Pergamon

0038-1101(94)00245-2

Solid-State Electronics Vol. 38, No. 7, pp. 1373–1379, 1995

0038-1101/95 $9.50 + 0.00

A SIMPLE AND ACCURATE SIMULATION TECHNIQUE FOR FLASH EEPROM WRITING AND ITS RELIABILITY ISSUE

KUEI-SHAN WEN and CHING-YUAN WU

Advanced Semiconductor Device Research Laboratory and Institute of Electronics, National Chiao-Tung University, 1001 Ta-Hsueh Road, Hsinchu, Taiwan, R.O. China

(Received 5 August 1994; in revised form 21 October 1994)

Abstract—An efficient and accurate 2D analysis for gate-current is proposed for short channel n-MOSFETs, in which the channel hot-electron-enhanced injection probability is proposed and expressed in terms of the actual current path and its power density flow. The accuracy of our gate-current analysis has been verified by comparisons between simulation and experimental data. This well-established gate current analysis as well as the charge boundary condition on the floating gate have been implemented into the *su*b-*m*icron *MOS* (SUMMOS) two-dimensional device simulator for characterizing n-channel flash EEPROM writing. Comparisons with experimental EEPROM writing have been made, and quite good agreements have been obtained for test devices with different channel lengths ranging from 0.8 to 0.5 μm for wide range of applied biases. Moreover, computer simulation for EEPROM reliability issue caused by oxide electron traps has also been performed to characterize the endurance of flash EEPROM operation.

1. INTRODUCTION

Flash EEPROMs have been a prevalent category of submicrometer devices because they achieve a smaller cell size than conventional EEPROMs by having the block-ERASE capability[1–4]. The ETOX-based flash memory is programmed by charging the floating gate with channel hot-electrons injecting over the energy barrier at the Si/SiO_2 interface, and can be erased by a high drain- or source-voltage through Fowler–Nordheim tunneling current. Although device designs for high speed and reliable flash EEPROMs have been published by considering different cell structures or processing conditions[3–7], an accurate design tool based on either a 2D device simulator or analytic modeling is required to shorten the optimization design cycle.

A widespread simulation approach based on the non-Maxwellian form of electron energy distribution (EED) had been proposed to characterize the 2D hot-carrier injection problems[8–11]. Recently, Fiegna *et al.*[10] have used a conventional drift–diffusion (PISCES II) simulator coupled with a post-processor to calculate the gate-current and this method has been applied to the EEPROM writing[10]. With the electric field and the carrier density calculated from the solution of PISCES II, the electron temperature and the effective field concerning the non-local effect are obtained. The effective field is then fed to the empirical carrier distribution function calibrated by the results of Monte Carlo method in homogeneous high energy region[12]. However, there are many assumptions and fitting parameters in their formulation procedure, for example: the empirically fitted distribution function aforementioned and the power-law approximation for the non-parabolic band. With such a complicated post-processing method, the simulation results in the high field region, however, are overestimated (see Fig. 5 in [10]).

It is known that the injection probability is proportional to the integral function of hot-carrier distribution and density of states. Since this integral is insensitive to the detailed form of these functions, it is possible to treat the hot-carrier enhanced injection problems by an effective term. In this paper, a macroscopic approach is proposed to formulate a quasi-2D hot-carrier injection model by considering the effects of channel hot-carrier enhanced Si/SiO_2 interface barrier lowering. Calculation of the barrier height can be directly obtained from the available physical quantities deduced from a 2D MOS simulator-SUMMOS[13] without another post-processing steps. Only one fitting parameter is introduced to characterize this enhanced barrier lowering effects and this parameter is shown to be a universal constant for short-channel n-MOSFETs with different channel lengths for wide ranges of applied biases. The developed gate-current model and charge boundary condition on the floating gate have been implemented into SUMMOS simulator. Its accuracy has been well verified by comparing both the experimental MOSFET gate-current and the writing characteristics of EEPROMs with different channel

lengths ranging from 0.5 to 1.0 μm for wide ranges of external biases. Moreover, the EEPROM reliability issue caused by trapped electrons in the oxide during hot-carrier injection through oxide layer is also considered. The roles of total oxide trap density, trapping rate, applied biases, and the write/erase cycle on the degradation of EEPROM writing have been simulated. Therefore, our developed simulation technique can be used to perform device optimization design and lifetime prediction for flash EEPROM memory devices.

2. THE SIMULATOR

2.1. Models embedded in SUMMOS

A macroscopic impact ionization model considering the non-homogeneous electric field and surface scattering effects have been implemented into SUMMOS[14]. The accuracy of our new developed impact-ionization model has been verified by comparisons between experimental substrate current and simulation results of test devices fabricated by different technologies with different channel lengths (down to 0.36 μm) for wide ranges of applied external biases (drain, gate and substrate biases)[14]. For the gate-current modeling, a 2D channel hot-carrier enhanced injection probability is expressed as[15]:

$$P = A \exp\left(-\frac{\Phi_{B,2D}}{qE_{\perp} \cdot \lambda}\right), \tag{1}$$

where A is a normalization constant, λ is the mean-free-path of hot electron injection and has been determined experimentally to be about 91 Å[16] and $E_{\perp}$ is the vertical surface electric field in the substrate. The 2D effective barrier height considering an additional barrier lowering term (Φ_{BL}) due to channel hot-carriers is expressed by:

$$\Phi_{B,2D} = \Phi_{B0} - \alpha E_{ox}^{1/2} - \beta E_{ox}^{2/3} - \Phi_{BL}, \tag{2}$$

where $\Phi_{B0} = 3.1$ qV is the Si/SiO_2 interface barrier height; E_{ox} is the oxide electric field; $\alpha = 2.59 \times 10^{-4}$ q (V cm)$^{1/2}$ is the coefficient of barrier lowering due to the image force; and $\beta = 1.0 \times 10^{-5}$ q (V cm^2)$^{1/3}$ is the coefficient of barrier lowering due to Fowler–Nordheim tunneling[16]. The barrier lowering term Φ_{BL} is simply expressed as:

$$\Phi_{BL} = \gamma \frac{J \cdot E}{|J|}, \tag{3}$$

where J is the hot-electron current density vector; E is the electric field vector and γ is a proportional factor related to the average mean-free-path of channel hot carriers and is shown to be constant.

Physically, $J \cdot E$ in eqn (3) means the power density of hot electrons gained from the local electric field along the current flow path in the substrate. Let $\gamma \simeq q v_{sl} \langle \tau \rangle$ and $|J| = n q v_{sl}$ in the saturation region, where v_{sl} is the saturation velocity; n is the local density of hot electrons; $\langle \tau \rangle$ is the average energy relaxation time for hot electrons, Φ_{BL} can be simply expressed as $\Phi_{BL} = \langle \tau \rangle J \cdot E/n$. Therefore, Φ_{BL} represents the average energy gained by each hot-electron and is equivalent to the average barrier lowering of each channel hot electron looking from the substrate. This new developed injection probability has been verified by the experimental gate-currents of conventional n-MOSFETs with different oxide thicknesses and effective channel lengths for wide ranges of applied biases[15]. In addition to the gate injection probability, the kinetics of the Si/SiO_2 interface-trap generation/occupation due to hot-electron injection at high drain- and gate-biases have also included in SUMMOS. The simulated spatial distribution of the generated Si/SiO_2 interface traps has been well verified by the charge pumping measurements. The detailed descriptions and the simulation results can be found elsewhere[17].

2.2. Characterization of EEPROM writing

A special requirement for numerically simulating EEPROM writing is to treat the stored charges in the floating gate and its corresponding floating gate potential (V_{FG}). To handle this problem, we use the charge boundary condition to uniquely determine the floating gate potential (V_{FG}), and the condition is:

$$F = \oint_{FG} E \cdot \hat{n} W \, dl - \frac{Q_{FG}}{\epsilon_{ox}} = 0, \tag{4}$$

where Q_{FG} is the total charges in the floating gate; $E \cdot \hat{n}$ is the electric field normal to the floating gate surface. Note that the integral in eqn (4) is computed by considering all the floating gate surface with the floating gate width, W. By using the conventional Newton–Raphson (N–R) method, the next guess of V_{FG} is:

$$V_{FG}^{i+1} = V_{FG}^{i} - F \cdot \left(\frac{\partial F}{\partial V_{FG}}\right)^{-1}, \tag{5}$$

and the iteration is carried out until a satisfactory solution is obtained. An initial guess strategy based on a simple capacitance model[18,19] for EEPROM is performed to avoid divergence as well as to reduce the total number of iterations for the N–R method. The quasi-stationary procedure used to characterize the EEPROM writing is described by a flowchart shown in Fig. 1. Instead of partition in time domain[10], our simulator uses the quantity of accumulated charges to vary the time step in EEPROM writing. Because the stored charges directly determine the potential variation in the floating gate, the partition in charge domain is more efficient in CPU time than in time domain. In Fig. 1, the writing procedure starts at $t = t_0 = 0$ and $Q_{FG} = 0$, and at the beginning of each charging step i the amount of stored charges $(Q_{FG})_{i-1}$ in the floating gate is accumulated from all the previous steps. With this stored charges $(Q_{FG})_{i-1}$, the corresponding potential on the floating gate $(V_{FG})_i$ is obtained as mentioned before. With $(V_{FG})_i$ and the drain-bias of writing operation, the gate injection current $(I_g)_i$ of the cell

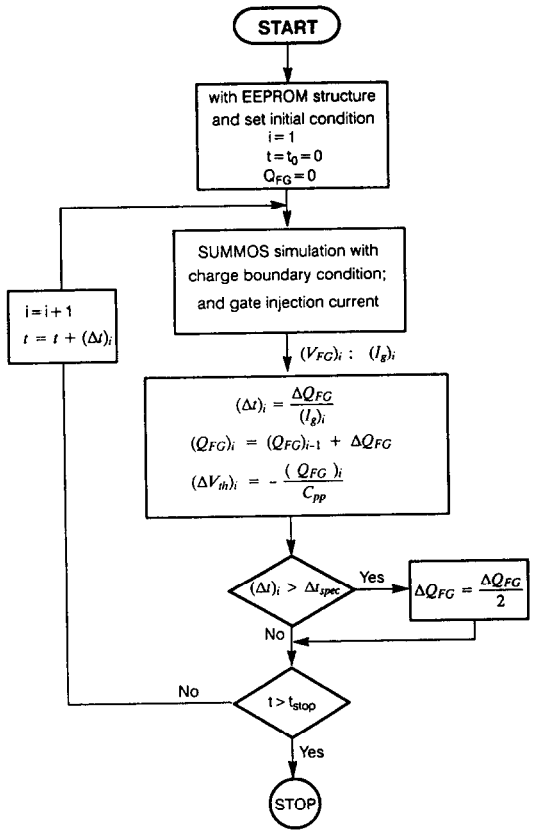

Fig. 1. A characterization procedure for EEPROM writing using the SUMMOS simulator.

transistor can be computed at this charging step by the developed 2D channel hot-electron injection model. The gate-current is assumed to be constant within this charging step so that the time interval $(\Delta t)_i$ used to accumulate stored charges ΔQ_{FG} can be easily obtained as:

$$(\Delta t)_i = \frac{\Delta Q_{FG}}{(I_g)_i}. \tag{6}$$

The writing time and the stored charges at the beginning of the next charging step are:

$$t_i = t_0 + \Sigma_i = {}_1\Delta t_i, \tag{7}$$

and

$$(Q_{FG})_i = (Q_{FG})_{i-1} + (\Delta Q_{FG})_i, \tag{8}$$

respectively. The threshold-voltage shift caused by the channel hot-carrier injection at the ith time step for the cell transistor can be simply calculated as[10,20]:

$$(\Delta V_{th})_i = -\frac{(Q_{FG})_i}{C_{pp}}, \tag{9}$$

where C_{pp} is the capacitance between the control gate and the floating gate. The way for choosing the value of ΔQ_{FG} is to meet a good trade-off between accuracy and computation time. Since larger ΔQ_{FG} may inevitably produce a worse agreement with the simplifying assumption of constant gate-current during the charging step. On the contrary, smaller ΔQ_{FG} requires more simulation steps, resulting in the consumption of CPU time. When the low current injection region is reached, the time interval is prolonged dramatically to maintain the desired ΔQ_{FG} specified for the initial charging period. In order to avoid this inappropriate long time interval, the SUMMOS simulator automatically halves the ΔQ_{FG} value when the calculated time interval is greater than the specified maximum time interval (Δt_{spec}).

2.3. *Oxide-electron-trap-induced degradation*

For EEPROM memory devices, it is inevitable that the threshold-voltage swing becomes narrow gradually. This phenomenon is attributed to the oxide trapping effect of the injected electrons through the thin oxide layer during repeated write/erase operations. In order to simulate the degradation of EEPROM writing caused by these oxide trapped electrons, the SUMMOS simulator has been further modified for this purpose. The whole writing cycle starts at writing cycle = 1 with the oxide trapped electron density of $\mathrm{Trap}_0 = 0$, and then enters the basic writing procedure mentioned in Fig. 1. At each time step of the present writing cycle, the increment of oxide trapped electron density $\Delta\mathrm{Trap}_i$ is calculated by the following rate equation:

$$\frac{\Delta\mathrm{Trap}_i}{(\Delta t)_i} = \sigma_{ox} \cdot (I_g)_i \cdot (N_{ox} - \mathrm{Trap}_{i-1}), \tag{10}$$

with

$$\mathrm{Trap}_i = \mathrm{Trap}_{i-1} + \Delta\mathrm{Trap}_i, \tag{11}$$

where N_{ox} and σ_{ox} are the total trapping density and the capture cross section of oxide electron trap, respectively. Poissons's equation in the oxide region is then solved by including these trapped charges to get the potential and field distributions for the next time step. This process is proceeded until the last time step, and the accumulated oxide trapped electrons are stored as the initial value of Trap_0 for the following writing cycle. This process is repeated until the specified writing cycle is reached for the reliability simulation procedure.

3. RESULTS AND DISCUSSIONS

Comparisons of the simulated and experimental gate-currents for conventional n-MOSFETs with different effective channel lengths of 0.8 and 1.0 μm at drain bias 5.5 V are shown in Fig. 2, in which the experimental data and the detailed device specifications are obtained from [10]. From Fig. 2, it is shown that the experimental gate-current vs gate-bias curves for different channel lengths can be well characterized by our developed gate-current model. It is clearly shown that our gate current model. though simple and macroscopic, has obtained better simulation

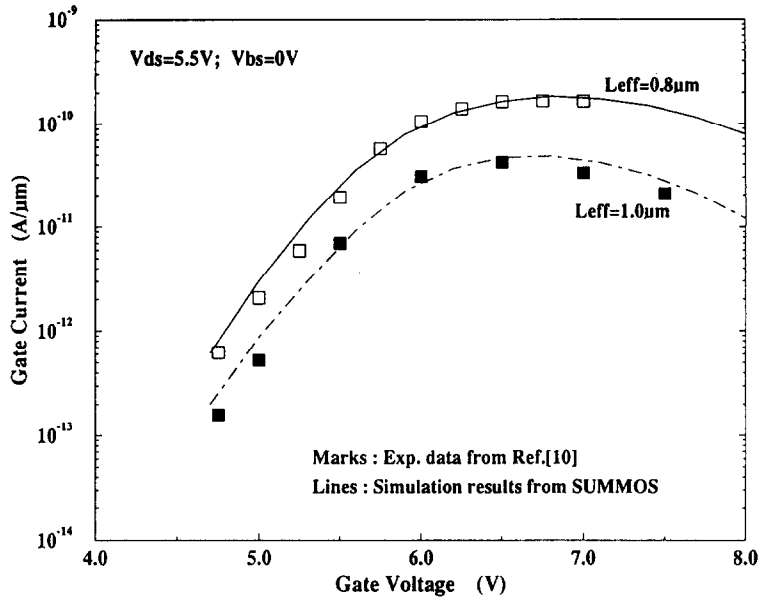

Fig. 2. Comparisons between simulated gate-currents and experimental data for conventional n-MOSFETs biased at $V_{DS} = 5.5$ V with different effective channel lengths.

results as compared to those shown in [10]. Note that, in our gate-current modeling, we have proposed that the 2D effective injection barrier should include the extra barrier lowering term because these injection electrons are "hot" and the term Φ_{BL} in eqn (3) can be used to accurately simulate the hot-carrier effects. It is quite interesting that the parameter γ in eqn (3) is a universal constant of $3.8 \times 10^{-6} q$ (cm) for different channel lengths. This fact has been proven by our previous results using different MOS technologies[15,17].

Modelings of EEPROM writing have also been verified by comparing with experimental writing characteristics of a EEPROM cell in [10]. Figure 3 shows comparisons between measurements and simulations of the output characteristics of a EEPROM cell with $L_{eff} = 0.5\,\mu$m, in which the figure refers to the case of no charge stored in the floating gate. The achieved agreement demonstrates that the device parameters used for SUMMOS simulator have been accurately extracted. Figure 4 and Fig. 5 show the dependence of writing characteristics on applied

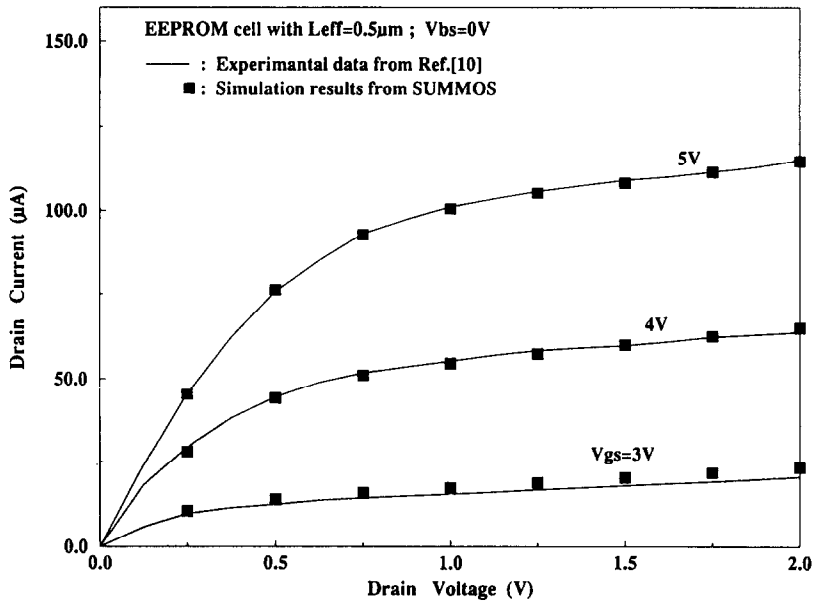

Fig. 3. Comparisons between calculated I_{DS}–V_{DS} characteristics and experimental data for an EEPROM cell with $L_{eff} = 0.5\,\mu$m.

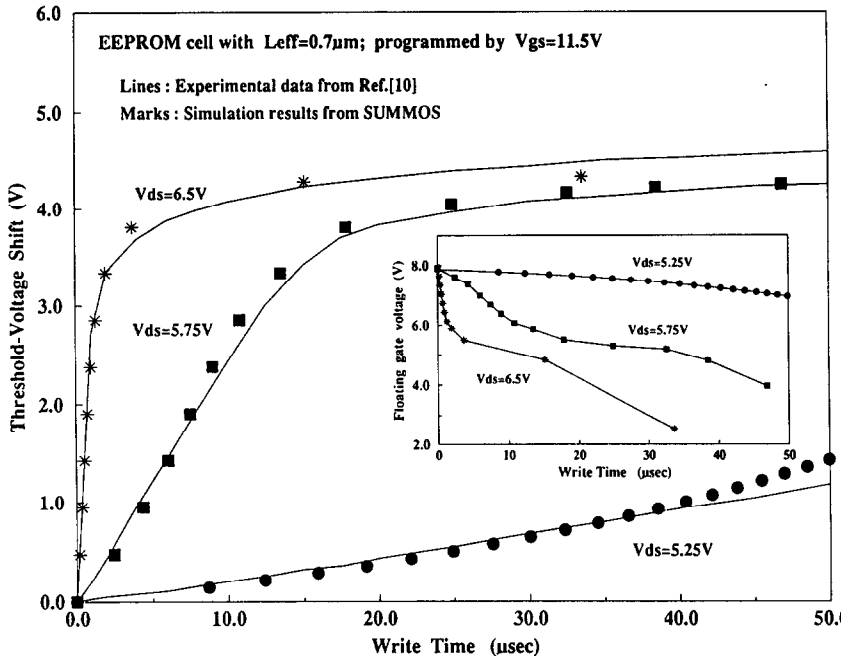

Fig. 4. The writing characteristics of a EEPROM cell with $L_{\text{eff}} = 0.7\,\mu\text{m}$ programmed by different drain-voltages. Lines: experimental data from [10]; Marks: simulation results from SUMMOS. Note that the variation of floating gate voltage V_{FG} during the writing process is inserted.

drain-bias and channel length, respectively. The whole ranges of experimental results with EEPROMs featuring different channel lengths (0.5–0.8 μm) and operated with wide range of applied biases have been simulated with the same γ value. From the achieved agreement shown in Figs 4 and 5, it indicates that our modeling is verified to accurately simulate the gate-current for the writing process with wide range of floating gate-voltages from 3 to 8 V, as shown in the inset of Fig. 4. Initially, the floating gate-voltage is higher than V_{DS}, which makes the threshold-voltage of the cell transistor increasing rapidly due to the high injection gate-current. When the injected charges are gradually accumulated in the floating gate, the charging current will first increase and then decrease abruptly due to the bell-shaped gate-current shown in Fig. 2. Thus, the threshold-voltage shift vs writing time curves tend to saturate when the floating gate-

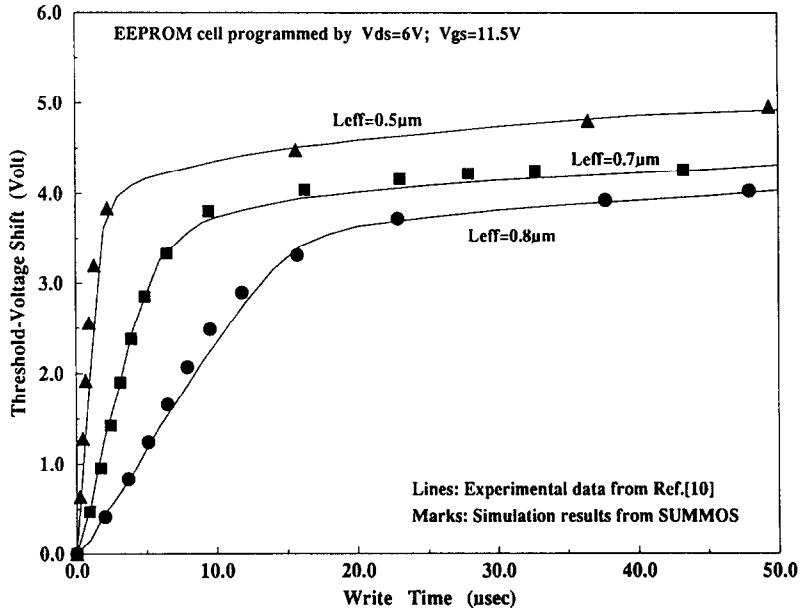

Fig. 5. The writing characteristics of EEPROM cells with different effective channel lengths. Lines: experimental data from [10]; Marks: simulation results from SUMMOS.

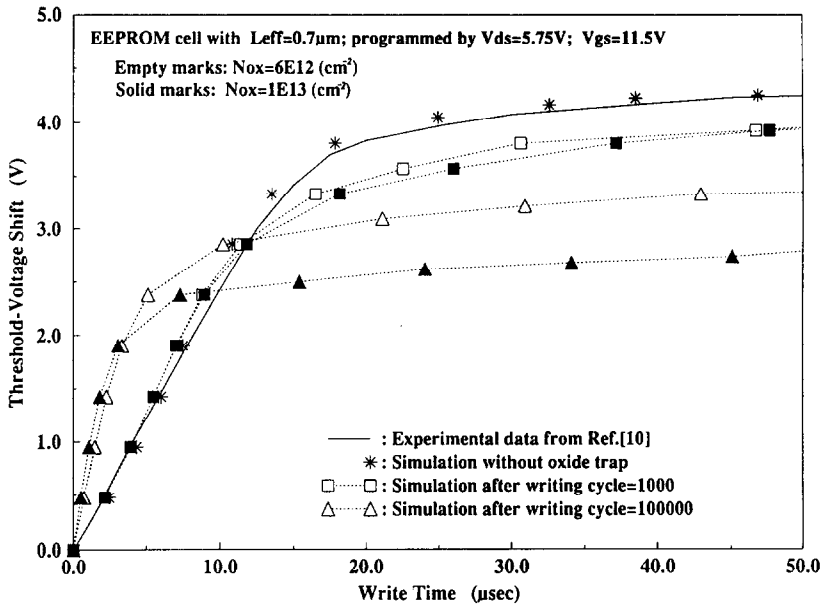

Fig. 6. The degradation of threshold-voltage shift for EEPROM writing. The total oxide trap density (N_{ox}) used in the rate equation are 6×10^{12} (cm^{-2}) and 1×10^{13} (cm^{-2}) for the capture cross section (σ_{ox}) of 1.5×10^{-18} (cm^2).

voltage is lower than V_{DS}, as shown in Figs 4 and 5.

The degradations of threshold-voltage swing for 0.7 μm EEPROM after different writing cycles of 1000 and 100,000 with different total trapping densities (N_{ox}) of 6.0E12 and 1.0E13 (cm^{-2}) are shown in Fig. 6. It is known that the trapped electrons in oxide layer will suppress the floating gate-voltage. With the same applied bias on the control gate, the floating gate-voltage V_{FG} will decrease with increasing the trapped electrons in the oxide. Therefore, the threshold-voltage shift will first

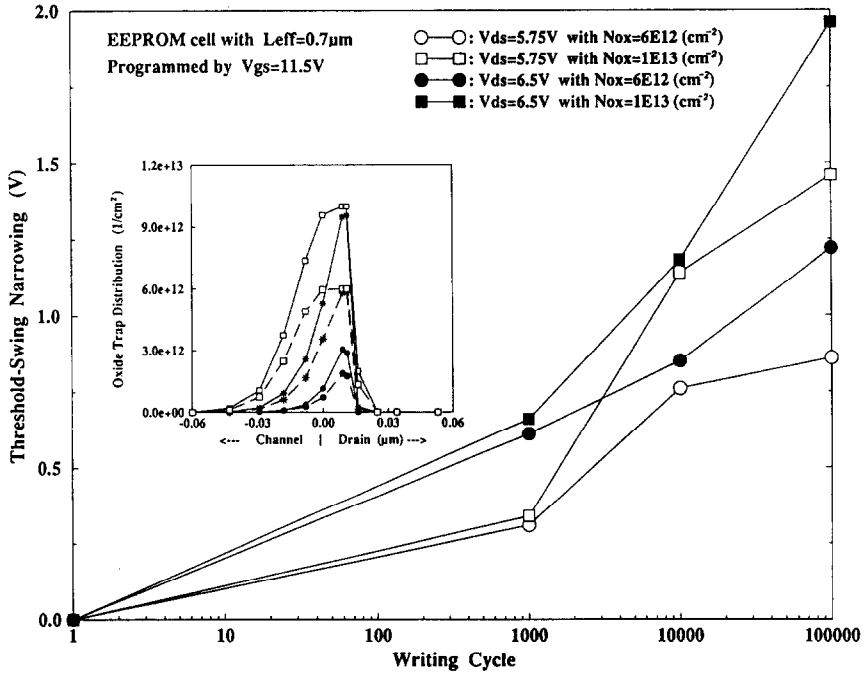

Fig. 7. The narrowing of threshold-voltage swing for a EEPROM cell operated with different drain-voltages for different total oxide-trap densities. The spatial distributions of oxide traps after writing cycles = 1000 (○), 10,000 (*), and 100,000 (□) programmed by $V_{DS} = 5.75$ V and $V_{GS} = 11.5$ are inserted with the dashed line: $N_{ox} = 6E12$ (cm^{-2}) and the solid line: $N_{ox} = 1E13$ (cm^{-2}).

speed up due to the bell-shaped gate injection current and then rapidly saturate to a lower value as compared to that of devices without the trapped electrons in the oxide. From the demonstration in Fig. 6, the trapped electrons in the oxide will narrow the threshold-voltage swing in EEPROM writing, and this effect is getting worse with increasing the writing cycle and the total oxide-trap density. Figure 7 summarizes the narrowing of threshold-voltage swing as compared to the case without the trapped electrons in the oxide. Note that, all the threshold-voltage swings are extracted at a writing time of 40 μs. It is clearly seen that the degradation is enhanced when the trapping rate (total trap density or the capture cross section), the number of writing cycle, and the writing bias are increased. The spatial distribution of oxide electron traps after different writing cycles of 1000, 10,000 and 100,000 is shown in the inset of Fig. 7. In Figs 6 and 7, the rate equations are calculated with a fixed capture cross section of $\sigma_{ox} = 1.5E\text{-}18$ (cm^2) and different total oxide trap densities of 6.0E12 and 1.0E13 (cm^{-2}). All these parameters are dependent on the fabrication process and can be determined from experimental measurements[21]. Then, substituting these parameters into our developed simulation technique, the optimized design for EEPROM can be obtained just before device fabrication.

4. CONCLUSIONS

In this paper, we have presented the developed gate-current modeling for hot-carrier effects and its applications to the writing and reliability problems of EEPROM devices. The gate current of short channel n-MOSFETs has been computed by a simple and accurate simulation technique, in which a 2D channel hot-carrier enhanced injection probability is proposed. This enhanced factor is modeled by an effective barrier lowering term which is simply expressed in terms of the actual hot-electron current flow path and its power density gained from the local electric field. The fitting parameter γ is shown to be a universal constant for all the simulation cases. With this gate injection model and charge boundary condition, characterization of n-channel EEPROM writing has been well verified to be consistent with the experimental data of different channel lengths (0.5–1.0 μm) for wide ranges of bias conditions. Moreover, the effects of oxide electron traps have been incorporated into our simulator to characterize the reliability problems during repeated EEPROM cell write/erase operations. It is shown that oxide electron traps will narrow the threshold-voltage swing in EEPROM writing the degradation is enhanced with increasing and the trapping rate and applied biases. With the efficient and accurate simulation technique proposed, the SUMMOS simulator can be used as a useful computer-aided-design (CAD) tool to support the development of scaled EEPROM memory devices in order to speed up the writing procedure and improve its long term reliability.

Acknowledgements—The authors would like to express our sincere thanks to the National Science Council, Taiwan, Republic of China for continuous grant support under the contract NSC-82-0404-E009-216. Special thanks are given to Dr Ruey-Kuen Perng and Dr Pole-Shang Lin for their helpful discussions.

REFERENCES

1. F. Masuoka, M. Asano, H. Iwahashi, T. Komuro and S. Tanaka, *IEDM Tech. Dig.*, p. 464 (1984).
2. G. Verma and N. Mielke, *Proc. IRPS*, p. 158 (1988).
3. N. Ajika, M. Ohi, H. Arima, T. Matsukawa and N. Tsubouchi, *IEDM Tech. Dig.*, p. 115 (1990).
4. Y. Yamauchi, K. Tanaka, H. Shibayama and R. Miyake, *IEDM Tech. Dig.*, p. 319 (1991).
5. J. S. Witters, G. Groeseneken and H. E. Maes, *IEDM Tech. Dig.*, p. 544 (1987).
6. D. A. Baglee, T. Sugawara, S. Fukawa and K. Mori, *Proc. IRPS*, p. 93 (1987).
7. R. C. Wijburg, G. J. Hemink, J. Middelhoek, H. Wallinga and T. J. Mouthaan, *IEEE Trans. Electron Devices* **ED-37,** 111 (1991).
8. E. Sangiorgi, B. Riccó and F. Venturi, *IEEE Trans. Computer-Aided Design* **CAD-7,** 259 (1988).
9. M. V. Fischetti and S. E. Laux, *Phys. Rev. B* **38,** 9721 (1988).
10. C. Fiegna, F. Venturi, M. Melanotte, E. Sangiorgi and B. Riccó, *IEEE Trans. Electron Devices* **ED-38,** 603 (1991).
11. C. Huang, T. Wang, C. N. Chen, M. C. Cheng and J. Fu, *IEEE Trans. Electron Devices* **ED-39,** 2562 (1992).
12. D. Cassi and B. Riccó, *IEEE Trans. Electron Devices* **ED-37,** 1514 (1990).
13. R. K. Perng, P. S. Lin and C. Y. Wu, *Solid-St. Electron.* **34,** 635 (1991).
14. K. S. Wen and C. Y. Wu, *IEEE Trans. Electron Devices.* Submitted.
15. K. S. Wen and C. Y. Wu, *IEEE Trans. Electron Devices.* Submitted.
16. T. H. Ning, C. M. Osburn and H. N. Yu, *J. appl. Phys.* **48,** 286 (1977).
17. K. S. Wen and C. Y. Wu, *Solid-St. Electron.* In press.
18. K. Prall, W. I. Kinney and J. Macro, *IEEE Trans. Electron Devices* **ED-34,** 2463 (1987).
19. B. J. Sheu, W. J. Hsu and P. K. Ko, *IEEE Trans. Computer-Aided Design* **CAD-7,** 520 (1988).
20. B. Eitan and D. F. Bentchkowsky, *IEEE Trans. Electron Devices* **ED-28,** 328 (1981).
21. C. F. Chen and C. Y. Wu, *IEEE Trans. Electron Devices* **ED-34,** 1540 (1987).