Simulation, fabrication and characterization of a 3.3 V flash ZE²PROM array implemented in a 0.8 μm CMOS process

J. Ranaweera, W.T. Ng, C.A.T. Salama *

Department of Electrical and Computer Engineering, University of Toronto, 10 Kings College Road, Toronto, Ont., Canada M5S 1A4

Received 3 February 1998; received in revised form 12 June 1998

Abstract

This paper describes a Zener based flash memory cell (ZE²PROM), programmed from hot electrons generated by a heavily doped reverse biased p⁺ n⁺ junction attached to the drain. The cell can be implemented in a NOR type memory array. It uses an orthogonal write technique to achieve fast programming with low power dissipation and reduced drain disturbance. The modeling of the charge transfer behavior of the flash ZE²PROM cell is also done to describe the charging and discharging of the floating gate during programming and erasing. The flash ZE²PROM arrays were implemented in a 0.8 μm lithography CMOS process flow in which the n-LDD step was replaced with a one sided p⁺ boron implant with a doping level of ~10¹⁹ cm⁻³. This minor change to a standard CMOS process, makes the concept highly attractive for embedded memory applications. A programming time of 850 ns at 3.3 V supply was achieved on fabricated test devices. © 1998 Elsevier Science Ltd. All rights reserved.

1. Introduction

One of the major drawbacks of flash E²PROM devices which are relying on conventional channel hot electron injection for programming, is the need for an external high-voltage power supply (>5 V) in order to program within a reasonable time. In this programming method, when the channel length is reduced to the sub-half micron region, the drain voltage must also be reduced to avoid punchthrough breakdown. This limits the electric field at the drain thus reducing the hot electron generation efficiency.

The use of a reverse biased junction in avalanche to inject electrons over a silicon dioxide barrier, has been previously used in programming a nonvolatile memory device [1]. However, this programming method was thought to have deleterious effects on cell reliability and endurance due to the large junction breakdown current and voltage, and the injection of hot holes into the gate oxide. To overcome these concerns, a cell that uses a reverse biased Zener junction for programming has been developed [2–4]. The performance limitations of this cell, operating at 5 V, was first presented by the authors [2]. In this paper, we report in detail on a 3.3 V only flash memory cell and its performance in an orthogonal write array architecture which minimizes drain disturb time and power dissipation while maintaining a programming speed of 850 ns.

2. Cell structure

The flash ZE²PROM cell, illustrated in Fig. 1, has one p⁺ pocket implant extending part-way across the width of the drain region to generate hot electrons for programming. The test devices were fabricated using a 0.8 μm lithography CMOS compatible process. The starting material for this process was a (1 Ω cm) p type (100) orientation silicon wafer. Prior to the formation of the gate oxide, a sacrificial thick oxide is grown in
steam at 1000°C to remove any Kooi effect caused by the nitride layer used for LOCOS formation [5]. This oxide is etched off and a 85 Å thin high quality gate oxide is grown by using diluted oxygen. A boron implant through this oxide was used to adjust the threshold voltage of the device. A layer of amorphous silicon is then deposited on the wafers to form the floating gate. A phosphorus implant is used to dope the gate. The interpoly dielectric consisting of an oxide–nitride–oxide sandwich with 200 Å of equivalent electric thickness is then fabricated [6]. The 80 Å thin bottom oxide was fabricated in diluted oxygen at 950°C for 8 min. Then a nitride layer of ~190 Å is deposited in the LPCVD at 750°C for 10 min. A wet oxidation of the silicon nitride layer was carried out at 975°C for 35 min to obtain a 40 Å thin top oxide. A second layer of amorphous silicon is then deposited, doped and annealed to form the control gate. The gate structure is then patterned by dry etching.

Next the p⁺ region of the Zener junction is formed by a boron implant. This boron implant replaces the conventional n− LDD implant in the CMOS process. A side wall spacer (SWS) is formed and the n⁺ source and drain regions are implanted with phosphorus and annealed to obtain 0.25 μm deep junctions. The SWS is necessary to form the self aligned n⁺ source and drain regions and also to avoid counter-doping of the p⁺ pocket. Finally aluminum silicon is deposited, patterned and annealed at 450°C for 25 min in forming gas to make the contacts.

3. Programming with Zener induced hot electrons

A typical programming bias condition for the flash ZE²PROM cell is \( V_{cg} = 12 \) V and \( V_d = 3.3 \) V. The reverse biased p⁺ n⁻ junction creates hot electrons from Zener breakdown as illustrated in Fig. 2. These
hot electrons have an energy much higher than the thermal equilibrium energy and the capacitive coupled floating gate voltage, creates an electric field across the thin tunnel oxide to pull these hot electrons towards the floating gate. Electrons with sufficient energy and momentum in the direction perpendicular to the Si/SiO₂ interface, cross the potential barrier formed by the gate oxide.

To gain insight into the basic charge-transfer characteristics of the flash ZÉ²PROM structure, a capacitive equivalent circuit was developed as shown in Fig. 3. $C_{FG}$ is the capacitance between the control gate and the floating gate, $C_p$, $C_{p+}$, $C_d$ and $C_s$ are the coupling capacitances between the floating gate and the P-well, the p⁺ region, the drain and the source, respectively. When the device is operating in the program mode, the current through the thin tunnel oxide is integrated and appears as a charge $Q_{fg}$ on the floating gate. The accumulation of the charge on the floating gate is reflected as a change in the threshold voltage $V_t$ of the device. The relationship between the $V_t$ of the device and $Q_{fg}$ is given by

$$V_t = V_{ti} - \frac{Q_{fg}}{C_{FG}}$$

where $V_{ti}$ corresponds to the intrinsic $V_t$ when $Q_{fg} = 0$. The threshold shift is proportional to the charge stored on the floating gate.

Charge balance implies that

$$C_{FG}(V_{CG} - V_{FG}) = C_{d}(V_{fg} - V_d) + C_{p}(V_{fg} - V_B) + C_{p+} + (V_{fg} - V_B) + C_{d}(V_{fg} - V_d)$$

From Eq. (2), the floating gate voltage $V_{fg}$, due to capacitive coupling, can be derived for a typical bias condition applied during programming the cell and expressed as

$$V_{fg} = \frac{C_{FG}V_{CG} + C_{d}V_d}{C_{FG} + C_d + C_p} = C_{p+}$$

or
\[ V_{fg} = \frac{C_{fg}V_{gs} + C_dV_d}{C_T} \]  
(4)

where

\[ C_T = C_{fg} + C_d + C_i + C_T + C_{p^*} \]  
(5)

is the total capacitance on the floating gate. Due to the extra capacitor \( C_{p^*} \), \( V_{fg} \) due to capacitive coupling in the flash ZE\(^2\)PROM cell is lower than in a conventional flash E\(^2\)PROM cell. This condition tends to slow the performance of the device. However, the flash ZE\(^2\)PROM cell is capable of high speed programming due to the energetic hot electrons created by the high electric field at the Zener injector.

4. Simulations

4.1. Programming

In order to demonstrate the ability of the Zener induced programming method to increase the injection efficiency as compared to the conventional hot electron programming method, numerical simulations of two device structures (conventional flash cell and the ZE\(^2\)PROM cell) were performed. These simulations were done by using the TMA MEDICI’s programmable device advance application module with Fowler–Nordheim and hot carrier gate current models. Impact

![Fig. 4. The simulated transient characteristics of the flash ZE\(^2\)PROM (cell A) and a conventional flash E\(^2\)PROM (cell B) during programming. Cells A and B consist of \( L = 0.8 \, \mu m, \, G_{ox1} = 85 \, \AA \) and \( G_{ox2} = 200 \, \AA \).](image-url)
ionization and band-to-band tunneling were included in simulating the flash ZE²PROM structure to obtain the total electron current contribution to the gate current.

In order to compare the two programming methods, 0.8 μm gate lengths ($L$), ~85 Å thin gate oxides ($G_{ox1}$) and ~200 Å thick interpoly dielectrics ($G_{ox2}$) were selected for both structures. The p-type substrate doping was $6 \times 10^{16}$ cm$^{-3}$ and the n$^+$ source/drain region junction depth was 0.3 μm with a peak doping concentration of $2 \times 10^{20}$ cm$^{-3}$ in both structures. The p$^+$ region of the flash ZE²PROM cell has a peak doping of $2 \times 10^{19}$ cm$^{-3}$ with a junction depth of 0.25 μm. The junction depths and doping concentrations of the n$^+$/p$^+$ regions, dimensions of the gate structure, and the substrate doping were selected to closely resemble the device parameters of the fabricated test structures.

To initiate the transient programming operation, time steps between 1 ns to 100 μs were specified. The simulation results showing the change in $V_{fg}$, $Q_{fg}$, hot electron current ($I_{he}$), drain current ($I_d$) and the injection efficiency ($I_{fg}/I_d$) during programming are shown in Fig. 4 for a conventional flash E²PROM cell as well as for the flash ZE²PROM cell during erasing (Fig. 5).

Fig. 4. Simulated transient characteristics of a conventional flash E²PROM cell during programming.

Fig. 5. Simulated transient characteristics of the flash ZE²PROM cell during erasing.

Fig. 6. The micrograph of a flash ZE²PROM array implemented in a NOR type architecture.
as for a flash ZE²PROM cell. In Fig. 4, the flash ZE²PROM cell is referred to as cell A and the conventional cell as cell B.

The drain voltage of 3.3 V and the doping level of the p⁺n⁺ junction contribute to the electric field responsible for the hot electron generation. At the beginning of the programming operation, the voltage coupled to the floating gate is dependent on the capacitance values of the gate stack. Fig. 4(a) shows that at 50 ns, with hot electron injection, the voltage of the floating gate in cell A decreases due to the negative charging of the floating gate while the floating gate potential of cell B remains unchanged. Due to the extra p⁺ region in cell A, its initial floating gate potential is lower compared to that of cell B as expected when Eqs. (3) and (4) were deduced.

At the beginning of the injection process, the field in the oxide causes the current to flow through it. Fig. 4(b) shows how the floating gate voltage is reduced when the floating gate charges up. This condition reduces electron injection causing the gate current to drop as the electron injection process becomes self-limiting. As the floating gate becomes fully charged, the gate current is reduced almost to zero because the oxide field prevents further injection of electrons. In cell A, the hot electron current is highest at the beginning of the writing process but begins to drop after about 400 ns as the negative charge injected onto the floating gate becomes significant (≈ \(1 \times 10^{-14} \) C) as shown in Fig. 4(c) and (d). This in turn leads to a lower floating gate potential and a decrease in hot carrier injection. The flash ZE²PROM cell can be programmed in 600 ns while the conventional cell remains unprogrammed at 100 μs.

Fig. 4(e) shows that during programming, the drain current in cell B is more than an order of magnitude larger than the drain current (or Zener breakdown current) in cell A. The low drain current of ≈40 μA/μm in

---

**Fig. 7.** (a) Orthogonal programming (cells A and C) of the flash ZE²PROM array (\(I_d\) is flowing in all the cells connected to BL1) and (b) threshold voltage versus programming time.
cell A is a result of the p+ n+ junction doping, depth and the applied drain voltage. The higher magnitude of the electric field (~10^6 V/cm) created by the heavily doped p+ n+ junction and the lower drain current lead to higher injection efficiency in cell A compared to cell B as illustrated in Fig. 4(f).

The simulation results show that programming with Zener induced hot electrons is more effective for high-speed programming with 3.3 V supply as compared to cells that programmed using the conventional channel hot electron method.

4.2. Erasing

Erasing of the flash ZE²PROM cell is performed by tunneling of electrons from the floating gate to the source. This is accomplished by applying a positive potential to the source region and a negative potential to the control gate. A typical erase characteristic of the simulated flash ZE²PROM cell is shown in Fig. 5. For this operation, 3.3 V at the source and −12 V at the control gate was used. Under these bias conditions, there is a sufficiently strong electric field in the oxide between the floating gate and the source for the negative charge to tunnel through the oxide. As in the case for the programming operation, Fowler–Nordheim and gate current models as well as time steps between 1 µs to 100 ms were specified. Fig. 5(a) and (b) indicate how the floating gate potential increases when the negative charge is removed by the erase operation. The Fowler–Nordheim tunneling current is highest at the beginning of erasing but decreases as the floating gate becomes less negative as shown in Fig. 5(c). This reduces the field in the thin gate oxide and $V_t$ of the cell decreases. After ~50 ms, the stored charge in the floating gate is removed (Fig. 5d).
5. Experimental results

5.1. Array architecture and orthogonal programming

In order to verify the validity of the simulations, flash ZE²PROM NOR type memory array, using a 0.8 \( \mu \)m CMOS compatible double polysilicon process, was implemented. These test structures occupy an area of 54 \( \mu \)m². The micrograph of the memory array is illustrated in Fig. 6. The rows of the array are accessible through polysilicon lines that contact the control gates of the cells. The source and drain regions of a column are connected to two dedicated metal bit and source lines.

When programming a flash ZE²PROM array with \( n \) word lines, all the \( p^+ n^+ \) junctions that are connected to the same bit-line will be conducting \( n \times I_d \) and dissipate a power of \( (n \times V_d \times I_d) \). The bit-line current and the power dissipation are independent of the number of cells being programmed simultaneously within the bit-line. Therefore, all the required cells in one bit-line can be programmed together (orthogonal programming), before selecting the next bit-line. This results in the programming time and power used for a single flash ZE²PROM cell.

To program a cell in the memory array, the selected bit-line is connected to 3.3 V, the substrate is grounded, the metal source line is left floating and 12 V is applied to the selected word lines, as shown in Fig. 7(a). This yields a small \( I_d \) of 50 \( \mu \)A per \( \mu \)m of device width. In the fabricated test devices, it was possible to program the cells within \(~850 \) ns, as shown in Fig. 7(b). This programming speed is slower than the 600 ns expected from the simulations. This can be attributed to possible discrepancies between the model parameters used in the simulator and the ones encountered in the experimental work. However, the trends confirmed by both simulations and experimental results show a significant improvement in programming time when compared to the tens of microseconds necessary to program conventional flash cells at a 5 V supply.

5.2. Drain and gate disturb

When a bit-line in the array is connected to 3.3 V to program selected memory cells, the drain regions associated with the selected bit-line will break down gen-
erating hot electron–hole pairs. The generated hot electrons will be injected into the floating gate of the cells with control gates biased at 12 V, thus programming these cells. Simultaneously, unselected cells along the bit-line ($V_{cg} = 0$ V and $V_d = 3.3$ V) will have hot holes injected into their gate oxide (drain disturb) as illustrated in Fig. 8(a). Drain disturb characteristics, due to the $p^+n^+$ junctions, in unselected programmed and erased flash ZE^2PROM cells are shown in Fig. 8(b). The programmed flash ZE^2PROM cells exhibit a soft erase after a drain disturb time of ~500 ms while the erased cell remains undisturbed at 100 s. This is due to the fact that the floating gate potential of the programmed cell is lower than that of the erased cell leading to hot hole injection in the programmed cells. This undesired injection of hot holes can be minimized by the orthogonal programming technique. In an array with $n$ word lines, the maximum drain disturb time using orthogonal programming is $n$ times less than in a conventional word-line by word-line programming technique.

Similarly to drain disturb, gate disturb can occur when an unselected cell is exposed to high word-line voltages as other cells along the same word line are being programmed. In this condition, unprogrammed cells can have electrons tunnel from the $n^+$ drain and source regions into the floating gate increasing the threshold voltage of the cell. With the orthogonal programming technique, the cells must be able to endure the word-line stress for the total times that it takes to program all the cells along the word-line. Fig. 9 indicates a minor gate disturb in the erased

Fig. 10. Experimental (a) read current as a function of the $p^+$ pocket width and (b) soft write endurance during reading.
cells after ~10 s with 12 V applied to the gate. Another form of gate disturb occurs when an unselected programmed cell is exposed to a programming voltage on the word line. In this condition, the stored electrons can tunnel from the floating gate to the control gate through the ONO inter-poly dielectric. This results in a reduced threshold voltage for the cell. Since the equivalent oxide thickness of the ONO interpoly is ~200 Å, a very small gate disturb was observed (after 1 s) in the programmed cells. This condition is also depicted in Fig. 9.

These excellent drain and gate disturb characteristics can be attributed to the use of low $V_d$ and $I_d$ during programming and the high quality of the gate oxide and ONO in the fabricated flash ZE²PROM.

5.3. Erasing

In a fabricated flash ZE²PROM cell, a bias condition of $V_s = 3.3$ V and $V_{cg} = -12$ V is sufficient to perform the erasure within 50 ms. However, as all cells in a sector are erased at the same time, the per bit erase time becomes very short. The erase time can be further decreased by using higher voltages at the source and at the control gate and by decreasing the thickness of the gate oxide and the interpoly ONO. When selecting the source voltage, care must be taken to avoid the source to substrate junction breakdown. Without junction breakdown, there is no avalanche current to supply and the source current during erasure is very small. Therefore, the source voltage can be generated on chip without the need for an additional power supply.

5.4. Reading

Since the p⁺ region adjacent to the drain exhibits a low breakdown voltage which contributes to soft programming, the dedicated metal source-line can be used in the read operation. In this cell design, the reduced channel width available for reading also yields lower read current compared to the conventional flash memory cells. The read current dependence on the width of the p⁺ diffusion is depicted in Fig. 10(a). Soft write endurance (time to cause a threshold $V_t$ shift of 0.5 V) during the read cycles indicates a greater than 10 year lifetime at $V_s = 1$ V, as shown in Fig. 10(b).

5.5. Endurance characteristics

The number of write/erase cycles of this memory cell is shown in Fig. 11. No $V_t$ degradation is observed for the initial 10,000 cycles. However, a $V_t$ window closure is apparent at 50,000 cycles. The $V_t$ window closure of
the programmed and erased states is due to electron trapping in the gate oxide at the drain edge as well as at the source edge. The observed closure represents a significant improvement in endurance as compared to previously reported results [4].

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Micronet, Nortel, Mitel and Gennum.

6. Conclusion

The simulation and experimental results obtained on the flash ZE2PROM cell show promising performance in programming speeds at 3.3 V supply. A programming time of 850 ns was achieved on fabricated test structures. The orthogonal write architecture can be used to speed up block writes and to reduce drain disturb. This write architecture also reduces the power dissipation of the flash ZE2PROM array. The cell was shown to be immune to gate and drain disturb while achieving greater than 10 year soft write immunity for \( V_s = 1 \) V.

References