Injector Design for Optimized Tunneling in Standard CMOS Floating-Gate Analog Memories

J. Madrenas, A. Ivorra, E. Alarcón and J.M. Moreno
Departament d'Enginyeria Electrònica. Universitat Politècnica de Catalunya.
C/ Gran Capità s/n. Campus Nord. Mòdul C4. 08034 Barcelona. SPAIN
contact author: madrenas@eel.upc.es

Proceedings of the 41st IEEE Midwest Symposium on Circuits and Systems (MWSCAS98), pp. 348-351, Univ of Notre-Dame, Indiana, USA, August 1998.

©1998 IEEE. Personal use of this material is permitted. However, permission to reprint or republish this material for advertising or promotional purposes or for creating new collecting works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reported without the explicit permission of the copyright holder.

Injector Design for Optimized Tunneling in Standard CMOS Floating-Gate Analog Memories

J. Madrenas, A. Ivorra, E. Alarcón and J.M. Moreno
Departament d'Enginyeria Electrònica. Universitat Politècnica de Catalunya

Abstract

Programming mechanisms in floating-gate non-volatile (EEPROM) standard-CMOS memories are briefly reviewed. A methodology to optimize the programming time in poly1-poly2 Fowler-Nordheim based structures is proposed. From design constraints, the optimum number of bumps and bootstrap capacitance value are obtained to maximize the programming speed for a given programming voltage.

1 Introduction

Non-volatile memories are key elements for permanent information-storage applications and adaptive or configurable systems. CMOS-compatible solid-state non-volatile memories are unique in portable and reduced-size systems, since they can be integrated in a monolithic substrate.

Electrically-Erasable Programmable Read-Only Memory (EEPROM) circuits have been well-known since long time in solid-state electronics, but in the recent last years EEPROMs are becoming more and more appreciated for emerging applications [1]. Recent advances in technology have allowed large integration of digital EEPROM cells in the so-called flash memories. Furthermore, in order to boost information capacity and reduce cost, four-valuated flash memories that store 2 bits per cell have been implemented [1,2], and probably more than 2 bits can be reliably stored. Flash memories are suitable for a wide range of mass-storage applications such as speech storage and recording, electronic cameras and personal computers.

Besides specific EEPROM processes that optimize the floating-gate programming process (using microtextures, and/or very thin oxides), it has been demonstrated that current standard CMOS processes, even though at a cost of efficiency, are suitable for the implementation of non-volatile memories. The reason is that the thin oxide layers in current standard CMOS processes combined with relatively high voltages enable the tunneling mechanisms employed for the floating-gate programming. Some applications of analog memories are bias and offset correction, parameter storage (as in neural networks or fuzzy systems), learning in adaptive systems and sensor aging compensation.

Because of the low tunneling charge rate, it is important to design the tunneling structures to maximize the tunneling current, thus reducing the programming time. In this paper we propose an injector structure for the Fowler-Nordheim tunneling effect and provide design rules for an improvement of the tunneling current.

In next section we will review the floating-gate charge injection mechanisms applicable to VLSI circuits. Section 3 is devoted to the analysis of the optimal parameters to reduce the floating-gate programming time. A new injector structure is proposed as well. The design of a test structure is reported in Section 4. Finally, in Section 5 conclusions and future work are discussed.

2 Floating-gate charging mechanisms

In standard CMOS technology, two mechanisms allow injection or extraction of electrical charge in a floating-gate: FN (Fowler-Nordheim) tunneling and hot-electron injection. Only FN tunneling is considered in this work.

Qualitatively, the FN tunneling effect appears in a MOS or conductor-SiO_2-conductor structure when a strong electric field is applied. Thin oxide layers, like gate oxide or poly1-poly2 oxide, achieve this required electric field at reasonable voltages. Because of thermal energy, electrons in the substrate have a small but existing probability of entering inside the insulator up to 5 nm. These electrons are pushed back to the substrate because of a high-energy barrier (near to 3.2 eV for 5 nm). However, if the applied electric field is strong enough then the electrons that penetrate 5 nm inside the oxide can overcome the 3.2 eV barrier, and they are attracted by the electric field towards the opposite conductor. The minimum value for the electric field to attract electrons penetrating the oxide more than 5 nm is thus $6.4\cdot10^6$ V/cm.

It is worth observing that the electric field has to be greater than E_{MIN} only locally at the substrate-insulator interface. This is necessary to help the electrons overcome the conductor-insulator energy barrier. Out of this zone, in the remaining oxide the electric field can be smaller. With a proper geometry, this fact will relax the

minimum voltage that has to be applied to the insulator and avoid reaching its breakdown value.

Equation (1) models the FN effect. Experimental results show good accuracy in more than ten decades of current density [3]:

$$J_{tun} = A_{FN} \cdot E^2 \cdot e^{\frac{-B_{FN}}{E}} \quad (1)$$

J_{tun} is the current density, E the electric field across the oxide and A_{FN} and B_{FN} are material and structure-dependent constants. Typically, $A_{FN} = 1.06 \cdot 10^{-6}\ A/V^2$ and $B_{FN} = 2.38 \cdot 10^{8}\ V/cm$ [3]. Of course, they have to be adjusted experimentally to model a particular geometry and process.

In a double-poly CMOS technology, the tunneling effect on a poly1 floating-gate of a MOS transistor can be produced either from the second polysilicon layer or from a substrate well. In both cases, the SiO_2 insulator is thin enough for tunneling. So either poly2-poly1 or n(p)well-poly1 injectors can be designed.

Depending on the application and the injection mechanism, two main floating-gate operating modes can be envisaged: discrete charging or continuous-time adaptation [4]. While continuous-time is more adapted to learning systems, discrete charging is suitable for analog memory applications.

Only discrete charging is considered in this paper. In this approach, the polysilicon floating gate of a MOS transistor is progressively charged or discharged by applying high-voltage pulses to a tunneling injector until the desired output current of the floating-gate MOS is programmed [5]. Unfortunately, MOS current measurement and floating-gate charging (discharging) cannot be simultaneous. The reason is that the voltage of the floating-gate changes when the injector high-voltage is removed because of capacitive coupling. So, usually the MOS current is measured after each pulse (or a small number of pulses), and by means of a simple feedback the pulses are interrupted when the desired MOS current is reached. This system is simple to implement and, once programmed, the high-voltage source is not needed.

3. Injector structures and optimal parameters

When using only FN tunneling to charge and discharge a floating-gate, a poly1-poly2 injector is most suitable, because a well-poly1 injector will only work either for charging or discharging (unless true twin-well technology is available). Since oxide in poly1-poly2 is normally thicker than gate oxide, tunneling has to be enhanced by increasing the local electric field at the electron source. Overlapping poly2 on poly1 (Fig. 1) boosts the tunneling effect. The bumps and corners that appear from the wafer processing produce sharp zones of locally increased electric field. To further increase the electric field, even sharper structures that violate the process design rules have been proposed [5]. However, such structures are not considered here for unpredictability reasons.

The basic injector composed of a single bump poly1-poly2 has a serious drawback. It requires a two-polarity high-voltage source to be applied to the poly2 terminal to be able to charge and discharge the poly1 floating-gate. Furthermore, the tunneling efficiency depends on the floating-gate voltage, which is not known a priori. By including a bootstrap control capacitor (C_B), this problem is solved (Fig. 2) [6]. Since C_B is designed for a capacitance much larger than that of the injector (C_I), MOS gate (C_G) and parasitic (C_P) capacitance, the voltage at the floating-gate is controlled by V_C. When the floating gate is to be charged with electrons (writing), then Vc is set to a high-voltage $+V_{PP}$ and V_I to 0 V. To remove electrons from the floating gate (erasure), then $V_C = 0\ V$ and $V_I = +V_{PP}$.

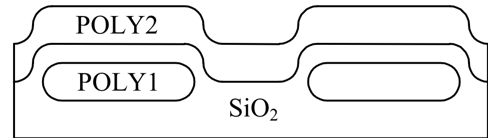

Fig. 1. Two-bump poly1-poly2 injector

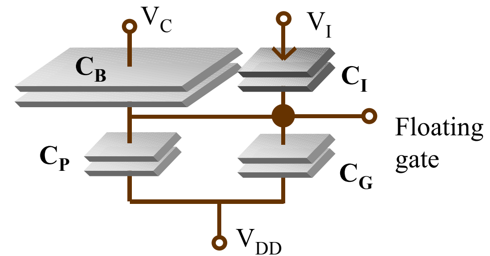

Fig. 2. Control capacitor model

In the following, the optimization process to reduce tunneling voltage and programming time is described. The available design parameters are the bootstrap capacitance and the number of injector bumps [7]. A PMOS transistor is considered because its substrate is connected to V_{DD}, reducing the load of parasitic capacitance.

From Fig. 2, the total capacitance C_T connected to the floating gate is:

$$C_T = C_B + b \cdot C_I + C_P + C_G \quad (2)$$

b is the number of bumps. The floating gate voltage is:

$$V_{fg\,w} = V_{PP}\left(\frac{C_B}{C_T} + m \cdot \frac{C_G + C_P}{C_T}\right) + \frac{Q_{prog}}{C_T} = \quad (3a)$$
$$= V_{PP} \cdot r_w + V_{prog}$$

$$V_{fge} = V_{PP} \cdot \left(\frac{C_I}{C_T} + m \cdot \frac{C_G + C_P}{C_T} \right) + \frac{Q_{prog}}{C_T} = \qquad (3b)$$
$$= V_{PP} \cdot (1 - r_e) + V_{prog}$$

Where V_{fgw} and V_{fge} are the writing (charging) and erasure (discharging) floating-gate voltages, respectively. V_{PP} is the programming high-voltage, $m = V_{DD}/V_{PP}$ is the power supply to programming voltage ratio. r_w and r_e are the writing and erasure control ratios that reduce the effective programming voltage V_{PP} by a factor smaller than unity. V_{prog} is the programmed floating-gate voltage (due to the injected charge). Starting from a minimum-sized injector structure, C_I, C_B and C_P have been estimated from layout. The gate MOS capacitance is approximated by $C_G = C'_{OX} \cdot W \cdot L$. In a 0.8-micron technology, it was found that C_I = 3.23 *fF*. The other capacitance values, related to C_I, are the following: $C_B = n \cdot C_I$, $C_P = n \cdot C_I / 10$, $C_G = 4 \cdot C_I$. Thus r_w becomes:

$$r_w = \frac{n\left(1 + \frac{m}{10}\right) + 4m}{1.1 \cdot n + 5} \underset{n \to \infty}{\to} \frac{10 + m}{11} \qquad (4)$$

r_e has a similar expression. Since similar results are obtained for r_e, and for space reasons, from now on only results for $r = r_w$ are presented.

To achieve a maximum field across the oxide, r has to be maximized. The optima will be for $n \to \infty$. Taking into account that $0 < m < 0.5$, for $n > 90$ the figure r is under 5% near to its asymptotic value. These calculations hold for other technologies, simply by changing the capacitance values. Because of scaling reasons, significant changes are not expected for other CMOS technologies.

Next step is, given a value for V_{PP}, to obtain the number of bumps and the C_B value to maximize the programming speed.

Obviously, as the number of bumps is increased, the injected current grows proportionally. However, since C_I also grows, the voltage drop at the injectors is reduced and so the tunneling current per bump. Equation (4) becomes now:

$$r_w = \frac{n\left(1 + \frac{m}{10}\right) + 4m}{1.1 \cdot n + b + 4} \qquad (5)$$

Combining (1), (3) and (5), the tunnel current is:

$$I_{tun}(b) = b \cdot S \cdot A_{FN} \cdot E_0^2 \cdot r^2 . e^{\frac{-B_{FN}}{E_0 \cdot r}} \qquad (6)$$

where S is the effective surface of current tunneling in one injector. The voltage programming rate of the floating-gate is:

$$\frac{dV_{fg}}{dt} = \frac{I_{tun}(b)}{C_T} = \frac{b \cdot I_{tun}(1)}{C_T} \qquad (7)$$

This equation has been calculated for several values of E_0 and n, under a fixed area constraint for C_B, the largest element in the memory cell. For $E_0 = 8 \cdot 10^6$ V/cm and C_B near to $100C_I$ (n = 100), the optimal number of bumps is between two and three, for both writing and erasure. Extending to other reasonable values E_0 and C_B, b is between 1 and 4. The reason for this finite minimum is that although the injector area increases with b, r is reduced, and thus the electric field across the injectors and tunneling efficiency.

However, a more general case is analyzed when no constraints are fixed on the area. Instead, a fixed r value is imposed independently on the number of bumps. The condition is imposed by equating expression (5) for one bump and a capacitance ratio n, to the same expression with b bumps: $r_w(1,n) = r_w(b,n_b)$. The new n_b ratio is:

$$n_b = \frac{40[m(b-1) + n(1-m)] + b \cdot n(10+m)}{50 - 39m} \qquad (8)$$

n_b guarantees a constant electric field across the injectors regardless to the number of bumps. Therefore, voltage rate $\frac{dV_{fg}}{dt}$ only depends on $\frac{b}{C_T}$. In Fig. 3, the voltage rate is represented for b=1~5 as a function of n and for a particular electric field.. For each value of b there exists an optimal n for maximum programming speed. Also, programming speed is higher with the number of bumps, and follows a saturation curve. If the memory cell area (that grows linearly with b) is divided by the optimal voltage rate for each b value, an area-delay figure of merit is obtained. The curve varies with $m = V_{DD}/V_{PP}$ ratio. For $V_{DD} = 5$, the optimal value is b = 2 for $V_{PP} <$ 13.2 V and b = 1 for larger programming voltages. A curve example is shown in Fig. 4 for m = 0.4.

The maxima of Fig. 3 that correspond to $E = 8 \cdot 10^6$ *V/cm*, have been evaluated for a broad range of E and several bumps (Fig. 5). Results show that n has to be increased about 50 units for each bump added. Also, dependance on E is rather flat.

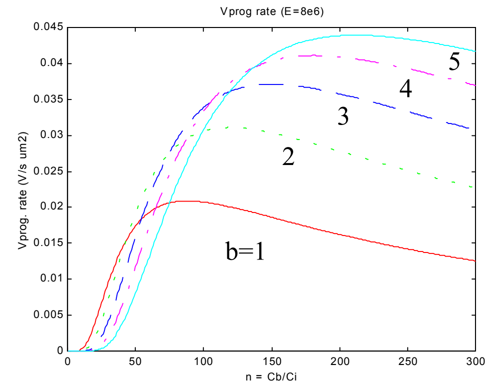

Figure 3. V_{fg} rate versus n

Assuming that two bumps is an optimal value for a large part of the E working range, a double-bump

injector is proposed. The bootstrap capacitance should be about C_B = 150 C_I. Moreover, in order to equalize both writing and erasure, a complementary injector is proposed. This injector is composed of a poly1-poly2 bump and a poly2-poly1 bump connected by metal lines. The geometrical symmetry guarantees a similar tunneling efficiency for both writing and erasure.

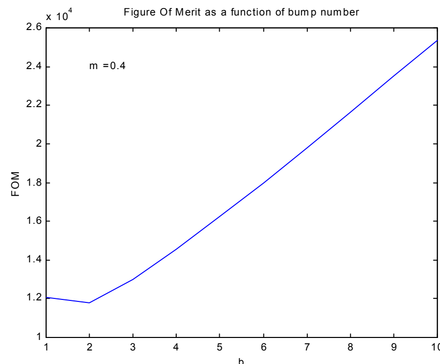

Figure 4. An area-delay cost curve

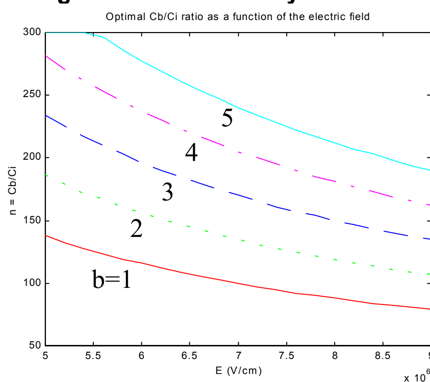

Figure 5. Optimal *n(E)* for *b* bumps

4. Test structure design

To verify the results obtained in the previous section, the layout of several test memory cells has been designed and will be sent to fabrication in a near run. CMOS 0.8-micron technology has been used. In this technology, the poly1-poly2 oxide depth is 20 nm. A sample layout is shown in Fig. 6. Several injectors have been implemented, among them: simple injector, Chai-Johnson injector [5], complementary injector (Fig. 7), double injector and poly1/n-well injector. Also, one of the transistors has an independent drain connection to measure the hot-electron injection.

5. Conclusions and future work

From the analysis of the FN tunneling in floating-gate CMOS-compatible EEPROM memory cells, design guidelines have been provided for the poly1-poly2 structures. It has been shown that there exists an optimum number of injector bumps as a function of the electric field and the bootstrap capacitance. Moreover, this capacitance can be adjusted to optimize the programming time.

A layout in a 0.8-micron CMOS technology has been designed, and will be sent to manufacturing. Upon reception of the samples, the different injector efficiencies will be measured and the theoretical results presented here will be experimentally checked .

Acknowledgement. This work has been done under the support of CICYT Spanish Project TIC96-1195.

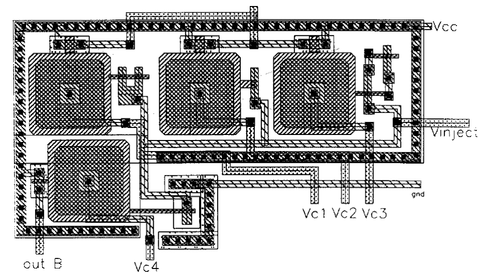

Figure 6. Layout of some test cells

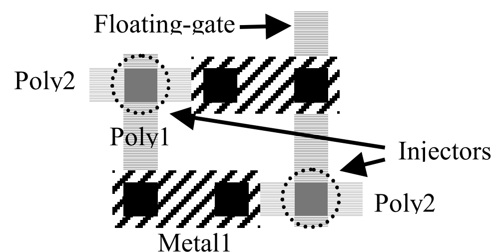

Figure 7. Complementary injector

References

[1] *High-Density Flash Memories Tackle Mass-Storage Needs*, D. Bursky, Electronic Design, Sept. 3, 1996, pp. 80.

[2] *A 117-mm² 3.3-V Only 128-Mb Multilevel NAND Flash Memory for Mass Storage Applications*, T.S. Jung et al., IEEE JSSC, Vol. 31 No. 11, Nov. 1996, pp.1575-1583.

[3] *On tunneling in metal-oxide-silicon structures*, Z.A. Weinberg, J. Appl.Phys., Vol. 53, July 82, pp. 5052-5056.

[4] *Single Transistor Learning Synapse with Long-Term Storage*, P. Hasler, C. Diorio, B.A. Minch and C. Mead, Proc. of IEEE ISCAS'95.

[5] *A 2 × 2 Analog Memory Implemented with a Special Layout Injector*, Y.Y. Chai, L.G. Johnson, IEEE Jour. Solid-State Circuits, Vol. 31 No. 6, June 1996, pp.856-859.

[6] *Floating gate MOSFET with reduced programming voltage*, Y.Y. Chai, L.G. Johnson, IEE Electronics Letters, 1 Sept. 1994, vol. 30, No.18, pp. 1536-1537.

[7] *A Nonvolatile Analog Neural Memory Using Floating-Gate MOS Transistors*, H. Yang, B.J. Sheu, J.C. Lee, Analog Integrated Circuits and Signal Proc. 2, 1992, pp. 19-25.