Analysis of the Correlation Between the Programmed Threshold-Voltage Distribution Spread of NAND Flash Memory Devices and Floating-Gate Impurity Concentration

Riichiro Shirota, Yoshinori Sakamoto, Hung-Ming Hsueh, Jian-Ming Jaw, Wen-Chuan Chao, Chih-Ming Chao, Sheng-Fu Yang, and Hideki Arakawa

***Abstract*—The effect of the activated floating-gate (FG) impurity concentration on the programmed threshold-voltage (V_t) distribution was newly investigated and analyzed. The lower FG impurity concentration leads to a wider threshold-voltage distribution, which is explained by the time-dependent tunnel-oxide electric-field enhancement effect induced by the reduction of the depletion region in the FG as the programming time is lengthened. Initially, the FG is deeply depleted at the interface of the tunnel oxide. However, as the programming time is prolonged, electrons by Fowler–Nordheim (FN) tunneling in the FG generate electron–hole pairs, and generated holes are gathered at the interface of the tunnel oxide, which reduces the depletion region, and enhance the oxide electric filed. The enhancement effect of the electric field for the tunnel oxide is coupled to the FN tunneling statistics and enlarges the distribution of the programmed V_t. This effect is more clearly observed at the lower FG impurity concentration, which gives the limitation of the minimum impurity concentration in FG. Monte Carlo simulations considering both the tunnel-oxide electric-field enhancement effect and FN tunneling statistics were carried out and showed good agreement with the experiments.**

***Index Terms*—Electron injection statistics, Flash memory devices, Fowler–Nordheim (FN) tunneling, program efficiency, semiconductor device modeling.**

I. Introduction

In NAND Flash memory devices, the Fowler–Nordheim (FN) tunneling mechanism has been used for programming and erasing operation in order to realize high reliability and low power consumption, which enable more than 4-kB page programming at once [1]. Recently, the multilevel (2 bit or 3 bit/cell) technology has become indispensable to achieve high-density NAND Flash memory devices [2]. For multilevel NAND Flash memory devices, it is essential to achieve the tight threshold-voltage (V_t) distribution in each state. When the threshold-voltage distribution becomes narrower, the V_t margin between neighboring levels can be relatively wider, and higher reliability can be obtained. In order to secure the tight V_t distribution for programming, a step-up program scheme is commonly used to serve a constant FN current per each step [3], [4]. In the step-up programming method, the control gate (CG) voltage is not kept constant but increases as a staircase with a fixed step amplitude of V_{step}. After each step of the staircase, the programmed V_t of the cell is sensed and compared with a verify level. If the sensed V_t is larger than the verify level, the programming is concluded; otherwise, another CG step is applied to the memory cell. This allows the tight V_t distribution. When the applied CG voltage is kept constant, the tunnel-oxide electric field is reduced due to the increasingly stored charges in the floating gate (FG) after the program time becomes longer. However, by applying the staircase programming pulse with the fixed step amplitude, the tunnel-oxide electric field can be kept constant and results in the almost constant FN current from the substrate to the FG in each step. Therefore, the V_t increase per each short pulse is almost constant, and V_t increases proportional to the number of steps due to the constant FN-current programming steps. Moreover, the V_t increase per each step is nearly equal to the CG step-up height. This result comes from the FN-tunneling programming characteristics [5]. For ideal case, the V_t distribution width should become the same as V_{step}. However, the real V_t distribution width becomes larger than V_{step}, due to the noise during programming and reading. The main program noise is the FN tunneling statistics [6]–[9], and the main read noise is the random telegraph noise (RTN) [10]–[13]. Due to the low probability of electron tunneling, the FN tunneling probability is considered to follow sub-Poisson statistics [7]. The effect of electron tunneling current fluctuation is becoming serious as cell size scaling due to the reduction of the capacitance between the CG to the FG. The V_t deviation from the neutral one is expressed as the injected charge in the FG divided by FG–CG capacitance $\delta V_t = \delta Q_{\mathrm{FG}}/C_{\mathrm{PP}}$, where Q_{FG} is the charge in the FG and C_{PP} is the capacitance between the FG and the CG. Therefore, when the capacitance of C_{PP} is reduced, the V_t variation becomes larger, i.e., proportional to the inverse of C_{PP}. In this paper, we newly investigate the effect of the FG impurity (phosphorus) concentration on the

Manuscript received December 4, 2010; revised May 7, 2011 and June 26, 2011; accepted July 26, 2011. Date of publication September 22, 2011; date of current version October 21, 2011. The review of this paper was arranged by Editor S. Deleonibus.

R. Shirota is with the Department of Electrical Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan (e-mail: rshirota@faculty.nctu.edu.tw).

Y. Sakamoto, H.-M. Hsueh, J.-M. Jaw, W.-C. Chao, C.-M. Chao, S.-F. Yang, and H. Arakawa are with the Powerchip Technology Corporation, Hsinchu 300, Taiwan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2011.2165073

programmed ΔV_t distribution. We show that the V_t distribution width after programming has strong dependence on the FG impurity active doping concentration, and the V_t distribution width becomes wider at the lower active FG doping. Here, the definition of doping is the same as the active phosphorus doping in the FG as a matter of practical convenience. It was reported that the erase efficiency has both temperature and erase pulsewidth dependence, which was explained by the FG polysilicon deep-depletion effect during erasing [14]. In the case of high temperature and long erase pulsewidth, the thermally generated holes are stored in the FG. These holes reduce the depletion width in the FG so that the erase efficiency is recovered. The deep-depletion effect in FG also needs to be considered for programming. The V_t distribution broadening at the lower phosphorus doping closely relates to the time-dependent variation of band bending in the FG depletion region at the interface of the tunnel oxide during programming. At the early stage of the one-shot programming pulse, a wide band bending in FG occurs and reduces the electric field of the tunnel oxide, which results in the reduction of the program efficiency. However, as the programming time becomes longer, the FN tunneling electrons in the FG generate electron–hole (e–h) pairs in the FG; these holes form an inversion layer at the tunnel-oxide interface, and the amount of the band bending is reduced. This mechanism of the e–h pair generation in the FG for programming is different from that of erasing. In the case of erasing, only the thermally activated e–h generation occurs. However, in the case of programming, the tunneling electrons into the FG become hot and then create e–h pairs.

This reduction of depletion width makes the tunnel-oxide field reduction slower. As FG doping is lowered, the band bending in the FG becomes larger. Therefore, the tunnel-oxide field enhancement effect appears more clearly than that with the higher FG doping. This feedback mechanism in the case of the lower phosphorus doping causes the broadening of programmed V_t distribution. In this paper, we propose the analytical model to explain these phenomena considering both effects of FN tunneling statistics and oxide field enhancement by the hole amassment in the FG. A numerical calculation was done by the Monte Carlo method, which showed good agreement with the measured data. It has been well known about the relationship between phosphorus concentration in the FG and cell reliability. Generally, the high concentration of phosphorus must be poured in the FG in order to get the active high doping state, which degrades the tunnel-oxide quality [15], [16]. Therefore, the phosphorus dosage in the FG must be carefully designed, considering both two important effects of the reliability and the V_t distribution margin.

II. Experimental Data

We conducted an experimental investigation of the V_t distribution spread using the step-up programming scheme, as shown in Fig. 1. The measurements of the programmed V_t distributions have carried out by using the 40-nm-feature-size 16-Gbit NAND Flash device.

Fig. 2 shows the V_t distributions of one page (4 kB), which were arrayed in the same word line and programmed at the same time using the step-up programming, where the step-up height of V_{step} is constant and the program pulsewidth of τ_s is 20 μs. V_t distributions with the higher and lower FG phosphorus doping are compared. For a fair comparison, exactly the same fabrication processes were done except the FG doping, and exactly the same program sequence, which consists of the initial CG voltage, step-up height, pulsewidth, and verify level, was used. The V_t distribution with the lower FG doping is wider than that of the higher one. Fig. 3 shows the V_t transients of several cells applying the step-up programming pulses from the

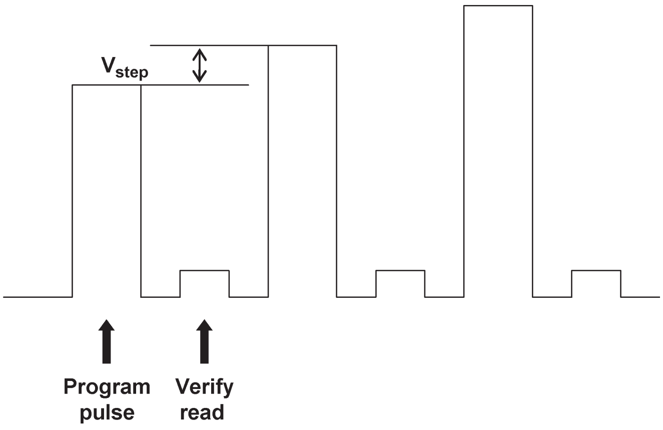


Fig. 1. Program sequence of the step-up programming. Step-up programming pulse with verify read is applied to the 40-nm-rule NAND device. The staircase program pulse is applied to the CG where pulsewidth is fixed at 20 μs. The verify read sequence is inserted to monitor if each cell Vt exceeds the verify level or not, i.e., between two program pulses.

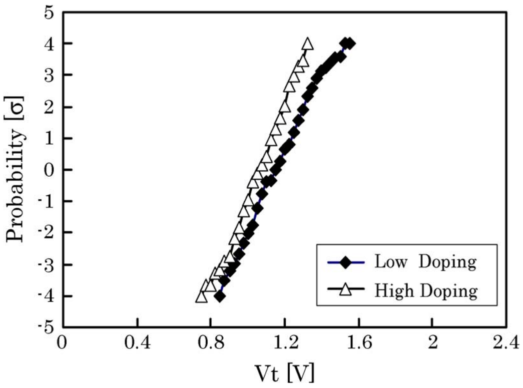


Fig. 2. One-page V_t distribution with verify programming is shown with a fixed V_{step} value. NAND cell array with low FG phosphorus doping shows a wider V_t distribution than that with high doping.

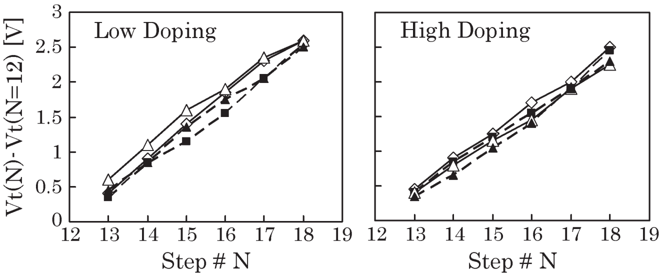


Fig. 3. Example of ΔV_t transients measured on the 40-nm-rule NAND cells. Left- and right-hand graphs show the data using low and high FG phosphorus doping, respectively.

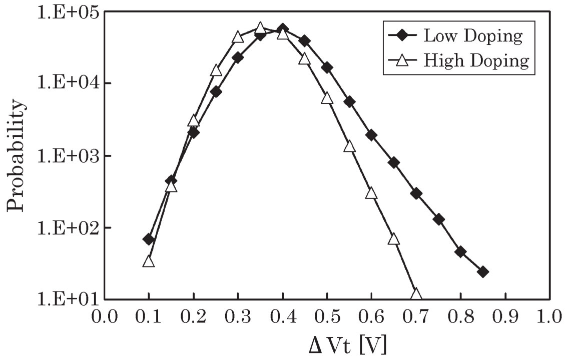


Fig. 4. Bit-by-bit V_t transient (ΔV_t) distribution from i to $i+1$ staircase programming pulses. ΔV_t distribution sampling points are accumulated by adding the data of $i = 12 \sim 17$. V_{step} is 400 mV. The NAND cell array with low FG phosphorus doping shows wider ΔV_t distribution than that with high doping.

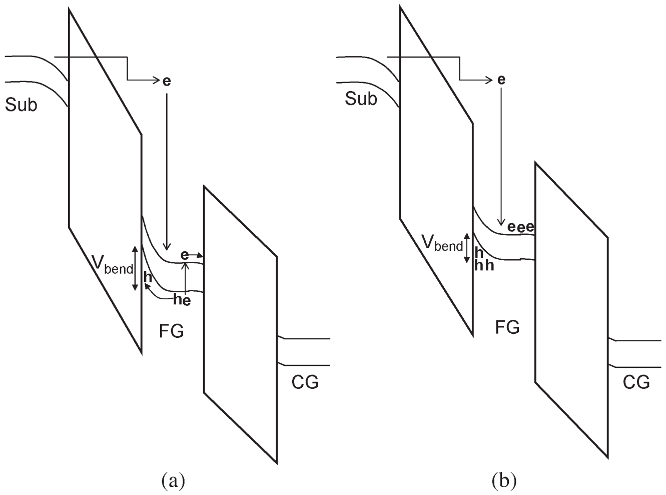


Fig. 5. Schematic 1-D band diagram for the NAND cell during programming. (a) Band diagram at the initial stage of programming and the process of FN tunneling currents that generate e–h pairs. Generated holes move to the tunnel-oxide interface, and generated electrons move to the interface of ONO. (b) Band diagram after the passage of programming time. The inversion layer of holes has formed and degenerated the deep-depletion region (Band bending height becomes smaller, as compared with that of initial case).

12th to the $(12+i)$th, where $V_t(j)$ means the V_t after the jth number of program steps. The average increment of the V_t in both FG doping is almost the same as V_{step}. Fig. 4 shows the cumulative ΔV_t distributions in one page, where ΔV_t means the V_t shift from j to $j+1$ step-up programming pulses; and $[\Delta V_t \equiv V_t(j+1) - V_t(j)]$, where j is 12–17. For the ideal case without the program noise and the read noise, ΔV_t should become almost the same as V_{step}. However, the distribution spreads wider than V_{step} due to program noise and read noise. Moreover, the ΔV_t distribution of the low FG doping shows wider spread than that of the high doping, and tail bits are observed at the higher bound in the case of the low FG doping.

In order to avoid the V_t distribution increase due to erratic bits, a NAND integrated circuit without W/E endurance is used to measure. Accordingly, as shown in Fig. 4, the data of ΔV_t distribution have no discontinuous bits, which are usually monitored when erratic bits exist. Therefore, the ΔV_t distribution characteristics should be analyzed not by tunnel-oxide quality but by statistical point of view.

III. Model for the Electron Injection Process

The constant-current FN programming of the NAND Flash device is achieved by the application of a staircase CG voltage of the V_{CG} waveform with the fixed step amplitude of V_{step} and pulsewidth τ_s. The periodic increase in V_{CG} during a ramp-programming scheme is intended to balance the field reduction that follows the electron injection into the FG, maintaining a nearly constant tunnel-oxide electric field. The dosage of phosphorus in the FG is usually in the order of $10^{20}/\text{cm}^3$, and some of them become electrically active doping. The detail of the time dependence of the FG potential is considered in each programming step. At the beginning of the Nth programming pulse duration, the tunnel-oxide interface in the FG is deeply depleted due to the large electric field at the tunnel oxide and has a large band bending, as schematically shown in Fig. 5(a). The tunneling electrons become energetic at FG and generate e–h pairs. These generated holes are gathered at the interface of the tunnel oxide. Therefore, as the programming time has become longer, gathered holes create the inversion layer and reduce the depletion width in the FG, which results in the reduction of the band bending voltage V_{bend}. Generally, after the programming time is prolonged, the charges stored in the FG reduce the tunnel-oxide electric field. However, the V_{bend} reduction makes the tunnel-oxide field reduction slower. This V_{bend} reduction is schematically shown in Fig. 5(b). Thereby, the electric-field enhancement through the tunnel oxide increases the FN tunneling current at the latter period of the Nth programming pulse duration. Before applying the next $N+1$th programming pulse, a verify read sequence is inserted. Then, generated holes during programming diffuse into the whole FG area and will almost recombine with electrons during the long verify read period. Therefore, the deep depletion in the FG repeatedly occurs at the beginning of the next $N+1$th programming pulse duration. This electric-field enhancement effect through the tunnel oxide appears more severely in the case of lower phosphorus doping in the FG because of the larger V_{bend} at the beginning of the programming pulse duration. The wider ΔV_t distribution in the lower doping FG can be analyzed by this effect. The tunnel-oxide field is expressed as

$$V_{\text{FG}} - V_{\text{bend}} = E_{\text{ox}} \cdot t_{\text{ox}} \tag{1}$$

where V_{FG} is the FG voltage at the edge near to the CG, V_{bend} is the band bending value of the FG at the interface of the tunnel oxide, t_{ox} is the tunnel-oxide thickness, and E_{ox} is the electric field of the tunnel oxide.

The electric field in the FG–CG interpoly dielectric (E_{ONO}) is expressed as

$$V_{\text{CG}} = V_{\text{FG}} + E_{\text{ONO}} \cdot t_{\text{ONO}} \tag{2}$$

where t_{ONO} is the thickness of the interpoly dielectric.

The charge stored in the FG Q_{FG} is expressed as

$$\begin{aligned} Q_{\mathrm{FG}} =& C_{\mathrm{pp}}(V_{\mathrm{FG}} - V_{\mathrm{CG}}) + C_{\mathrm{ox}}(V_{\mathrm{FG}} - V_{\mathrm{bend}}) \\ &+ C_{\alpha}(V_{\mathrm{FG}} - V_{\alpha}) \\ =& C_{\mathrm{tot}} \cdot V_{\mathrm{FG}} - C_{\mathrm{pp}} \cdot V_{\mathrm{CG}} - C_{\mathrm{ox}} \cdot V_{\mathrm{bend}} \\ &+ \mathrm{constant} \end{aligned} \tag{3}$$

where C_{ox} is the capacitance between the FG and Sub (source/drain (S/D) + channel capacitance), C_{α} is the capacitance between the FG and the other node of neighboring cells, and C_{tot} is the total capacitance between the FG and the others. Bias V_{α} at the neighboring cells has no time dependence during programming. Thus, V_{α} is treated as a constant voltage. V_{bend} is expressed as

$$V_{\mathrm{bend}} = \frac{q}{\varepsilon_{\mathrm{ox}}} \int_{0}^{X_d} x N_{\mathrm{FG}}(x) dx \tag{4}$$

where X_d is the depletion width in the FG and N_{FG} is the activated phosphorus concentration in the FG. During programming, electrons are injected from the substrate by FN tunneling, become hot electrons after passing the tunnel barrier, and arrive to the FG. These hot electrons generate e–h pairs. Auger recombination will reduce the number of holes. However, remaining holes move to the interface of the tunnel oxide and form the inversion layer, which reduces the depletion width of X_d in the FG. The generation rate of e–h pairs can be measured by using the charge separation method of the MOS transistor applying negative bias to the gate with the p-type substrate (p-substrate) grounded. The monitored p-substrate current is the net hole current, which is deducted by Auger recombination. The number of generated holes is fixed to 1.5 times FN-injected electrons J_G [17] by hypothesizing that the low-doping Si substrate case can be extended to the FG case.

Using this value, we can estimate the gathered holes per cm^2 (Q_{hole}) at the interface of oxide as

$$Q_{\mathrm{hole}} = 1.5 \int_{0}^{\tau_s} J_G \times dt. \tag{5}$$

Then, the relation between Q_{hole} and the band bending in the FG is expressed as

$$\frac{\varepsilon_{\mathrm{ox}}}{t_{\mathrm{ox}}}(V_{\mathrm{FG}} - V_{\mathrm{bend}}) = \frac{Q_{\mathrm{hole}}}{A_{\mathrm{ox}}} + q \int_{0}^{Xd} N_{\mathrm{FG}}(x) dx \tag{6}$$

where A_{ox} is the FG tunneling area faced to the S/D and the channel.

From (1)–(6), we can analyze the time-dependent charge amassment in the FG by FN tunneling current, where the tunneling current is expressed as $J_G = \mathrm{AE}_{\mathrm{ox}}^2 \cdot \exp(-\beta/E_{\mathrm{ox}})$ and A and β are fitted by the measured tunneling current. In order to estimate the ΔV_t spread for programming, we need to consider the electron tunneling statistics [5], [6]. FN tunneling is assumed to follow the Poisson statistics. Combining the effect of the band bending reduction due to the holes stored in the FG and the FN tunneling statistics, we can analyze the ΔV_t distribution of each step-up program pulse. Monte Carlo simulation was carried out where each program pulse is divided into many small segments, and the calculation is carried out by each segment. At the first segment, there is no hole in the FG. Thus, the tunnel oxide filed is derived as a function of charges in the FG and the band bending without hole amassment. In this first segment, the probability of FN tunneling follows the Poisson statistics, and Monte Carlo calculation are used to evaluate the FN tunneling probability. In the next segment, the band bending value is modified according to the generation of holes by the injected electrons in the FG at the first segment. The tunnel-oxide field has a feed back of charges in the FG as well as the band bending in FG. Then, FN tunneling probability is calculated again. These procedures are repeated.

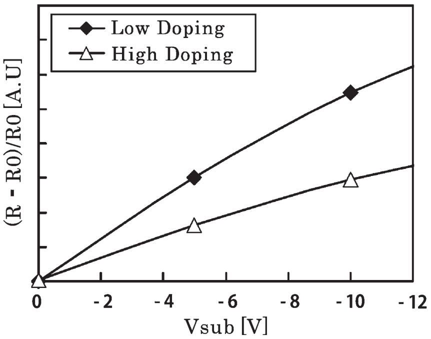


Fig. 6. Measured FG resistivity as a function of the p-substrate bias with a thick gate oxide (∼40 nm). Both FG resistances with low FG and high FG phosphorus doping are compared.

IV. Estimation of Phosphorus Concentration in the FG

In order to execute a numerical calculation of ΔV_t distribution, we must know the phosphorus doping concentration in the FG, which is electrically activated. Two methods were used to evaluate the phosphorus doping. One is the C–V measurement method. The other is to measure the gate resistance variation dependent on the p-substrate bias, as shown in Fig. 6. Both of them can monitor the gate depletion effect, where the gate materials are the same as the FG. In the case of an n-channel MOS capacitor, the inversion layer is formed, and the value of the capacitance increases when the gate bias is swept from negative to positive with the S/D grounded. However, as the gate bias is over ∼3 V, the gate at the oxide interface is depleted, and the value of the capacitance turns to be reduced. Therefore, from the C–V data, phosphorus doping can be estimated. 1-D simulation is done to calibrate the channel boron profile and is fitted to the experimental C–V data. The estimated value is also supported by the resistance measurement method. When the negative bias is applied to the p-substrate, the gate-depletion-area spread and the depletion area will not contribute to the resistor. Therefore, the gate resistance becomes larger as the p-substrate bias shifts to the negative side. From the resistance increment ratio as a function of the p-substrate bias, FG doping can be also estimated.

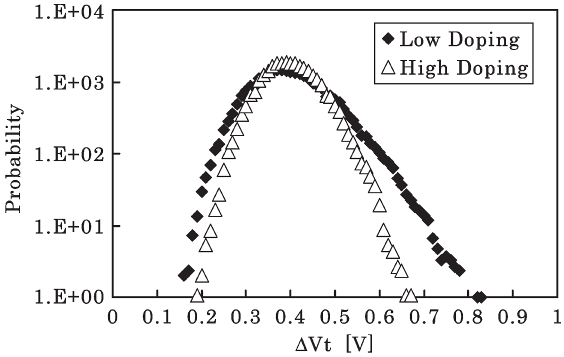


Fig. 7. Calculated ΔV_t distribution considering the effect of FN tunneling statistics in both cases with low and high FG phosphorus doping.

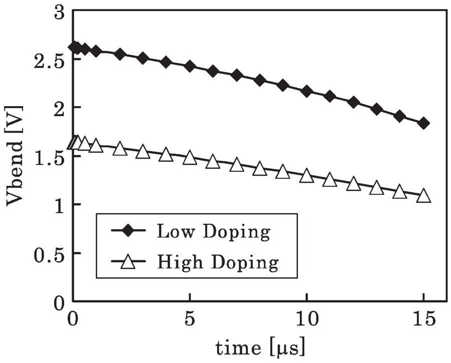


Fig. 8. Calculated time-dependent FG band bending in both cases with low and high FG doping. This graph shows the average value of the time dependence of V_{bend}.

V. Results

The calculated ΔV_t distributions with high and low FG phosphorus doping are shown in Fig. 7. It is clear that ΔV_t distribution with the low FG doping shows a wider spread in comparison with the case of the high doping, due to the tunnel-oxide electric-field enhancement effect. Fig. 8 shows the calculated result of the average V_{bend} as a function of programming time. As the programming pulsewidth becomes longer, the band bending of V_{bend} is monotonically reduced due to the hole amassment in the FG. The reduction of band bending ΔV_{bend} is larger at the low FG doping, which causes the larger tunnel-oxide electric-field enhancement effect and results in the wider ΔV_t distribution. Before the comparison of ΔV_t distribution with the experimental data, we need to taken into account the effect of the RTN. In the case of the 40-nm cell feature size, the V_t fluctuation due to RTN cannot be neglected, as shown in Fig. 9. The probability of ΔV_t considering both effects becomes

$$P_{\mathrm{tot}}(V_t) = \int P_{\mathrm{RTN}}\left(V_t'\right) \cdot P_{\mathrm{FN}}\left(V_t - V_t'\right) \cdot dV_t' \qquad (7)$$

where P_{RTN} and P_{FN} are the probability of ΔV_t variation come from RTN and FN tunneling statistics, respectively.

A. Comparison With Experiment and Calculated Results

Finally, the calculated ΔV_t distributions considering both effects of FN current statistics and RTN are obtained, as shown in Fig. 10. The calculated results agree well with experiments in both cases of the high and low FG doping, which verify the accuracy of the model. Fig. 11 shows the phosphorus doping dependence of $\sigma(\Delta V_t)$, where the feature size of the memory cell and V_{step} is fixed at 40 nm and 400 mV, respectively. In this calculation, the effect of FN tunneling statistics is considered, and the effect of RTN is not considered. The value of "high" and "low" doping used in Figs. 2–4, Figs. 6 –10 is also illustrated in Fig. 11. The ΔV_t distribution significantly spreads with the reduction of the FG phosphorus doping. This ΔV_t distribution widening mainly comes from the existence of the tail bits at the high ΔV_t side, as shown in Fig. 7. Reversely, the amount of tail bits at the high ΔV_t side can be reduced as the FG doping increases. It means that, when the FG doping becomes heavier,

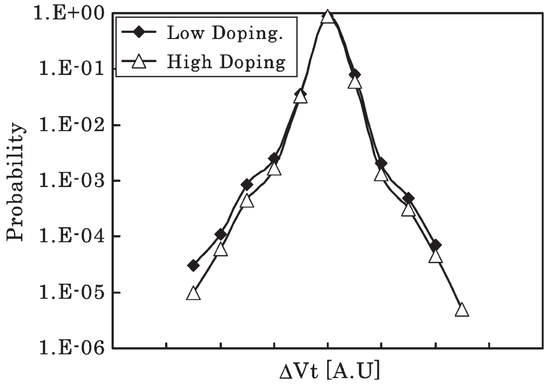


Fig. 9. Measured RTN noise amplitude of a 40-nm-feature-size NAND cell array. Noise levels between low and high FG phosphorus doping have no difference.

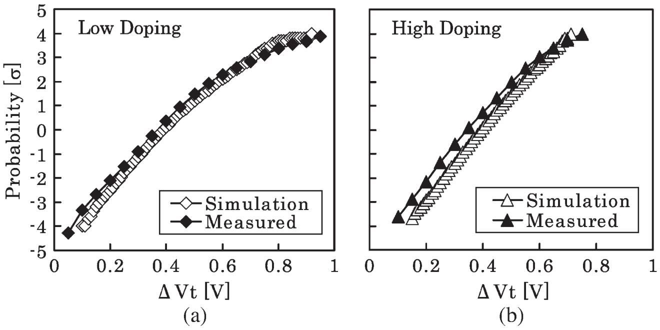


Fig. 10. Measured and calculated ΔV_t distributions. Results refer to the 40-nm-feature-size NAND technology. Both effects of FN tunneling statistics and RTN are considered for the calculation.

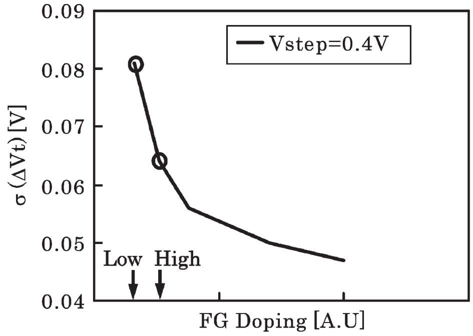


Fig. 11. Calculated $\sigma(\Delta V_t)$ as a function of FG phosphorus doping. V_{step} is fixed to 400 mV.

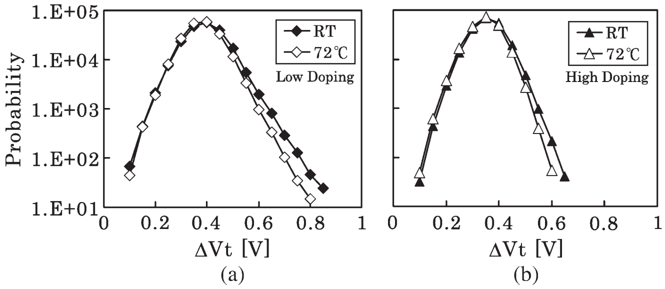


Fig. 12. Measured ΔV_t distribution with different temperature in both cases with low and high FG phosphorus doping.

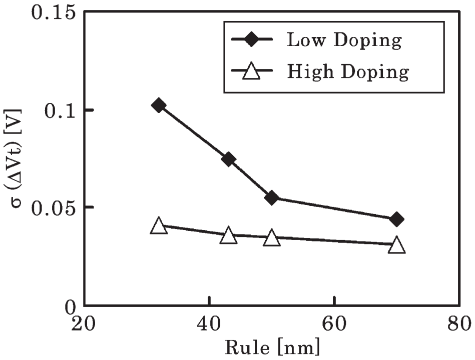


Fig. 13. Calculated $\sigma(\Delta V_t)$ as a function of technology nodes. $\sigma(\Delta V_t)$ values with low and high FG phosphorus doping are compared, where V_{step} is fixed 400 mV.

the σ plot of the ΔV_t distribution shows clearer linearity, as shown in Fig. 10. This result provides one new guide line for the design of NAND cell process/integration.

B. Temperature Dependence of ΔV_t

Fig. 12 shows a measured ΔV_t distribution in both high- and room-temperature cases. The ΔV_t distribution in high temperature becomes narrower in comparison with that in room temperature, i.e., in both cases of the low and high FG doping. At high temperature, the thermal generation of e–h pairs in the depletion region of the FG occurs, which enhances not only the amassment of holes in the FG but also the generation of e–h pairs by FN tunneling electrons. Therefore, the reduction of the depletion region is accelerated and makes the ΔV_t distribution tighten [14]. The ratio of ΔV_t distribution width reduction at high temperature is slightly larger at the lower FG doping case. The wider depletion region at the lower FG doping will help the thermal e–h pair generation. Therefore, the ΔV_t distribution widening by the tunnel-oxide electric-field enhancement effect becomes relatively smaller at high temperature.

VI. Scaling Projections

Generally, the miniaturization of the NAND cell involves the reduction of C_{pp}, which determines an increase in $\sigma(\Delta V_t)$. Furthermore, the tunnel-oxide electric-field enhancement effect at the lower FG impurity doping accelerates the increase in $\sigma(\Delta V_t)$ as the cell size scaling, as shown in Fig. 13, where the effect of FN tunneling statistics is considered. The rapid increase in $\sigma(\Delta V_t)$ resulting from the figure introduces a new reliability constraint to the design of the V_t levels of the future NAND technologies. The upper limit of the impurity doping will come from the tunnel-oxide reliability degradation. Conversely, the lower limit may come from this ΔV_t distribution spread.

VII. Conclusion

This paper has investigated the effect of the activated FG impurity doping on the NAND Flash programmed V_t distribution and provided the accurate physical model to explain the phenomena. NAND cells show the wider threshold-voltage distribution as FG impurity doping lowers. The phenomenon is explained by the time dependence of the band bending in the FG. At the beginning of the programming pulse applied, the FG is in deep-depletion condition. However, as the programming time is prolonged, FN tunneling electrons generate e–h pairs, and then generated holes form the inversion layer, which reduces the depletion region and enhances the oxide electric filed. Therefore, FN tunneling current recovered as electrons are accumulated in the FG. The NAND cells with the lower impurity doping have a wider depletion region, and the tunnel-oxide field enhancement effect appears more severely. Monte Carlo simulation considering the tunnel-oxide electric-field enhancement effect and FN tunneling statistics were carried out and showed good agreement with the experiments. The analysis of V_t distribution broadening strongly correlates with FG impurity doping and gives the new aspect for the NAND cell design.

Acknowledgment

The authors would like to thank S. Lu, S. Pittikoun, C.-M. Chen, Y.-T. Shen, and T.-K. Teng and all members of the R&D business group in Powerchip technology corporation for the discussions and supports.

References

[1] R. Kirisawa, S. Aritome, R. Nakayama, T. Endoh, R. Shirota, and F. Masuoka, "A NAND structured cell with a new programming technology for high reliable 5 V-only Flash EEPROM," in *VLSI Symp. Tech. Dig.*, 1990, pp. 129–130.

[2] K. Imamiya, H. Nakamura, T. Himeno, T. Yamamura, T. Ikehashi, K. Takeuchi, K. Kanda, K. Hosono, T. Futatsuyama, K. Kawai, R. Shirota, N. Arai, F. Arai, K. Hatakeyama, H. Hazama, M. Saito, H. Meguro, K. Conley, K. Quader, and J. J. Chen, "A 125-mm^2 1-Gb NAND Flash memory with 10-MByte/s program speed," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1493–1501, Nov. 2002.

[3] G. J. Hemink, T. Tanaka, T. Endoh, S. Aritome, and R. Shirota, "Fast and accurate programming method for multi-level NAND EEPROMs," in *VLSI Symp. Tech. Dig.*, 1995, pp. 129–130.

[4] K. D. Suh, B. H. Suh, Y. H. Lim, J. K. Kim, Y. J. Choi, Y. N. Koh, S. S. Lee, S. C. Kwon, B. S. Choi, J. S. Yum, J. H. Choi, J. R. Kim, and H. K. Lim, "A 3.3 V 32 Mb NAND Flash memory with incremental step pulse programming scheme," in *Proc. ISSCC Tech. Dig.*, 1995, pp. 128–129.

[5] A. Kolondy, S. T. K. Nieh, B. Eitan, and J. Shappir, "Analysis and modeling of floating-gate EEPROM cells," *IEEE Trans. Electron Devices*, vol. ED-33, no. 6, pp. 835–844, Jun. 1986.

[6] C. M. Compagnoni, A. S. Spinelli, R. Gusmeroli, A. L. Lacaita, S. Beltrami, A. Ghetti, and A. Visconti, "First evidence for injection statistics accuracy limitations in NAND Flash constant current Fowler–Nordheim programming," in *IEDM Tech. Dig.*, 2007, pp. 165–168.

[7] C. M. Compagnoni, A. S. Spinelli, R. Gusmeroli, S. Beltrami, A. Ghetti, and A. Visconti, “Ultimate accuracy for the NAND Flash program algorithm due to the electron injection statistics,” *IEEE Trans. Electron Devices*, vol. 55, no. 10, pp. 2695–2702, Oct. 2008.

[8] C. Friederich, J. Hayek, A. Kux, T. Muller, N. Chan, G. Kobernik, M. Specht, D. Richter, and D. Schmitt-Landsiedel, “Novel model for cell-system interaction (MCSI) in NAND Flash,” in *IEDM Tech. Dig.*, 2008, pp. 831–834.

[9] K. Prall and K. Parat, “25 nm 64 Gb NAND technology and scaling challenges,” in *IEDM Tech. Dig.*, 2010, pp. 102–105.

[10] K. S. Ralls, W. J. Skocpol, L. D. Jackel, R. E. Howard, L. A. Fetter, R. W. Epworth, and D. M. Tennant, “Discrete resistance switching in submicrometer silicon inversion layers: Individual interface traps and low-frequency (1/f?) noise,” *Phys. Rev. Lett.*, vol. 52, no. 3, pp. 228–231, Jan. 1984.

[11] H. Kurata, K. Otsuga, A. Kotabe, S. Kajiyama, T. Osabe, Y. Sasago, S. Narumi, K. Tokami, S. Kamohara, and O. Tsuchiya, “The impact of random telegraph signals on the scaling of multilevel Flash memories,” in *VLSI Symp. Tech. Dig.*, 2006, pp. 140–141.

[12] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, “Anomalously large threshold voltage fluctuation by complex random telegraph signal in floating gate Flash memory,” in *IEDM Tech. Dig.*, 2006, pp. 491–494.

[13] A. Ghettii, C. M. Compagnoni, A. S. Spinelli, and A. Visconti, “Comprehensive analysis of random telegraph noise instability and its scaling in deca-nanometer Flash memories,” *IEEE Trans. Electron Devices*, vol. 56, no. 8, pp. 1746–1752, Aug. 2009.

[14] A. Spessot, C. M. Compagnoni, F. Farina, A. Calderoni, A. S. Spinelli, and P. Fantini, “Effect of floating-gate polysilicon depletion on the erase efficiency of NAND Flash memories,” *IEEE Electron Device Lett.*, vol. 31, no. 7, pp. 647–649, Jul. 2010.

[15] H. Watanabe, S. Aritome, G. J. Hemink, T. Maruyama, and R. Shirota, “Scaling of tunnel oxide thickness for Flash EEPROMs realizing stress-induced leakage current reduction,” in *VLSI Symp. Tech. Dig.*, 1994, pp. 47–48.

[16] T. Kubota, K. Ando, and S. Muramatsu, “The effect of the floating gate/tunnel oxide interface on Flash memory data retention reliability,” in *Proc. IRPS*, 1996, pp. 12–16.

[17] C. Chang, C. Hu, and R. W. Broderson, “Quantum yield of electron impact ionization in Silicon,” *J. Appl. Phys.*, vol. 57, no. 2, pp. 302–309, Jan. 1985.

[18] B. Yu, D. H. Ju, W. C. Lee, N. Kepler, T. J. King, and C. Hu, “Gate engineering for deep-submicron CMOS transistors,” *IEEE Trans. Electron Devices*, vol. 45, no. 6, pp. 1253–1262, Jun. 1998.

Riichiro Shirota was born in Hyougo, Japan, in 1954. He received the B.S., M.S., and Ph.D., degrees from Nagoya University, Nagoya, Japan, in 1977, 1979, and 1982, respectively, all in physics.

From 1982 to 2004, he was with Toshiba Corporation, Kawasaki, Japan, where he engaged in the research and development on the DRAM, EEPROM, and Flash memory devices. In 1987, he started to develop NAND Flash memory devices. He has been a Technical Supervisor with the Research and Development Center, Toshiba Corporation. He has developed NAND Flash memory devices from 0.7-μm to 32-nm design rule. From 2004 to 2006, he was the Chief Specialist of the Flash business strategy development with the Memory Division, Toshiba Corporation. In 2005, he became a Guest Professor with the Beijing Institute of Technology, Beijing, China. In 2006, he became a Chair Professor with National Tsing Hua University, Hsinchu, Taiwan. In 2010, he has moved with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu. His current research interests include the scaling of Flash memory devices and the physical modeling of the reliability issues of memory cells and systems of solid-state drives using NAND Flash memory devices.

Dr. Shirota was the recipient of the Ichimura Award in Japan for his contribution in large-scale NAND Flash development and pioneering its application in 2000 and the Tanashi Award of the Japanese Electro-Chemical Society entitled “Development of the Giga bit NAND Flash Memory.”

Yoshinori Sakamoto was born in Nagano, Japan, in 1969. He received the B.S. and M.S. degrees in physics from Tokyo Science University, Tokyo, Japan, in 1992 and 1994, respectively.

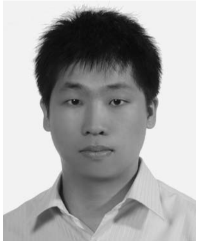
In 1994, he joined the Device Development Center, Hitachi Ltd., Tokyo. From 1994 to 2003, he engaged in the development of Flash memory circuit design and devices. From 2003 to 2009, he was with Renesas Technology Corporation, Tokyo, and engaged in the design and development of NAND Flash memory devices. From 2009 to 2010, he was with Power Flash Japan, Tokyo, and has engaged in the product development of Flash memory devices. In April 2010, he joined Powerchip technology Corporation, Hsinchu, Taiwan, and has been engaged in the product development of Flash memory devices as a Technical Manager.

Hung-Ming Hsueh was born in Hsinchu, Taiwan, in 1974. He received the B.S. degree in electronics from Vanung University, Taoyuan, Taiwan, in 1995 and the M.S. degree from Minghsin University, Hsinchu, in 2005.

In 1996, he joined the Department of Foundry Product, United Microelectronics Corporation, Hsinchu. From 1996 to 2005, he engaged in the development of memory and logic products. From 2005 to 2010, he was with the Powerchip Semiconductor Corporation, Hsinchu, and engaged in the development of Flash memory devices. In 2010, He joined Powerchip Technology Corporation, Hsinchu, and has been engaged in the product development of Flash memory devices as a Technical Deputy Manager.

Jian-Ming Jaw was born in Taipei, Taiwan, in 1972. He received the B.S. and M.S. degrees in physics from Fu Jen Catholic University, New Taipei, Taiwan, in 1995 and 1997, respectively.

In 1999, he joined the Department of DRAM Integration, Winbond Electronic Corporation, Hsinchu, Taiwan. From 1999 to 2001, he engaged in the development of DRAM memory integration and process. From 2001 to 2004, he engaged in the development of NOR Flash memory integration and process. From 2004 to 2006, he joined the Powerchip Semiconductor Corporation, Hsinchu, and engaged in the development of NAND Flash memory devices. From 2006 to 2008, he engaged in the development of the SONOS/SST NOR/PRAM Flash memory integration and process. From 2008 to present, he has engaged in the product development of NAND Flash memory devices as a Technical Section Manager.



Wen-Chuan Chao was born in Hsinchu, Taiwan, in 1979. He received the B.S. degree in electrical engineering from Chung Hua University, Hsinchu, and the M.S. degree in electrical engineering from National Chiao Tung University, Hsinchu.

In 2008, he joined the Powerchip Technology Corporation, Hsinchu, and has been engaged in the product development of Flash memory devices.

Chih-Ming Chao was born in Taipei, Taiwan, in 1971. He received the B.S. and M.S. degrees in physics from the National Sun Yat-Sen University, Kaohsiung, Taiwan, in 1994 and 1996, respectively.

In 1996, he joined the Department of Physics, National Sun Yat-Sen University, as a Teaching Assistant. From 1997 to 2001, he was a Device Engineer for DRAM development with Mosel Vitelic Inc., Hsinchu, Taiwan, and then transferred to ProMOS Technologies Inc., Hsinchu, as a Process Integration Engineer for eDRAM/6T-SRAM front-end CMOS process development. From 2002 to 2005, he was with the Device Technology Development Division, Winbond Electronics Corporation, Hsinchu, where he worked on the characterization and modeling of DRAM, LCD drivers, and Flash memory cells. In 2005, he joined the Technology Development Division II, Powerchip Technology Corporation, Hsinchu. He is currently the Deputy Manager with the Department of Device Technology I, Powerchip Technology Corporation, and responsible for CMOS device and nonvolatile memory device development.

Sheng-Fu Yang was born in Tainan, Taiwan, in 1976. He received the B.S. and M.S. degrees from I-Shou University, Kaoshiung, Taiwan, in 1999 and 2001, respectively.

In 2003, he joined the Device Group, Powerchip Technology Corporation, Hsinchu, Taiwan, where he was engaged in the development and characterization of high-power devices. His research interests include the TCAD simulation of Si MOSFET and NAND Flash memory devices.

Hideki Arakawa was born in Yamagata, Japan, in 1951. He received the B.S. degree in physics from Tokyo Metropolitan University, Tokyo, Japan, in 1980.

He has about 30 years of experience on nonvolatile memory design. In 1972, he began his career with Fujitsu Laboratory, Fujitsu Corporation, Japan. In 1987, he joined the Memory Division, Sony Corporation, and worked for developing EEPROM and Flash memory devices. In 2000, he moved again to Memory Division, Fujitsu Corporation, and in 2003, that division and the Flash Memory Division, AMD, were merged as Spansion. Prior to joining Power Chip Technology Corporation (PTC), Hsinchu, Taiwan, in 2005, he was working with Spansion as a Director for developing NAND and Mirror-Bit products. He worked for the development of GaAs MESFETs and EEPROM devices. He is currently a Program Director with PTC and a Technical Advisor with PowerMemory Inc., Japan. During his 30 years, he has been responsible for nonvolatile memory product design, technology development, and future product architecture and strategy. He has the experience of developing MNOS, Flotox and DINOR-type EEPROM, NOR and NAND Flash, and Mirror-Bit products. His current research interest include Flash memory design and reliability.

Mr. Arakawa is a member of IEEE Solid-State Circuits Society and Electron Devices Society.