An Analytical Model for Optimization of Programming Efficiency and Uniformity of Split Gate Source-Side Injection Superflash Memory

Huinan Guan, *Member, IEEE*, Dana Lee, *Member, IEEE*, and G. P. Li

***Abstract*—An analytical model for evaluating the programming efficiency and uniformity of SST SuperFlash cells is developed for the first time. Starting with two-dimensional electric field analysis, this model calculates the effective hot electron injection-induced gate current during programming. Based on full transient simulation of the calculated gate current, the time to program is then developed and used as a figure of merit to evaluate SST cells programming. The time-to-program model predicts the nonlinear transformation from control-floating gate coupling ratios to the programming speed and that the programming distribution broadening correlates with coupling ratios. The model also suggests that higher bias voltage of $(V_d - V_{cg})$ and a lower coupling ratio should result in better programming efficiency and uniformity.**

***Index Terms*—EEPROM, hot carriers, programming speed, time to program model, two-dimensional (2-D) analysis.**

I. Introduction

FLASH memory is becoming commonly used as memory storage in replacement of EPROM, E^2PROM, and magnetic memory storage devices, because it is nonvolatile and reprogrammable. Many electronic devices on the market today have already incorporated flash memory. Phone answering machines with flash memory as the message storage media are very popular now. Telecommunication products such as cellular phones and fax machines can use flash memory to store numbers during battery replacement. PC manufactures are planning to install operating systems, such as DOS and Windows, in the embedded flash memory on motherboards. Portable PCs use flash memory cards as bulk data storage instead of magnetic disks, because flash memory cards are lighter, more tolerant to shock, and smaller in size.

Among various designs of flash memory, split gate cells with source-side injection outperform stacked gate cells with higher programming efficiency and overerase immunity [1]–[4]. While the programming operation in a stacked-gate cell is well analyzed [5], that of the split gate cell has not yet been investigated sufficiently. The existing model for programming split triple-gate flash memory cells resembles the stack gate model with a virtual drain assumption [1]. As shown in Fig. 1(a), an inversion layer is formed under the floating gate channel and acts like a virtual drain, when the triple-gate cell is being programmed at floating gate voltages higher than the drain. Hot electron injection into the select gate channel was analyzed the same way as programming in the stack gate flash memory. However, this modeling approach cannot be applied directly to the split dual-gate flash cells. Instead of using an extra control gate, the dual-gate cell controls the floating gate potential by a highly doped diffusion region on the drain side via the

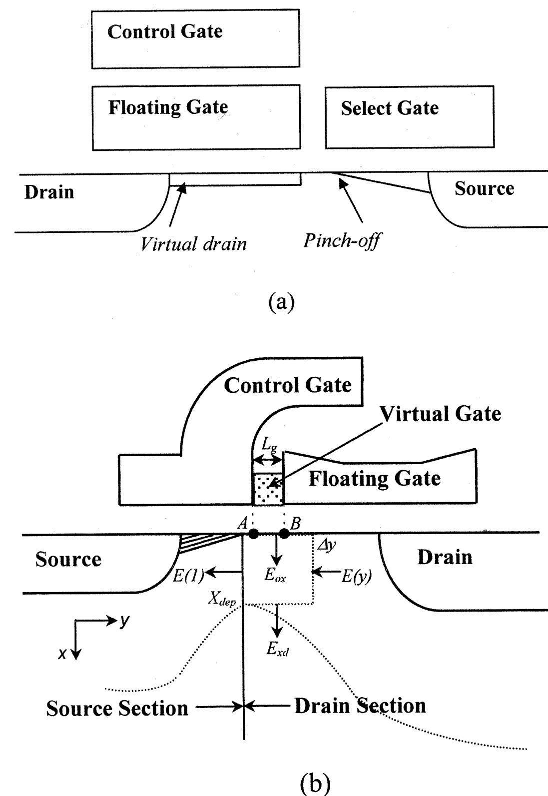

Fig. 1. Cross-sectional views of two source-side injection flash cells under programming condition. (a). Triple-gate configuration. The channel under the floating gate is in the linear region, which serves as a virtual drain for the select gate. (b) Dual-gate configuration (SST cell). The channel under the floating gate is fully depleted. The schematic of a two-dimensional analysis is also illustrated in this figure.

Manuscript received August 23, 2002; revised December 6, 2002. The review of this paper was arranged by Editor J. Vasi.

H. Guan and G. P. Li are with the Integrated Nanosystems Research Facility, Department of Electrical and Computer Engineering, University of California, Irvine, CA 92697 USA (e-mail: hguan@ece.uci.edu).

D. Lee is with the Silicon Storage Technology, Inc., Sunnyvale, CA 94086 USA.

Digital Object Identifier 10.1109/TED.2003.811416

0018-9383/03$17.00 © 2003 IEEE

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.

capacitive coupling as shown in Fig. 1(b) for the SuperFlash[1] SST cell. Due to the fact that other electrodes providing coupling voltage to the floating gate are at much lower bias than the drain, the floating gate voltage in SST cells is lower than drain voltage. Consequently, an SST cell's floating gate channel is in the saturation condition during programming. Therefore, the virtual drain assumption is not valid for the dual-gate flash cell. It is clear by now that an alternative hot electronic injection model for SuperFlash SST cell programming is needed to determine the cell programming efficiency.

Furthermore, as the flash memory cell further scales down in dimension and the memory density continues to increase, the power supply voltage and process variation may pose new constraints on the cell programming conditions. In order to reliably screen the high-density flash memory chips and to program them at the lowest possible voltage, not only the programming efficiency dependence on process variation and programming voltage but also their effects on the programming uniformity need to be fully understood. As a distribution of the coupling ratio between control and floating gates is often expected due to process variation in SST flash memory products, an analytical model to predict the programming efficiency and uniformity is desirable for the success of future generation of low-voltage high-density SuperFlash memory chips.

In this paper, an analytical model for evaluating the programming efficiency and uniformity of SST SuperFlash cells will be developed. Two-dimensional (2-D) analysis and lucky electron models are used to develop the electric field distribution, gate current, and substrate current models as SST cells being programmed. The time to program, which is derived from a full transient simulation of the gate current, provides a direct measure relating the programming speed of SST cells to bias conditions and device parameters. The programming uniformity due to process variations is also predicted and optimized based on the time-to-program model.

II. SST SuperFlash Technology

A cross-sectional view of SuperFlash SST cell along the channel is shown in Fig. 1(b), consisting of a control and floating gate channels. In order to achieve a higher coupling ratio to the floating gate for improving programming efficiency, the drain diffusion is doped deeper than that of the source. During programming, both channels are operating under saturation conditions. The use of a source-side injection produces the vertical electric field in favor of electron injection, which results in a high programming efficiency for SST cells. Poly control gate to poly floating gate tunneling via a field-enhancing tunneling injector is used for cell erasing [6]. The typical operating conditions of the SST cell are listed in Table I. As a result of independent control of the control and floating gate, there is no overerase or overprogramming problem. This simplifies the peripheral circuitry design, resulting in a high array efficiency [7].

The floating and control gate lengths are both equal to 0.25 μm in this study. The control gate oxide and the interpoly oxide in the sidewall are grown at the same time, having a thickness of 18 nm. The floating gate oxide thickness is 8 nm. While the coupling ratio between the floating gate and source is around 70%, that between the floating gate and the control gate is about 25%. The details of the SST technology and operation can be found in [6].

[1]SuperFlash is a registered trademark of Silicon Storage Technology, Inc.

TABLE I
BIAS CONDITION FOR THE OPERATION OF THE SST CELL

Bias Condition	Drain	Source	Control Gate	Substrate
Program	9V	-5μA	1.7V	Ground
Erase	Ground	Ground	12V	Ground
Read	Ground	1V	1.7V	Ground

III. Floating Gate Current Model Development and Verification

The floating gate current is the most direct measure of hot carrier injection and cell programming efficiency for the SST cell. In order to model the gate current accurately, both lateral and vertical electric field distribution along the channel should be determined. A 2-D analysis for solving the electric field distribution is commonly used to determine hot carrier generation [5], [8].

A. Electric Field Modeling

Fig. 1(b) shows a schematic diagram for electric field calculation. The flash memory cell channel is divided into source and drain sections for electric field modeling. A virtual gate connecting the control gate and the floating gate with linearly increased potential from A to B is assumed when solving the Poisson equation. To facilitate the modeling, the oxide thickness is taken as a constant of 18 nm across the channel. As results of biasing conditions used in programming, the entire channel under the floating gate is actually depleted, resulting in two lateral electric field peaks located on each side of the floating gate. The formation of two electric field peaks can be understood by noting the pinch-off condition at the control gate and at the floating gate. The continuous electric field distribution in between these two peaks and its strength higher than velocity saturation field suggest that mobile charge (electrons) should not accumulate in this region, resulting in a depletion of electrons in the entire region from the control gate pinch-off point to the drain junction. Thus, the SST dual-gate cell can be simplified as one NMOS with a source section and a drain section. According to the gate potential, the drain section is divided into three parts: 1) $V_g = V_{\rm cg}$; 2) gate voltage increases linearly from $V_{\rm cg}$ to $V_{\rm fg}$; and 3) $V_g = V_{\rm fg}$.

By applying Gauss's law to a rectangular box of height $X_{\rm dep}$ and length Δy in the channel depletion region, the electric field distribution along the channel can be solved. The detailed derivation and results will be published elsewhere [9]. The peak value of lateral electric field in the gap region can be approximated by

$$\begin{aligned} E_m \approx& k_g \left[1 - \frac{1}{\sqrt{1 + AL_g + \frac{1}{2}(AL_g)^2}}\right] = \frac{C_1}{L_g}(V_{\rm fg} - V_{\rm cg}) \\ C_1 =& 1 - \frac{1}{\sqrt{1 + AL_g + \frac{1}{2}(AL_g)^2}} \\ A =& \sqrt{\frac{C_{\rm ox} n_{\rm sp}}{X_{\rm dep}\varepsilon_s}} \end{aligned} \tag{1}$$

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.

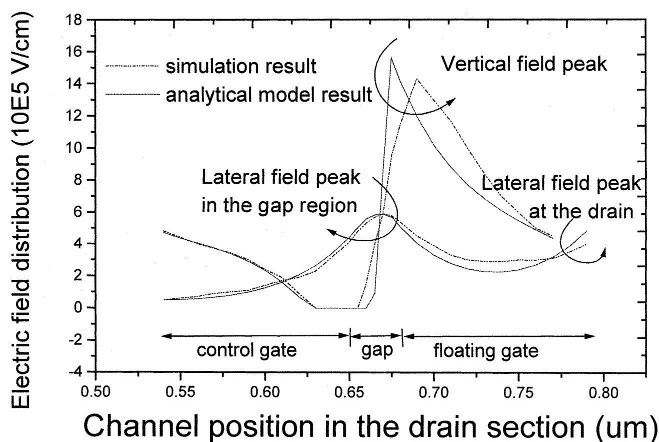

Fig. 2. The lateral and vertical electric field along the channel surface under programming conditions: $V_{\rm fg} = 7$ V, $V_d = 9$ V, and $V_{\rm cg} = 1.6$.

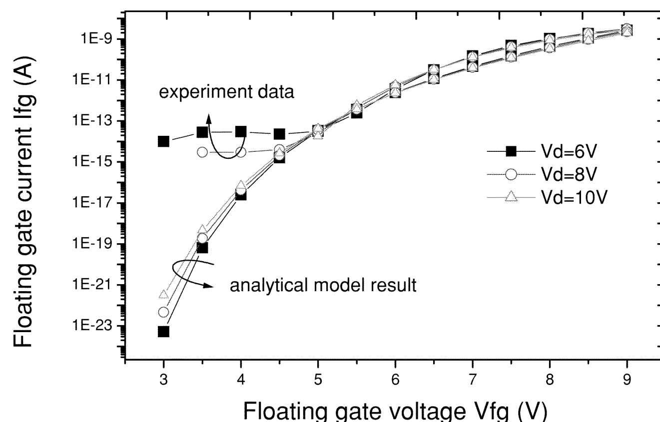

Fig. 3. Analytical model result compared with the experimental data for the floating gate current $I_{\rm fg}$ as a function of the floating gate voltage.

where $X_{\rm dep}$ is the depletion depth at the control gate pinch-off point, L_g is the gap region width, ε_s is the permittivity of Si, and $n_{\rm sp}$ is a fitting parameter as discussed in [10]. Equation (1) suggests that, for SST cells, the maximum lateral electric field E_m is independent of drain voltage and is approximately proportional to the difference of the floating and control gate potential for a given gap region width. This is in contrast with the field distribution of the typical NMOS, where the lateral field peaks at the drain end, and its value can be approximated by $(V_d - V_{\rm dsat})/l$ [11]. However, in real cells, the drain voltage will affect the electric field distribution via capacitive coupling to the floating gate. It is noted that the gap region width is a critical device parameter that determines E_m. A narrower gap region is desirable for higher E_m that results in a higher gate current and more efficient cell programming. However, the gap region optimization is constrained by dielectric control issues including tunneling control and capacitive coupling optimization.

Fig. 2 plots results of lateral and vertical electric field modeling compared with the 2-D simulation (using Avant! TSUPREM and MEDICI). Even with the approximation made above, the analytical model predicts the peak lateral field (E_m) in the gap region and peak vertical field $(E_{\rm ox\,max})$ fairly close to simulation results. The second lateral field peak at the drain end is due to the saturation condition in the floating gate channel. Since the vertical field at the drain side is not in favor of electron injection to the floating gate, the hot carrier population in the drain does not contribute significantly to the gate current.

B. Gate Current and Substrate Current Modeling

The electric fields derived above are used to model the gate and substrate current. Based on the Lucky-Electron Model (LEM), the gate current can be derived by integration of hot electron injection probability over the channel as follows, and the details will be published in [9]:

$$I_g = \frac{I_d \lambda^2 E_m^2}{4\lambda_r A \phi_b^2 \left[1 + \frac{E_m \lambda}{\phi_b (2-m)}\right]} \exp\left(-\frac{\phi_b}{\lambda E_m} m\right) \cdot P(E_{\rm ox}). \tag{2}$$

By assuming $m = E_m(2/E_m - 1/E_A)$ as a constant, where E_A is the lateral field at point A, the best fitting for m is equal to 0.89 over the electric field range of interest. $P(E_{\rm ox})$ can be treated as a constant in the region of interest. The potential barrier ϕ_b is found to be [12]: $\phi_b = 3 - \beta\sqrt{E_{\rm ox}} - \vartheta E_{\rm ox}^{2/3}$, where β is $2.59E - 4\ ({\rm V \cdot cm})^{1/2}$. ϑ is determined by the comparison with the experiment data. The gate current is measured on a modified design of the SST cell with an additional accessible contact to the floating gate. Fig. 3 plots the measured gate current as a function of the floating gate voltage together with the calculated results for comparison. Good fit is obtained over four decades of gate current. As the floating gate potential rises, the floating gate channel starts operating into linear mode, invalidating the assumption for calculating the electric field in the drain section. As a result, the gate current model deviates from the experiment data. It is noted that the drain voltage does not play a major role in changing gate current, due to the lateral electric field not being strongly dependent on it as illustrated in part A.

In this modeling, ϑ is chosen to be $1.0E - 4\ {\rm V}^{1/2}{\rm cm}^{2/3}$, which is higher than the result of $4E - 5\ {\rm V}^{1/2}{\rm cm}^{2/3}$ in [13]. The term of $\vartheta E_{\rm ox}^{2/3}$ represents the barrier lowering due to the tunneling effect. Higher ϑ suggests poorer quality of gate oxide. For SST cells, hot carrier injection concentrates at the channel region under the left edge of the floating gate. Part of the oxide at the injection region is formed by oxidation of highly doped floating gate polysilicon, possibly resulting in oxide with a higher probability for tunneling. In addition, the underestimated oxide thickness around the gap region could be another reason that magnifies the tunneling coefficient. It is noted that the method using normalized $I_g/I_{\rm sub}$ over $E_{\rm ox}$ to derive ϑ [1], [13] cannot be easily realized for SST cells, because $E_{\rm ox}$ cannot be measured directly from the drain side bias. In (2), the electron mean free path λ over the variation of energy is assumed to be a constant of 3 nm, which is shorter than the commonly used ones of 5–9.2 nm [11], [13], [14]. The use of low value λ could be justified by the following.

1) The electron mean free path actually decreases with the increasing carriers energy according to Monte Carlo simulation of more accurate channel hot carrier phenomena [14]–[16]. 3 nm used here for λ could be the average value of the mean free path over the carriers' energy range of interest.

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.

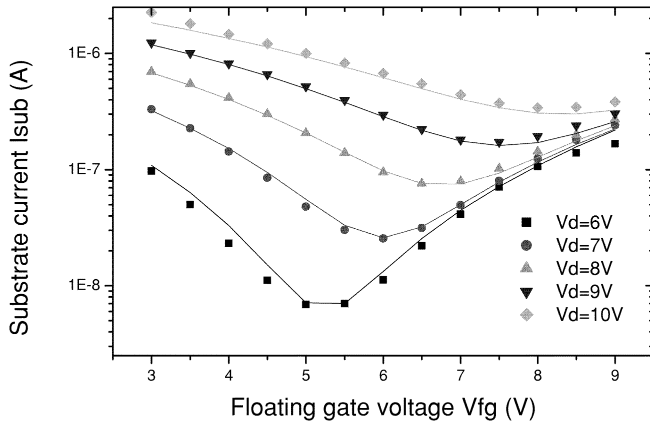

Fig. 4. Validation of the analytical model for the substrate current as a function of the floating gate voltage by experimental data. The solid lines represent theoretical result, and the dots are experimental data.

2) SST cells drain current is basically a surface current. With the extra scattering at the Si–SiO_2 interface, the electron mean free path at the surface is expected to be smaller than that in the bulk mean free path [17].
3) The assumption of the Maxwellian distribution function used in the LEM modeling is not accurate for a realistic device especially as the device scaling down. With an increasing electric field gradient, the nonlocal effects become significant to lower the efficiency of hot carriers generation [15], thus a shorter λ chosen here.

While a single λ value used in (2) is over simplified for calculating the gate current, it reflects the overall contribution from different carriers energy, carriers' location, and local electric field and provides a good engineering insight for assessing the device design in flash memory cells.

Substrate current is also a good indicator to study the lateral electric field and mean free path λ. Since there are two peaks of lateral field along the channel generating impact ionization current, both need to be included in substrate current calculation. By integrating the ionization coefficient $\alpha \exp(-\beta/E)$ over the depletion region, the current can be approximated as [18]

$$I_{\rm sub} = I_s \frac{A_{\rm ion}}{B_{\rm ion}} \left[\frac{E_{m1}^2}{\left|\frac{dE}{dy}\right|_{\max 1}} \exp\left(\frac{-B_{\rm ion}}{E_{m1}}\right) + \frac{E_{m2}^2}{\left|\frac{dE}{dy}\right|_{\max 2}} \exp\left(\frac{-B_{\rm ion}}{E_{m2}}\right) \right]. \quad (3)$$

As shown in Fig. 4, the modeling results of substrate current versus floating gate potential are compared with the experimental data. The bell shape substrate current curve indicates two components of the substrate current having opposite dependence on the floating gate potential. The substrate current is dominated by impact ionization at the drain end at low floating gate potential, while it is controlled by impact ionization at gap region at higher potential. The impact-ionization coefficient $B_{\rm ion}$ is chosen to be to 4.1E6 (V/cm) at $\lambda = 3$ nm and $E_p = 1.23$ V. Good experimental data fits for both gate and substrate current with the same mean free path suggests that a low value of λ is a reasonable approximation for SST cells.

IV. Modeling the Time to Program

For the SST cells, the accessible floating gate contact is not available for measurement. On the other hand, the substrate current is not a good indicator of SST cell programming performance, because it contains the additional drain side impact-ionization current, not directly related to programming gate current. Thus, an alternative measure of programming, the time to program, is used in this work as the figure of merit to evaluate programming speed of SST cells. The time to program is defined as the time needed to program a cell such that the read current after programming is at a certain level (e.g., 5%) compared with the preprogramming read current. This measurable quantity can be derived by integrating gate current over time with the previous developed gate current model.

To solve the time to program, the integration equation is written as

$$\frac{dV_{\rm fg}(t)}{dt} = -\frac{1}{C_{\rm fg}} I_{\rm fg}(t) \quad (4)$$

where $C_{\rm fg}$ is the total capacitance between the floating gate and other electrode nodes. $V_{\rm fg}(t)$ is the function of bias and coupling ratio α

$$V_{\rm fg}(t) = \alpha V_{\rm cg} + (1-\alpha)V_d + V_q(t) \quad (5a)$$

$$\alpha = \frac{C_{fc}}{C_{fc}+C_{fd}+C_{fs}} \approx \frac{C_{fc}}{C_{fc}+C_{fd}} \quad (5b)$$

where V_q is the floating gate potential due to the existing charge and C_{fc}, C_{fd}, and C_{fs} are the capacitance between the floating and control gate, the floating gate and the drain, and the floating gate and the substrate, respectively. Substituting (2) into (4), the time to program can be solved with a triangular approximation

$$\Delta t(V_{q2}) = C_g \frac{\Delta V_q}{(V_{\rm fg2}-V_{\rm cg})(V_{\rm fg1}-V_{\rm cg})} \cdot \frac{y_2^2 + a y_2}{(y_2+b)^2} e^{y_2} \quad (6)$$

where the subscript "1" represents the initial bias condition, the subscript "2" represents the post-programming bias at each node, and

$$y = \frac{m\phi_b}{\lambda E_m} \approx \frac{\frac{m\left[3 - C_4 - C_3(V_{\rm fg}(t) - V_{\rm cg})\right]}{C_1 \lambda}}{L_g(V_{\rm fg}(t) - V_{\rm cg})} = \frac{mL_g(3-C_4)}{C_1 A} \cdot \frac{1}{(1-\alpha)(V_d - V_{\rm cg}) + V_q} - \frac{mL_g C_3}{C_1 A} \quad (6a)$$

$$C_g = \frac{2C_{\rm fg}\lambda_r L_g^2 A}{C_1^2 \lambda^2 I_d P(E_{\rm ox})} \quad (6b)$$

where $a = m/(2-m)$, $b = C_3 m L_g/(\lambda C_1)$, and C_3 and C_4 are related to ϕ_b. The fact that the exponential term e^y in (1) is about 2% of e^y in (2) when the floating gate potential is changed from 7 V to 5.5 V, a typical case for 0.25-μm SST cells operation, justifies the use of the triangular approximation. This suggests that the postprogramming stage of cell potential is the determining factor for the time to program.

It is noted that the only uncertain quantity in the expression of y is V_q, since the relation between V_q and the read current at

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.

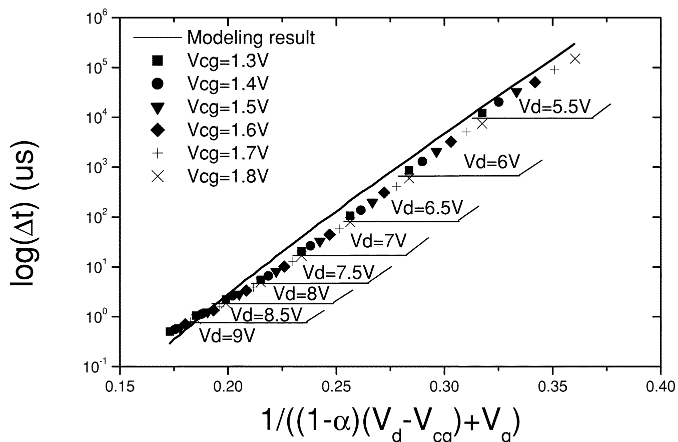

Fig. 5. The linear dependence of time to program in log scale on $1/((1-\alpha)(V_d-V_{\rm cg})+V_q)$. The solid line is from modeling results based on 6(a) and 6(b). The scattering dots represent the experiment data for time-to-program measurement at different bias conditions. V_q is assumed to decrease from 1.5 to 0 V by the programming.

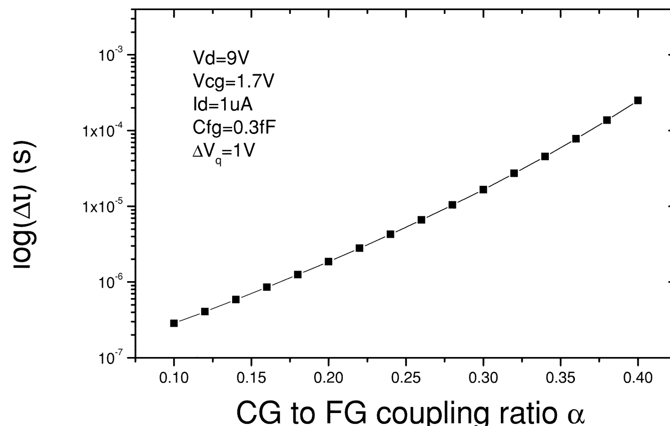

Fig. 6. The time to program increases with the control gate to the floating gate coupling ratio based on a full transient simulation for given gap region width.

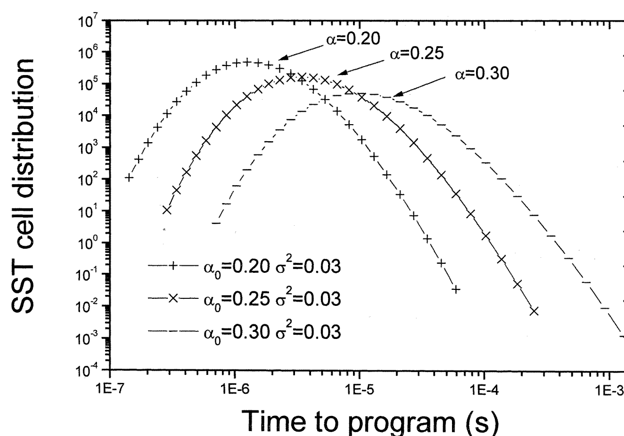

Fig. 7. The simulation of time-to-program distribution for three groups of SST cells: those having Gaussian distribution in coupling ratio with different mean of 0.20, 0.25 and 0.30, and the same deviation of 0.03.

time to program is unknown. Based on the floating gate potential expression, previous data in [6] indicate the postprogramming V_q around 0 V with the error bar of 0–0.5 V. V_q of 0 V and $\Delta V_q = 1.5$ V due to the programming, a typical case for 0.25 μm SST cells, are chosen for both modeling and plotting experimental data. Fig. 5 depicts the curves of $\log(\Delta t)$ versus $1/((1-\alpha)(V_d-V_{\rm cg})+V_q)$ for both calculated results from (6) and experimental results measured at various V_d and $V_{\rm cg}$ bias combinations for programming. An almost straight-line result is shown in Fig. 5, indicating the exponential term e^y dominating the time to program over all the other terms in (6). The slope of line is given by $mL_g(3-C_4)/(C_1A)$. The intersection with the x axis can be found from the combination of (6) and 6(a). A good match between theory and experiment results confirms the validation of (6).

V. Programming Speed Distribution

The time-to-program model suggests that the coupling ratio between the control gate and floating gate (α) is a critical parameter dictating the programming speed. Following (6), the time needed to program a cell increases almost exponentially with α, as shown in Fig. 6. Under typical programming conditions, every 10% increment in coupling ratio induces about one decade's prolonging in time. Furthermore, the rate of increase in time to program as a function of α is accelerated as the α value is greater than 0.25. This also suggests that α should be smaller in order to have a tighter control in time to program. However, unlike other device parameters, α cannot be easily controlled through process tuning or measured directly with specific physical characterizations, because it is a combined function of several parameters, e.g., the gap region width, the overlap of the control gate to the floating gate, drain doping, and channel length. It is conceivable to have a considerably broad coupling ratio distribution with process variance in each device parameter. Since the programming speed is very sensitive to the change in coupling ratio α, a relatively wide range of programming speed distribution is also expected, which is a critical concern of programming performance uniformity for the memory chip design.

The time-to-program model renders an analytical approach to predict the distribution of the programming speed for a memory chip when the coupling ratio distribution is known. Since α from a group of split gate cells is a random variate, and since α from each individual cell has an arbitrary probability distribution with a mean and a finite variance, a Gaussian distribution for α can be assumed for the group of cells. For 0.25-μm SST cells, the mean of α and the deviation σ^2 has a typical value of about 0.25 and 0.03 respectively, representing over 90% of cells with α falling into the range from 0.20 to 0.30. Fig. 7 plots the corresponding time-to-program distribution under the same bias conditions as those used in Fig. 6. It is observed that, through the nonlinear transformation from α to the time to program, the resulting distribution is not a Gaussian distribution any more. This nonlinear transformation can be understood by a simple mathematic deviation: $dn = g(\alpha)\,d\alpha = g(\alpha)\,(d\alpha/dt)dt$ where dn is the number of cells within the range of coupling ratio $d\alpha$, $g(\alpha)$ represents the Gaussian distribution function with respect to α, and t is the time to program. The nonlinear term $d\alpha/dt$ tends to squeeze the time-to-program distribution at lower coupling ratio values, while broadening the distribution at higher α values. The peak of the distribution in the time to program has a corresponding coupling ratio smaller than 0.25. The time-to-program distribution shows that a lower coupling ratio mean leads

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.

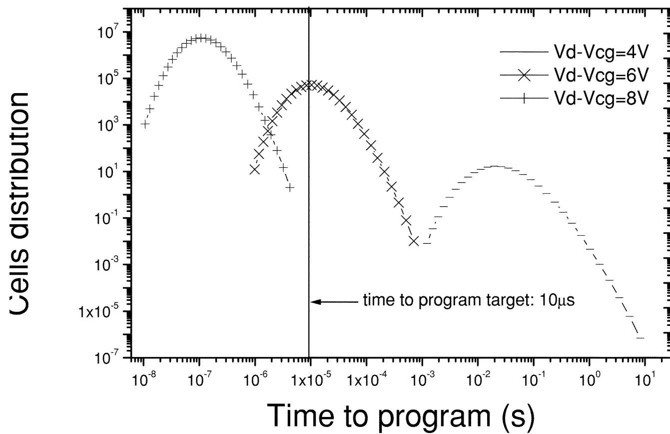

Fig. 8. The simulation of time-to-program distribution under different bias conditions $(V_d - V_{cg})$. The cell coupling ratio has a Gaussian distribution with its mean equal to 0.25 and the standard deviation equal to 0.03.

to better programming speed uniformity and, intuitively, lower coupling ratio deviation also gives tighter distribution. The effects of programming bias conditions on the cell programming distribution can also be investigated by the analytical model. As shown in Fig. 8, distributions of time to program are plotted for different values of $(V_d - V_{cg})$ for SST cells with the same α Gaussian distribution. As a result of different bias voltages, the time-to-program distributions are separated from each other. Another important implication from Fig. 8 is that a higher bias of $(V_d - V_{cg})$ is desirable for tightening the time-to-program distribution. Noticing a log scale of the time to program in the x axis, the bias difference has a more significant effect on the time-to-program distribution than appears in Fig. 8. This also suggests that our analytical model is consistent with the current SST design of choosing a higher V_d and lower V_{cg} during programming with the objective of improving cell programmability.

It is also noted that, for a given coupling ratio distribution and a requirement for cells to be programmed within 10 μs, as shown in Fig. 8, the bias voltage $(V_d - V_{cg})$ of 6.8 V and 6.5 V will result in 99% and 90% of the cells meeting the program specification, respectively. On the other hand, $(V_d - V_{cg})$ of 7 V will be able to program 99% of the cells within 10 μs even though the mean of α increases to 0.27. These results indicate that our analytical model can facilitate a guideline for proper choices of programming bias conditions and proper process screening requirement in high-yield manufacturing production.

VI. Conclusion

The 2-D lateral and vertical electric field distribution are analyzed for the SuperFlash SST cell programming. Hot electron injection is found to occur in the gap region where the peak lateral and vertical fields coincide. Gate current and substrate current models, derived by integration of hot electron generation and injection probability over the channel with the electron mean free path of 3 nm, showed a good correlation with experimental results. A full transient simulation of gate current for programming is developed for construction of the model for time to program that can be directly measured experimentally. This model relates the programming efficiency and uniformity to the bias conditions and some key device parameters, such as the gap region width and control-floating gate coupling ratio. Due to a nonlinear transformation between the coupling ratio and the time to program, the time to program has a non-Gaussian distribution even though a Gaussian distribution in the coupling ratio is assumed. The model predicts that a better programming speed and uniformity can be achieved by using a higher bias of $(V_d - V_{cg})$ or a lower coupling ratio. The programming speed distribution model can also provide a guideline to choose properly programming biases and define the process requirements for high manufacturing yield of memory chips.

Acknowledgment

The authors gratefully acknowledge N. Do, F. Gao, A. Levi, and S. Kianian from Silicon Storage Technology, Inc., for their technical supports and stimulating discussions.

References

[1] J. V. Houdt, G. Groeseneken, and H. E. Maes, "An analytical model for the optimization of source-side injection flash EEPROM devices," *IEEE Trans. Electron Devices*, vol. 42, pp. 1314–1320, 1995.

[2] Y. Ma, C. S. Pang, J. Pathak, S. C. Tsao, and C. F. Chang, "A novel high density contactless flash memory array using split-gate source-side-injection cell for 5 V-only application," in *1994 Symp. VLSI Technology Dig. Tech. Papers*, vol. 5A, pp. 49–50.

[3] K. Chang, W. Chen, C. Swift, J. M. Higman, W. M. Paulson, and K. Chang, "A new SONOS memory using source-side injection for programming," *IEEE Electron Device Lett.*, vol. 19, pp. 253–255, 1998.

[4] A. T. Wu, T.-Y. Chan, P.-K. Ko, and C. Hu, "A novel high-speed, 5-volt programming EPROM structure with source-side injection," in *IEDM Tech. Dig.*, 1986, pp. 584–587.

[5] Y. A. El-Mansy and A. R. Boothroyd, "A simple two-dimensional model for IGFET operation in the saturation region," *IEEE Trans. Electron Devices*, vol. ED-24, pp. 254–262, 1977.

[6] *SST Data Book Flash Memory*, 1997.

[7] S. Kianian, A. Levi, D. Lee, and Y.-W. Hu, "A novel 3 volts-only, small sector erase, high density flash E^2PROM," in *1994 Symp. VLSI Technology Dig. Tech. Papers*, vol. 6A, pp. 71–72.

[8] M. El-Banna and M. A. El-Nokali, "A simple analytical model for hot-carrier MOSFET's," *IEEE Trans. Electron Devices*, vol. 36, pp. 979–986, 1989.

[9] H. Guan, D. Lee, and G. P. Lee, "Analytical model for the programming of source side injection SST SuperFlash split-gate cell using two-dimensional analysis," *Solid State Electron.*, submitted for publication.

[10] P. K. Ko, "Hot-electron effect in MOSFET," Ph.D. dissertation, Univ. California, Berkeley, 1982.

[11] C. Hu, S. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K. W. Terrill, "Hot-electron-induced MOSFET degradation—Model, monitor, and improvement," *IEEE Trans. Electron Devices*, vol. ED-32, pp. 375–385, 1985.

[12] T. H. Ning, C. M. Osburn, and H. N. Yu, "Emission probability of hot electrons from silicon into silicon dioxides," *J. Appl. Phys.*, vol. 48, pp. 286–293, 1975.

[13] S. Tam, P.-K. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-31, pp. 1116–1125, 1984.

[14] N. Goldsman, L. Henrickson, and J. Frey, "Reconciliation of a hot-electron distribution function with the lucky electron exponential model in silicon," *IEEE Trans. Electron Device Lett.*, vol. 11, pp. 472–474, 1990.

[15] J. M. Higman, K. Hess, C. G. Hwang, and R. W. Dutton, "Coupled Monte Carlo-drift diffusion analysis of hot-electron effects in MOSFET's," *IEEE Trans. Electron Devices*, vol. 36, pp. 930–937, 1989.

[16] E. Bringuier, "High-field transport statistics and impact excitation in semiconductor," *Phys. Rev. B*, vol. 48, pp. 7974–7989, 1994.

[17] J. W. Slotboom, G. Streutker, G. J. T. Davids, and P. B. Hartog, "Surface impact ionization in silicon devices," in *IEDM Tech. Dig.*, 1986, pp. 292–295.

[18] T.-Y. Chan, P.-K. Ko, and C. Hu, "A simple method to characterize substrate current in MOSFET's," *IEEE Electron Device Lett.*, vol. EDL-5, pp. 505–507, 1984.

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.

Huinan Guan (S'98–M'03) received the B.S. degree from Tongji University, Shanghai, China, in 1994, the M.S. degree from the Chinese Academy of Science in 1997, and the Ph.D. degree from the University of California, Irvine, in 2002. Her doctoral dissertation concerned the model and optimization of the split-gate flash memory programming.

She is now a Post-Doctoral Researcher in the Department of Electrical and Computer Engineering, University of California, Irvine. Her current research interests are in high-speed device modeling and RF circuit design.

Dana Lee (M'93) received the B.S. degree in electrical engineering and computer sciences from the University of California, Berkeley, and the M.S. degree in electrical engineering from Stanford University, Stanford, CA.

He has worked since 1993 as a Device Engineer at Silicon Storage Technology (SST), Sunnyvale, CA, working on deep-submicrometer generations of the SuperFlash technology. He is currently a Principal Engineer at SST whose responsibilities include the development, optimization, and production realization of nonvolatile semiconductor memories.

G. P. Li has published over 170 research papers involving semiconductor materials, devices, technologies, polymer-based Bio-MEMS systems, RF-MEMS, and circuit systems. He worked on developing a silicon silicide molecular beam epitaxy (MBE) system while at the University of California, Los Angeles (UCLA), and then in the area of silicon bipolar VLSI technology and process-related device physics while at the IBM T. J. Watson Research Center, Yorktown Heights, NY. During his tenure as a Staff Member and Manger of the technology group at IBM, he coordinated and conducted research efforts in technology development of high-performance and scaled-dimension (0.5-μm and 0.25-μm) bipolar devices and ICs, as well as research into optical switches and optoelectronics for ultrahigh-speed IC measurements. In 1987, he also chaired the committee for defining IBM semiconductor technology for beyond the year 2000. He also led a research/development team to transfer the semiconductor chip technology to manufacturing in IBM. He joined the University of California, Irvine, in 1988 and is currently a Professor in Electrical and Computer Engineering and Director of the Integrated Nanosystems Research Facility (INRF). His current research interests include novel microelectromechanical (MEM) devices design and fabrication for RF wireless communication, biomedical and environmental sensing applications, novel materials and processes for high-speed electronic/optoelectronic device fabrication for network and wireless communication applications, novel electrooptical probing of semiconductor materials, devices and circuits for *in situ* wafer quality evaluation and ultrahigh-speed chip level testing, and design and fabrication of novel electronic/optoelectronic devices for low-power technology.

Prof. Li has received an outstanding research contribution award from IBM (1987) and and outstanding engineering professor awards from the University of California, Irvine (1997 and 2001).

Authorized licensed use limited to: Viktor Markov. Downloaded on March 31, 2009 at 20:00 from IEEE Xplore. Restrictions apply.