxide as adopted elsewhere [12], [17], the saturation voltage shift $\Delta V_{G,SAT}$ from the optical filling method can be directly linked to the occupied trap density $N_{ox}$ through $\Delta V_{G,SAT} = \frac{q^2 \tau_{ox} N_{ox}/2 \varepsilon_{ox}}{\varepsilon_{ox}}$ where $\varepsilon_{ox}$ is the oxide permittivity. The resulting $N_{ox}$ in 1/cm$^2$ for different $Q_e$ in C/cm$^2$ is displayed in the inset of Fig. 2, showing a power-law relation

$$N_{ox} = 1.62 \times 10^{18} Q_e^{0.5}. \tag{2}$$

A percolation oxide breakdown model [13] formulates explicitly the generated neutral electron trap density $N_t$ as a function of three physical controlling factors: the possible minimum trap number ($= (t_{ox} - 2t)/2\pi r^2$) [14], the possible minimum area ($= \pi r^2$) [10], [11], [14] of the locally conductive path, and the ultimate thickness limit of 2.5 nm for breakdown [18]. Here, $r$ is the trap radius and $t$ ($= 0.5$ nm [13]-[15]) is the transition layer thickness. $N_t$ can be related directly to $N_{ox}$ via filling fraction $p$: $N_t = N_{ox}/p$. The physical origin behind $p$ is Coulomb repulsion [19]. $p$ can be quantified from a Weibull
plot of charge-to-breakdown $Q_{BD}$ statistical data as shown in Fig. 3 from the secondary sample with two oxide areas. Here $Q_{BD}$ is the $Q_e$ at the onset of oxide breakdown. The Weibull slope $s$ and the modal (63%) $Q_{BD(m)}$ of the charge-to-breakdown distributions can be expressed as [13]

$$s = 0.7821 \times \frac{\tau_{ox} - 24}{r} \times 0.5$$

(3)

$$Q_{BD(m)} = \left(\frac{pC_0}{1.62 \times 10^{18}}\right)^2 \left(\frac{A_0}{A_{ox}}\right)^{1/s}.$$  

(4)

Here, $C_0$, the modal of the $N_t$ distributions at a specific area $A_0$ of 900 nm$^2$, reads [13]

$$C_0 = -1.5 \times 10^{20} + 7.611 \times 10^{19} \left(\frac{t_{ox} - 225}{\tau^2}\right)^{1/3}.$$  

(5)

$A_{ox}$ is the oxide area in nm$^2$. Extracting the values of $s$ (through two fitting lines of equal slope) and $Q_{BD(m)}$ from Fig. 3 and substituting into the above expressions, a unique solution for two different areas yields $p = 4.75\%$ for $r = 0.42$ nm. Note that such a very low occupancy fraction of the total neutral trap density is essential in addressing ITAT. Therefore, we achieve the goal of quantifying in advance the amount of generated neutral electron traps for given stress electron fluence.

IV. PARAMETER EXTRACTION AND COMPARISONS

Fig. 4 shows the evolution of gate voltage for the third sample subject to constant F–N tunneling stressing. It can be seen that gate voltage gradually increases with time until at around 1400 s a large drop down to 2 V occurs, indicating a hard breakdown event. The built power-law relationship between $N_t$ and $Q_e$ reproduces excellently such breakdown event, regardless of areas used. This is achieved by substituting the stress current density of 38.5 mA/cm$^2$ and area of $5.2 \times 10^{-6}$ cm$^2$ into (4) and (5). The resulting time to breakdown is 1332 sec, quite close to the spontaneous point in Fig. 4, as expected by Poisson area scaling [20]. The increment of gate voltage with respect to the initial value is magnified in the inset of Fig. 4 for stress time prior to breakdown, validating the power-law expression of $\Delta V_G \propto \ell^{1/2}$. Obviously, the traps generated are spatially randomly distributed within the whole oxide film, indicating that SILC magnitude obeys a linear relation with oxide area.

The measured SILC $I–V$ before and after stress is displayed in Fig. 5. This figure reveals that the SILC $I–V$ curves prevailing in the low voltage regime are raised up for increasing electron stress fluence, whereas in the high voltage region the $I–V$ is intact, an indicative of F–N tunneling dominating. The latter property facilitates transformation from gate voltage to oxide field $E_{ox}$. F–N tunneling fitting [21] was applied to fresh $I–V$ in Fig. 5 to access oxide field $E_{ox}$. The resulting $I = E_{ox}$ is depicted in Fig. 6. Prior to dealing with SILC $I–V$, a calculation work was carried out to furnish a set of $\tau$ and $E_{th}$ to meet the condition of $P_1 = P_2$ for the maximum overall tunneling probability. Fig. 7 shows such calculation results using formula in [8]. From Fig. 7, a specific set of $\tau = 4.0 \times 10^{-13}$ s and $E_{th} = 3.4$ eV is rigorously selected since they are able to offer a best reproduction of SILC $I–V$ or equivalently $I = E_{ox}$ in Fig. 6. Note that the main role of $E_{th}$ and $\tau$ is to adjust the slope and magnitude of SILC $I–V$, respectively. The same power-law relation between $N_t$ and $Q_e$ is also involved in the way. This explains why we call the procedure of balancing systematically both the modeling and experimental works. This procedure comprises the quantification process for the trap density, the verification process for gate voltage evolution and time-to-breakdown, and the parameter extraction process. Finally, a fundamental parameter of interest is examined. Fig. 8 re-plots experimental tunnel time versus oxide thickness from [9], where a straight line drawn through all data points is extrapolated far to zero oxide thickness (the authors in [9] suggested that the most right side data would be adjusted upward for measurement reasons). As we place our extracted $\tau$ in the figure corresponding to zero oxide thickness,
strikingly it matches exactly the line. Therefore, the expected coincidence turns out and it can be acknowledged unambiguously that current understandings involved in SILC mechanism indeed stand on the ground; for example, ITAT does favor the maximum likelihood of tunneling to and from the traps at the specific position; and the occupancy fraction of $N_T$ is very low, ensuring the possibility of ITAT.

Several physical parameters involved in SILC $I-V$ are also available in the open literature [5], [8], and are quoted here to explore differences and similarities of the involved parameters. Firstly, in the original work [8] pioneering the SILC model (1), $\tau$ was set at $1 \times 10^{-15}$ s without particular reasons, far away from ours by about two orders of magnitude. However, our extracted $E_t$ does not show such huge difference with that (3.6~4.0 eV) in [8]. This is not strange as one can recognize from (1) that the exponent factor and the pre-exponent factor are essentially independent of each other, and only in the pre-exponent factor, $\tau$ and $N_T$ are merged together and are thereby affected each other if extracted simultaneously. Secondly, a sophisticated modeling of SILC $I-V$ curves in [5] reported the trap energy level $= 2.3 \sim 2.4$ eV and the trap cross section $= 10^{-15} \sim 10^{-16}$ cm$^2$. The latter is somewhat comparable with the trap sphere area ($= \pi r^2$) of $5.6 \times 10^{-15}$ cm$^2$ in our work. Only the trap energy level is quite spread, which is likely attributed to different process technologies used. That is, the trap energy level would be considered a measure of the inelastic scattering property and different oxides from different process technologies reflect different scattering behaviors. Such arguments involved in inelastic scattering process certainly need further research.
V. CONCLUSION

A procedure of balancing systematically both the modeling and experimental works covering SILC and oxide breakdown has been comprehensively carried. Involved in this procedure are the quantification process for the trap density, the verification process for gate voltage evolution and time-to-breakdown, and the parameter extraction process for SILC I–V. Eventually the tunneling relaxation time viewed as a fundamental parameter has been exactly confirmed by existing data, thus clarifying the tunneling relaxation time as a fundamental parameter for SILC mechanism:

1) ITAT does favor the maximum likelihood of tunneling and from the traps at the specific position;
2) the occupancy fraction of the total generated neutral density is very low ensuring the possibility of ITAT.

Differences and similarities of physical parameters between different studies have been compared as well, suggesting that the trap energy level would be considered a measure of the inelastic scattering property and is process technology dependent.

REFERENCES


Shou-Kong Fan received the Sc.D. degree in materials science from the Massachusetts Institute of Technology, Cambridge, in 1987.

From 1986 to 1988, he was a Process Engineer at the Semiconductor Sector, Texas Instruments (TI). From 1988 to 1994, he was Member of Technical Staff, TI’s Central Research Lab, working on III-V based HBT microwave device. He also served as an Adjunct Professor at Southern Methodist University, Dallas, TX, from 1989 to 1990, teaching senior level electrodynamics. In 1994, he joined TI-Acer as Project Manager to work on a 16M DRAM design. Later he led a device design group to convert the 16M DRAM process to 0.5 μm logic process for foundry service using TCAD tool. From 1998 to 2000, he first worked as Deputy Director on TCAD and later on reliability as Specialist at WSMC. Since June 2000, he is with UMC as the Senior Reliability Assurance Department Manager. His responsibilities include the execution or implementation of reliability assessment, process/product qualification and the wafer-level reliability monitoring scheme among all UMC fabs.