Extended Abstracts of the 1995 International Conference on Solid State Devices and Materials, Osaka, 1995, pp. 67-69

High-Reliability Programming Method Suitable for Flash Memories of More Than 256 Mb

Naoki Miyamoto, Takayuki Kawahara*, Syun-ichi Saeki, Yusuke Jyouno**, Masataka Kato**, and Katsutaka Kimura*

Hitachi Device Engineering, Co., Ltd., Mobara, Chiba 297, Japan.
* Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo 185, Japan.
** Semiconductor & Integrated Circuits Division, Hitachi Ltd., Kodaira, Tokyo 187, Japan.

Both the variable word-line voltage programming (VVP) method and the VVP method with variable pulse width (VVWP) make it possible to achieve high reliability with a sufficient disturb margin while maintaining high-speed programming. A simulation shows that both methods reduce the maximum Fowler-Nordheim tunnel current density by 1.4 orders of magnitude compared to the conventional method with a programming time of about 1 ms. This is expected to triple the charge-to-breakdown.

1. Introduction

For a single low-voltage flash memory, incremental step pulse programming method has been proposed that use Fowler-Nordheim (F-N) tunneling and a variable word-line voltage [1]. Although this method enables to achieve high-speed programming and decrease the threshold voltage distribution, reliability performance is not discussed. As the program/erase endurance increases, the thin tunnel oxide is expected to degrade. Recently, it has been shown that the charge-to-breakdown (Qbd) in the intrinsic breakdown region decreases as the tunnel area increases [2]. Therefore, as the density increases, it is possible that Qbd at a 99.99% yield falls below the injection charge (Qinj) of the tunnel oxide after 10^6 cycles of endurance as shown in Fig. 1.

This paper, therefore, discusses reliability performance (increased Qbd by reducing the F-N tunnel current density) and programming characteristics by using a variable word-line voltage programming method.

2. Concept of variable word-line voltage for programming

Figure 2 shows the concept of variable word-line voltage for programming. To increase Qbd, F-N tunnel current density is reduced[3] by decreasing word-line voltage. However, this increases the constant word-line voltage programming time in the conventional method [4]. In the method 1, the first program word-line voltage (|Vcg1|) is set lower than the last (|VcgL|) with a constant programming pulse width of Δtpw. Here, |VcgL| is determined by the disturb margin. The program/verify-read sequence is repeated N times. Then |Vcg1| increases to |Vcg2| (< |VcgL|) and the program/verify-read sequence is again repeated N times. When the word-line voltage reaches |VcgL|, the program/verify-read sequence is repeated under constant voltage. The method 1 reduces the maximum F-N tunnel current density by about one or two orders of magnitude by reducing the electric-field (Eox) of the tunnel oxide. However the constant pulse width method leads to a larger number of programming verifications. The method 2 introduces the variable pulse width method in |VcgL| to reduce the number of verifications.

3. Simulated programming method

The effect of these methods are studied by simulating the programming characteristics. A simplified equivalent circuit, as shown in Fig. 3, is used for the simulation. The

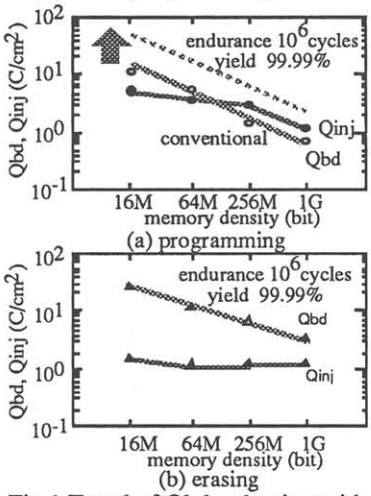


Fig.1 Trend of Qbd reduction with increasing flash memory densities

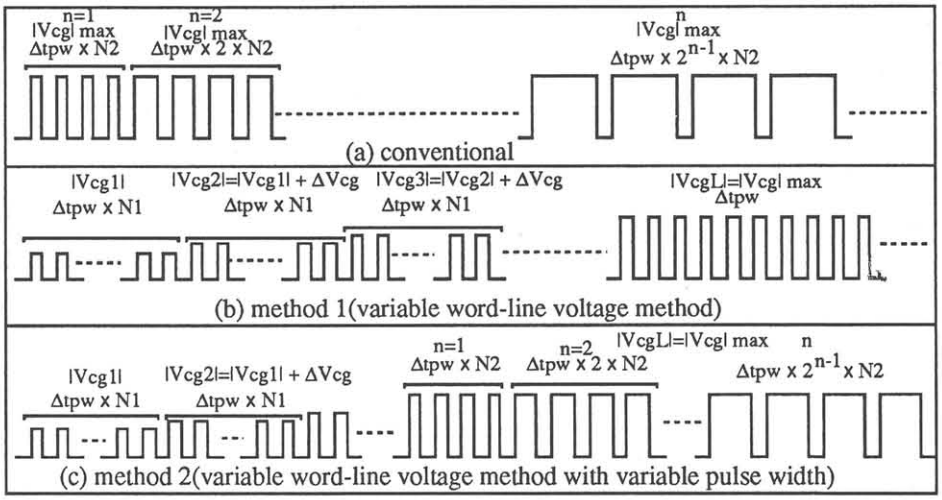


Fig.2 Concept of variable word-line voltage method for programming

tunnel current density is approximated by the F-N equation

$$J_{FN}=A \cdot Eox^2 \cdot (\exp(-B/Eox)) \quad (1)$$

where Eox is the electric field in the oxide, and A and B are constants. The tunnel oxide field Eox is given by

$$Eox=(Vedge - Vfg)/tox \quad (2)$$

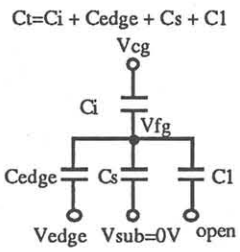
where Vfg is the floating gate voltage during the programming operation, and is given by

$$Vfg=(Ci/Ct) \cdot (Vcg - (Vth - \Delta Vth - Vthi)) + (Cedge/Ct) \cdot Vedge \quad (3)$$

where Ci is the inter poly capacitance, Ct is the total capacitance of the inter poly and tunnel oxide, Cedge is the capacitance of the edge tunnel oxide, and the threshold voltage shift ΔVth according to the program is given by

$$\Delta Vth=(1/Ci) \cdot J_{FN} \cdot A_{FN} \cdot \Delta tpw \quad (4)$$

where Δtpw is the programming time, A_{FN} is the F-N tunnel area. The simulated programming characteristics can be obtained by using the expressions in (1), (2), (3), and (4) as well as the memory cell parameters shown in Fig. 3.



Tunnel oxide	7.5 nm
Inter poly-Si dielectric	15 nm
Gate length	0.3 μm
Gate width	0.25 μm
Tunnel A (fastest)	3.36×10^{-7}
Tunnel A (typical)	3.36×10^{-8}
Tunnel A (slowest)	3.36×10^{-9}
Tunnel B	2.33×10^{-8}
Ci /Ct	0.47
Cedge /Ct	0.18
Vthi	2.0 V

Fig.3 A simplified capacitive equivalent circuit of the flash memories and memory cell parameters for simulated programming

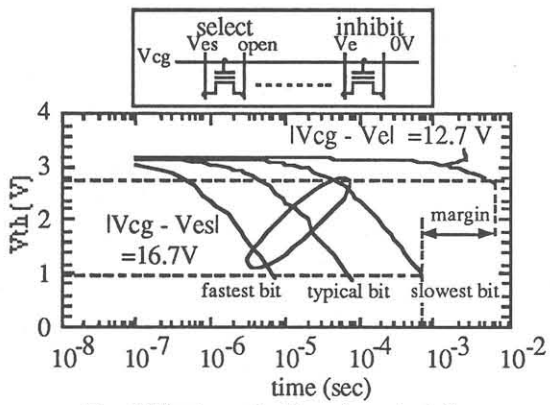


Fig. 4 Electron ejection characteristics

4. Method of setting the operating voltage

The programming voltage (|Vcg-Vesl|) is determined by the disturb margin. Figure 4 shows the electron ejection characteristics. The disturb margin is the time between the slowest bit programming time and the fastest bit disturbance time. Here, |Vcg-Vesl|=16.7 V and |Vcg-Vel|=12.7 V are the values needed to maintain the disturb margin, even though the programming time degradation in the slowest bit becomes five times the initial value after 10^6 cycles of endurance. This |Vcgl| is chosen as |VcgLl|.

5. Performance Comparison

Figure 5 shows the simulated programming characteristics. Here, |Vcg1-Vesl| is set to 13.7 V, ΔVcg=|Vcg1-Vcg2l| is 0.5 V, Δtpw is 10 μs, N1 is 8, N2 is 4. These values were selected to control the threshold voltage of the fastest bit within an accuracy ($\Delta Vthf$) of 0.2 V and suppress the number of verifications to about 100 in both methods. High-speed programming of about 1 ms can be obtained by using an operating voltage of 16.7 V with both methods. Figure 6 shows the simulated F-N tunnel current density during programming. Table 1 compares the performance. Both methods can reduce the maximum F-N tunnel current density by about 1.4 orders of magnitude compared to that of the conventional method, and the method 2 increases the number of verifications by less than the method 1. The reduced F-N tunnel current density will approximately triple Qbd compared to that of the conventional method [3] making Qbd higher than Qinj. Decreasing |Vfg-Vsubl also improves the reliability of the tunnel oxide by reducing band to band tunneling.

Figure 7 shows J_{FN} versus the total programming time (tpw). Using either proposed method allows us to decrease J_{FN} while suppressing the increase in the total programming time.

6. Conclusions

Both the variable word-line voltage programming (VVP) method and the VVP method with variable pulse width (VVWP) make it possible to achieve high reliability while maintaining the total programming time of the conventional method, and both provide a sufficient disturb margin. Simulation results show that both methods reduce the maximum F-N tunnel current density (J_{FN}) by 1.4 orders of magnitude compared to that of the conventional method and the VVWP method increases the number of verifications by less than the VVP method. This is expected to triple the Qbd. A Qbd higher than the Qinj is obtained for flash memories of more than 256 Mb.

Acknowledgments

We thank H. Kawamoto, M. Isihara, R. Hori, K. Miyazawa, E. Takeda, H. Kume, Y. Ohji, T. Tanaka, and M. Ushiyama for their suggestions.

References

[1] K. Suh et al., ISSCC95, pp.128-129, 1995
[2] A. Teramoto et al., IEICE.SDM94-38, pp.29-34, 1994
[3] P.P. Apte et al., IEEE Trans.ED, pp. 1595-1601, 1994
[4] T. Tanaka et al., 1994 Symp. on VLSI Ckt., pp. 61-62

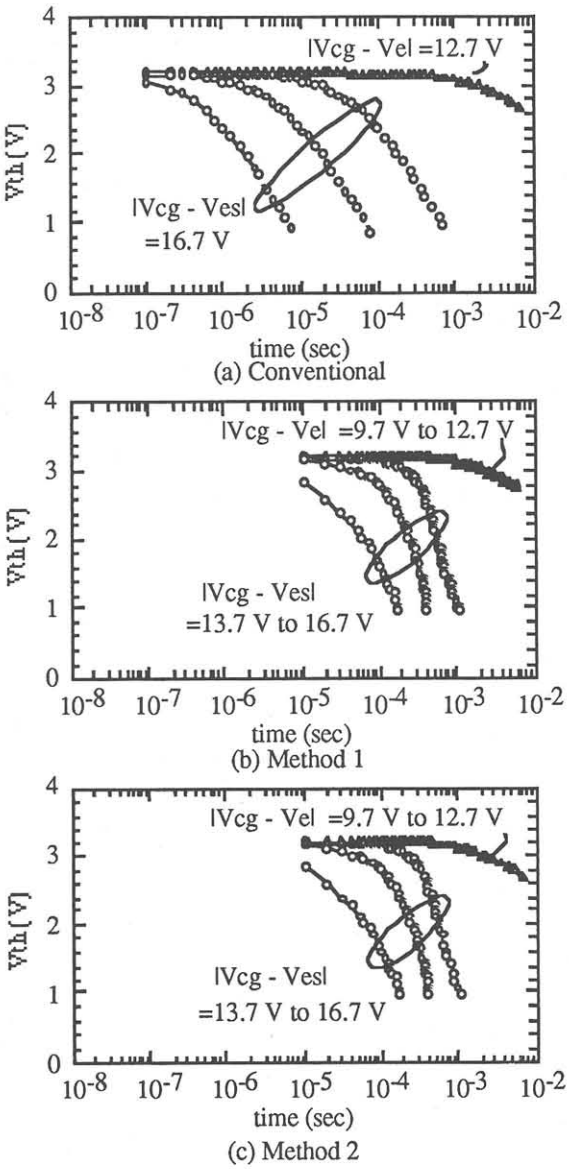


Fig.5 Simulated program characteristics

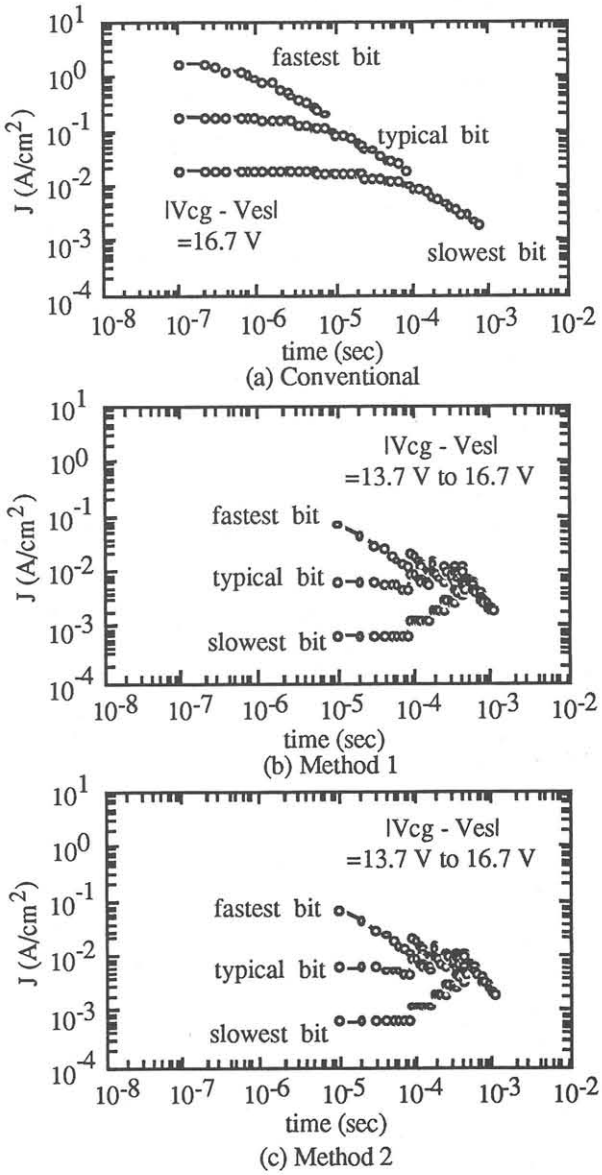


Fig.6 Simulated F-N tunnel current density

Table 1 Performance comparison

	Conventional	Method1	Method 2
Number of verifications (slowest bit)	43	105	64
Maximum F-N tunnel density J_{FN} max (fastest bit)	1.78 A/cm^2	0.07 A/cm^2	0.07 A/cm^2
Maximum electric field Eox max (fastest bit)	13.43 MV/cm	11.48 MV/cm	11.48 MV/cm
Maximum voltage between floating gate and substrate |Vfg - Vsubl max (fastest bit)	6.08 V	4.61 V	4.61 V
ΔVthf (fastest bit)	0.19 V	0.11 V	0.11 V

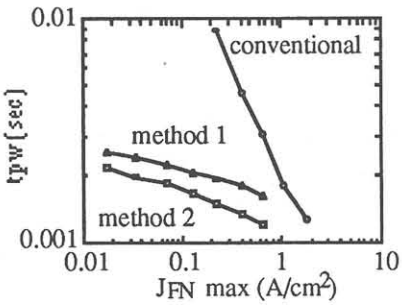


Fig.7 J_{FN} max versus total programming time
(tpw=tpw(slowest) + 100 μs + 10 μs x (Number of verifications))