

OPTIMIZATION AND CONSTRAINTS OF SINGLE-POLY NON-VOLATILE MEMORY CELLS FOR EMBEDDED APPLICATIONS

Paola Vega-Castillo^{1,2}, Wolfgang H. Krautschneider¹

¹Hamburg University of Technology, ²Instituto Tecnológico de Costa Rica

paola.vega@tu-harburg.de, krautschneider@tu-harburg.de

ABSTRACT:

The influence of the cell geometry in the programming and erasing characteristics of a fully CMOS compatible single-poly non-volatile memory cell for low cost embedded applications was investigated using cells fabricated in a standard CMOS 0.35 μm -technology with transistor sizes varying from 10 μm /0.3 μm to 0.7 μm /0.3 μm , being the 0.7 μm -wide cells the smallest single poly non-volatile memory cells reported until now. In addition to this, the impact of transistor separation is presented and discussed, together with the impact of the cell dimensions in the characteristics and endurance of the cells.

1. INTRODUCTION

One of the first proposals for a CMOS compatible single-poly non-volatile memory cell used both NMOS and PMOS transistors to form the storage cells, resulting in relatively large cell areas [1]. A single-poly EEPROM cell using only NMOS transistors was reported in [2]. In [3], measurement results for the NMOS and PMOS version of this type of cell are presented. In this paper, the cell dimensions were varied to study their impact on the programming and erasing operations. All investigated cells were fabricated using only one polysilicon level at the gate with a standard 0.35 μm standard CMOS process, with transistor widths varying from 10 μm to 0.7 μm . The gate oxide thickness is 7.5nm, which is appropriate for non-volatile memory applications.

This paper presents the impact of cell geometry in the reading, programming and erasing characteristics of the cell, and its influence in the cell endurance.

2. CELL OPERATION

2.1 Cell Structure and Operation

The cell consists of two transistors of the same type connected in series [2,3], as presented in figure 1. One transistor operates as access transistor, and the second transistor provides the non-volatile data storage, since its

single poly gate is left unconnected. Since no control gate is required, the cell can be easily integrated using a standard CMOS process, providing a low cost memory cell.

To read the cells, the bitline is grounded and the selectline and wordline are biased to turn on the access transistor. For the programming or erasing operation, the selectline is grounded and the wordline and bitline are biased. Cell operation is based on hot-electron and hot-hole injection for the erasing and programming operation, respectively. This cell can be optimized to favor one type of carrier injection by changing the cell sizing, so that the cell can operate as EEPROM or as OTP [3]. Favoring one specific type of carrier injection may decrease the operating voltages of the cell, and allows adapting it to a specific application.

2.2 Cell Model

Transistor models do not account for the variations of the cell geometry that are essential to cell optimization, neither correctly consider all capacitive components that contribute to couple voltage to the floating gate.

In order to model the voltage coupled to the floating gate correctly, two additional capacitive components must be considered. One of them is the coupling capacitance between the control gate and the floating gate (C_{WL}) and the other is the coupling capacitance between the bitline electrode and the floating gate [4]. None of these capacitances are included in the BSIM model. In addition to this, the modified characteristics of the shared diffusion node between the floating gate transistor and the access transistor caused by different transistor separation are also not considered in conventional transistor models.

The voltage coupled to the floating gate can be calculated by means of the following formula:

$$V_{FG} = \alpha_{BL} \cdot V_{BL} + \alpha_{WL} \cdot V_{WL} + \alpha_X \cdot V_X + \alpha_B \cdot V_B \quad (1)$$

where α_{BL} , α_{WL} , α_X , α_B are the coupling coefficients between the floating gate and the bitline, wordline, shared diffusion node and bulk, respectively.

V_{BL}, V_{WL}, V_X, V_B are the voltages at the bitline, wordline, shared diffusion node and bulk, respectively. Figure 2 shows a cross section and top view of the memory cell and the capacitances involved in calculating the coupled voltage to the floating gate. For simplicity, only the intrinsic capacitances of the bitline and the node X are not shown. The reader is referred to figure 2 for the description of the capacitances and their location, and to the detailed schematics of figure 3b) for derivation of the expressions for the coupling coefficients, which will be explained in the following.

The coupling coefficients are calculated according to

$$\alpha_i = \frac{C_i}{C_{BL} + C_{WL} + C_X + C_B} \quad (2)$$

where α_i is the coupling coefficient between the terminal i and the floating gate, and C_{BL}, C_{WL}, C_X, C_B are the capacitances between the floating gate and bitline, wordline, shared diffusion node and bulk, respectively, as shown in the simplified schematics of figure 3a).

The capacitance between bitline and floating gate can be expressed as

$$C_{BL} = C_{el} + C_{overlap} + C_{fringe} + C_{BLint} \quad (3)$$

where C_{el} represents the capacitive coupling between the metal electrode and the floating gate, $C_{overlap}$ includes the bias dependent and bias independent components of the overlap capacitance and C_{fringe} is the fringing capacitance between the floating gate and the diffusion region at the bitline side, given as the sum of the outer and inner fringing capacitance. C_{BLint} represents the intrinsic capacitance between the bitline and the floating gate, which is not shown in figure 2 for the sake of simplicity.

On the other hand, the capacitance between the shared diffusion node X and the floating gate capacitance is calculated as

$$C_X = C_{overlap} + C_{fringe} + C_{Xint} \quad (4)$$

since there is no metal electrode at the shared diffusion region. C_{Xint} represents the intrinsic capacitance between the shared diffusion node X and the floating gate, which is not shown in figure 2 for the sake of simplicity.

The gate to bulk capacitance is calculated as follows

$$C_B = \frac{C_{ox} \cdot C_{depl}}{C_{ox} + C_{depl}} + C_{overlap} \quad (5)$$

and reduces to $C_{overlap}$ once the inversion layer is formed at the floating gate transistor.

It is important to consider that the gate oxide capacitance is connected in series with the inversion layer capacitance arising from the non-zero charge thickness [5], yielding an effective oxide capacitance as follows,

$$C_{ox_effective} = \frac{C_{ox} \cdot C_{cen}}{C_{ox} + C_{cen}} \quad (6)$$

The simulator SAP[6] was applied to calculate the capacitance per unit length for these two additional capacitive components. To extract these capacitances using the simulator, several considerations must be taken into account when defining the structure file for simulation:

- 1) The separation between gate and drain contacts must consider the effective gate length data provided by the manufacturer. Since the effective gate length is larger than the drain and active length, it reduces the distance between the drain electrodes and the gate.
- 2) The metal layer on top of the contacts must also be considered, and its separation from the gate must also consider the effective gate length.
- 3) For the electrode capacitance, consider the actual metal extension and not the gate extension, since the gate extends beyond the metal.
- 4) For the gate to gate coupling capacitance, consider the actual gate extension, which means, not only the active width of the gate, but also the gate extension beyond the diffusion regions

Reference [4] calculates the gate to electrode capacitance using the formula for parallel plate capacitance. However, this approach leads to an overestimation of the electrode capacitance, since the contact region is not a continuous plate, but reaches the diffusion area only at the contacts. In addition to this, the metal wire on top of the contacts also contributes to the electrode capacitance. According to these remarks, two structures are defined and simulated to extract the mutual capacitance between the bitline electrode and the floating gate. The structures are shown in figure 4. The extracted capacitances per unit length are 21.171 pF/m for the structure of figure 4a) and 22.543pF/m for the structure of figure 4b). After these data, the total electrode capacitance for a 10 μ m-wide transistor is 207.6aF, instead of 359.5aF when using the parallel plate approximation of [4].

On the other hand, the gate to gate coupling capacitance was also extracted. The extracted capacitance was 26.2034pF/m for a drawn gate separation of 1 μ m.

Figure 5 presents the measured IV characteristics of a memory cell with 10 μ m-wide transistors and a separation of 0.7 μ m. The asymmetry of the IV curves of the memory cells can be explained when the influence of the body effect is considered. For biased bitline conditions, increasing wordline voltages lead to a decrease of voltage

at the shared diffusion node. Thus, the body effect at the floating transistor decreases as the wordline voltage increases. In addition to this, the access transistor does not suffer any body effect.

On the contrary, for biased selectline conditions, the access transistor suffers from a strong body effect, since the voltage at the shared diffusion node is given by the subtraction of the selectline voltage and the drain to source voltage drop of the access transistor. In addition to this, decreasing the voltage at the common node leads to a lower voltage coupled to the floating gate.

3. IMPACT OF CELL DIMENSIONS ON THE CELL'S CHARACTERISTICS

Figure 6 presents measurements of the gate current density for an NMOS transistor plotted as absolute value, and the regions of drain avalanche hot hole injection and hot electron injection. Hot hole injection occurs at low gate voltages, whereas hot electron injection occurs at higher gate voltages. The point minimum current density defines the beginning of the hot electron injection regime. Thus, optimization of the cells is based on the criteria of coupling the floating gate voltage that coincides with the points of maximum injection efficiency for hot hole injection and hot electron injection, respectively. In case of the hot electron injection, the maximum electron current would be placed in the gate voltage range of the channel hot electron injection regime. Due to the coupling characteristics of the cell, it is not possible to operate the cell in that regime. Hence, the optimization criterion for hot electron injection is to achieve maximum injection under the drain avalanche hot electron injection regime.

In case of the NMOS cell, lower wordline voltages lead to higher voltage coupled to the floating gate, since the voltage of the shared diffusion node increases. In such case, the NMOS cell operates in hot electron injection mode for cell erasing. Larger wordline voltages tend to decrease the voltage at the shared node, which decrease the coupled floating gate voltage, and the NMOS cell operates in hot hole injection mode for cell programming. In case of the PMOS cell, only hot electron injection is possible, since the energy barrier for injection of holes is higher than that of electrons.

Figure 7 shows a top view of the cell and the dimensions varied to study cell optimization. The influence of the transistor separation on the IV characteristics can be explained by considering the following facts:

- 1) As the separation between the gates decreases, the coupling between the gates is stronger
- 2) Shorter transistor separation leads to a decrease of the fringing fields and thus to the equivalent fringing field capacitance between the floating gate and the shared diffusion node X. This is also a consequence of 1).

- 3) Shorter distances between the transistors affect the shape, depth and concentration of the shared diffusion node, thus making the cell's transistors asymmetric

Gate proximity reduces the fringing capacitive component from node X and increases the capacitive coupling between the gates. The decrease of C_{fringe} at node X is larger than the increase of C_{WL} , leading to a decrease of the denominator of equation 2 and hence, to an increase of the coupling factors. The non-zero charge thickness capacitance also contributes to a decrease of the effective gate oxide capacitance, which also increases the coupling factors.

Decreasing the distance between gates also leads to a lower doping concentration and shallower junction at the junction depth in the highly doped region. This increases the series resistance and may influence to some extent the charge sharing at both transistors, thus affecting the DIBL and short channel effect parameters.

Table 1 summarizes the measured bitline current for cells with different transistor widths and transistor separation under the following bias conditions: $V_{bitline}=3V$, $V_{wordline}=3V$, $V_{selectline}=0$ and $V_{substrate}=0$.

Table 1. Measured NMOS-cell bitline current for different cell dimensions

Transistor width (μm)	Separation (μm)		
	0.45	0.7	1.0
0.7	141.35nA	3.27nA	33.7nA
1.0	12.78 μA	10.57 μA	8.53 μA
2.5	49.58 μA	35.32 μA	29.09 μA
5	112.12 μA	89.52 μA	78.55 μA
10	395.9 μA	237.26 μA	87.82 μA

Table 2. Equivalent coupled floating gate voltage for the measured NMOS-cell bitline current at different cell dimensions

Transistor width (μm)	Separation (μm)		
	0.45	0.7	1.0
1.0	0.8V	0.77V	0.74V
2.5	0.905V	0.84V	0.8V
5	0.94V	0.89V	0.86V
10	1.12V	0.955V	0.765V

From table 1 it can be noticed that the cells with shorter transistor separation present the highest cell current, and that this tendency is followed by the cells independently of the transistor width. It also shows that

cells with wider transistors tend to couple higher voltages to the floating gate.

Table 2 presents the coupled floating gate voltage according to an equivalent structure consisting of two transistors with both gates accessible, with a standard transistor separation defined by the design rules. Table 2 clearly shows the tendency presented on table 1, but in terms of coupled voltage. Table 1 also presents an important effect of the decrease of the transistor width, namely a wider statistical distribution of the cell characteristics for the transistor width of $0.7\mu\text{m}$ for the NMOS cells. This effect is presented by an example illustrated by figure 8, in which the cell current a group of samples with transistor width $0.7\mu\text{m}$ and cell separation of $0.7\mu\text{m}$ at biased bitline. This makes more difficult to distinguish between a 1 and a 0 stored at the cell, as shown by the spread in the current difference before and after programming shown in figure 8.

Shorter transistor separations lead to improvement of the erasing characteristics, whereas larger separations lead to improvement of the programming characteristics [3]. Thus, the separation between transistors is a parameter of key importance in order to optimize the single-poly memory cell. Figure 9 presents an example of programming of a $10\mu\text{m}$ cell with a transistor separation of $1\mu\text{m}$, and erasing of $10\mu\text{m}$ cell with a transistor separation of $0.45\mu\text{m}$, corresponding to the best cases of programming and erasing. It's also important to notice that independently of the cell width, the transistor separation has a significant influence on the cell I-V characteristics and on the programming and erasing properties of the cell.

The lower cell current of the narrower cells leads to a decrease of the channel current carriers that can contribute to the avalanche impact ionization at the depletion region in the bitline. Although it is possible to program the $0.7\mu\text{m}$ -wide NMOS cells, erasing becomes more difficult, not reaching the initial current levels before programming, making more difficult for the read stages of the memories to distinguish a '1' from a '0'. In addition to this, the statistical spread of the cells before and after programming may limit their practical application, for which transistor widths of $1\mu\text{m}$ or $2.5\mu\text{m}$ would be more advisable.

The PMOS cell is functional also at a transistor width of $0.7\mu\text{m}$ and presents a current of about $15\mu\text{A}$ at 3V, which is considerably higher current than that of the NMOS cells of the same width. The low gate voltage required for hot electron injection (4.5V) makes possible to program this type of cell, however the cell can only be used as EPROM or OTP, since it is not possible to inject hot holes to erase it, and the coupling characteristics of the cell makes not possible to apply Fowler-Nordheim tunneling.

Narrower cells are programmable at 6.5V. However, the programming time can be decreased by increasing the programming voltage. At 6.5V, the programming time is 4ms for a $10\mu\text{m}$ wide cell. The cells can also be

programmed at 8V to increase the programming speed, but it must be considered that the typical drain/source to bulk breakdown voltage of this technology is $>7\text{V}$.

4. IMPACT OF CELL GEOMETRY IN THE ENDURANCE

Carrier injection into the floating gate inevitably leads to oxide damage and degraded series resistance due to the presence of traps and interface states. It has been shown that hot hole injection causes more damage than hot electron injection [7]. Electrons are injected during the drain avalanche regime at higher gate voltage, although this regime also includes injection of holes. However, in this voltage range, electron injection predominates over hole injection and thus a net electron current is measured at the gate. The presence of this hot hole component increases the oxide damage of drain avalanche injection compared to that of channel hot electron injection.

Lower programming times and cell current could be achieved if the coupled voltage at the floating gate is near or at the maximum hole injection point. On the other hand, higher gate voltages would decrease the hole injection current and increase the programming time and the cell current during programming. In any case, the cell should be programmed only during the exactly necessary time to achieve the desired threshold voltage shift to avoid unnecessary damage, especially when considering that the programming characteristic of a non-volatile memory always reaches a saturation limit.

Regarding electron injection, the coupled gate voltage should be as high as possible not only to reduce the erasing time, but also to reduce the damage caused by the hot-hole injection component. The higher the coupled voltage, the lower the hot-hole component and thus its damage at the oxide. Better coupling leads to lower power consumption during programming and erasing, since the bitline voltage can be reduced to obtain an equivalent or improved injection efficiency. This explains the importance of the enhanced coupling at cells with shorter transistor separation.

The coupling mechanism at the floating gate also leads to the necessity of wordline pulses shorter than the bitline pulses, especially during erasing. In addition to this, the bitline pulse must start before the wordline pulse and finish after it. When the bitline voltage changes from zero Volt to the erasing bias, the floating gate voltage follows the bitline voltage from a low voltage -determined by the initial voltages and coupling conditions- to its maximum magnitude during erasing. During the transition, if the access transistor is active, current can flow from the bitline to the selectline and the floating gate transistor fulfills the hot hole injection conditions, that is, to the left of the minimum separating the hot hole and the hot electron regimes in figure 6. The same will happen when the bitline voltage decreases at the negative edge of the erasing

pulse. Thus, erasing is less efficient due to the additional hot hole injection during part of the positive and negative edge of the bitline pulse. Hence, to achieve appreciable erasing, the erasing time must be increased. In addition to this, unnecessary damage is caused to the oxide due to unintended hot hole injection.

To avoid this unwanted effect, the bitline voltage is increased before turning the access transistor on, preventing the current flow from bitline to selectline and increasing the coupled voltage due to the increased voltage at the shared node. The same procedure applies to the negative edge of the bitline voltage. An example of the importance of this procedure is shown by figure 10, which shows measurement results for an 10 μm -wide NMOS cell. The maximum number of program-erase cycles for simultaneous pulsing of wordline and bitline is about 1000 cycles, whereas delaying the wordline pulse with respect to the bitline leads to a maximum number of program-erase cycles of 40000. This makes evident the large oxide damage caused by simultaneous pulsing of the bitline and wordline. In addition to this, larger hot-hole injection efficiency increases the damage not only during the programming, but also increases the hole current component during erasing. Although improved hole injection efficiency would be desirable due to the lower magnitude of the current, there is a trade off between better hole injection efficiency and the endurance. In addition to this, coupling a higher voltage to the floating gate is also desirable due to the difficulty in reaching the gate voltage range of the hot electron injection regime.

The lower coupled floating gate voltage for the 0.7 μm -wide cells leads to a decrease of hot electron injection efficiency, and also to degraded endurance characteristics. It is important to remind that although the net measured gate current indicates electron injection, the gate current presents also a hot hole component during hot electron injection. This explains the degraded endurance characteristics observed for the 0.7 μm -wide cells.

5. APPLICATIONS

These single-poly cells present the advantage of the low cost due to the integration in a 100% standard CMOS process and the use of standard library components. As an example of the economical advantage of integrating this memory using a standard CMOS process, the integration cost per square millimeter of a memory system using the standard CMOS process can be reduced in 28% [8a, 8b] compared to the cost using an EEPROM process with the same feature size and from the same manufacturer. In some cases, the cost can be reduced up to 65% [8c]. Additionally, the understanding of the optimization criteria, physical mechanisms of operation and the constraints of these cells gives insights about:

- Minimum cell area, which is relevant for the integration costs

- Programming and erasing efficiency, defined by the magnitude of the coupled floating gate voltage
- Programming and erasing times, defined by the programming efficiency
- Minimum operation voltage for the reading, programming and erasing operations, defined by the magnitude of the coupled floating gate voltage
- Charge pump design requirements for the programming and erasing operations
- Maximum number of programming and erasing cycles and cell endurance, and the timing criteria and the order of application of the bias pulses for the program-verify and erase circuitry.

Thus, choosing a certain cell width and cell separation may depend on the specific application and process characteristics.

Low cost embedded systems are an important application of these cells, since with them it is not necessary to have an EEPROM fabrication process in order to integrate the non-volatile memory, neither create nor test the feasibility of integrating other devices or structures with the available integration process, thus relaxing the technological constraints of embedding non-volatile memories.

In [9], an analog calibration memory system for calibration of the impedance of the RF-Front of a smart label was designed applying the single-poly memory cells described in this paper.

Other possible applications of the single-poly memory cells include code and data storage for smart labels and smart cards, identification memories for wafer traceability and configurable commercial subproducts in which the features and/or settings of a product depends on the region, (e.g., language) and/or licensing.

6. CONCLUSIONS

Single poly non-volatile memory cells with transistor widths from 10 μm to 0.7 μm were fabricated and measured to study the constraints and potentials of this type of memory cells. The 0.7 μm -wide cells are the smallest single-poly non-volatile memory cells reported until now. It was confirmed by measurements that transistor separation is an important parameter to optimize the cells, regardless of the width of the cell's transistors. It was found that the large statistical spread of the 0.7 μm -wide cells may limit the practical application of cells with these dimensions. Therefore, it is recommendable to keep the minimum cell transistor width larger or equal than 1 μm .

Additionally, the relevance of the transistor width and transistor separation of the cells in the cell performance and the operation voltages was explained, and the appropriate way to apply the bias pulses during cell operation to avoid unwanted cell degradation was presented.

7. REFERENCES

- [1] Ohsaki, K. et al. "A Single Poly EEPROM Cell Structure for Use in Standard CMOS Processes", IEEE Journal of Solid-State Circuits, Vol. 29, No. 3, pages 311-315, March 1994.
- [2] Lee, K. et King, Y. "New Single-poly EEPROM with Cell Size down to 8F2 for High Density Embedded Non-volatile Memory Applications", 2003 Symposium on VLSI Technology Digest of Technical Papers.
- [3] Vega-Castillo et al. "Non Volatile Memory Cells integrable using standard CMOS Processes", Proceedings of the 7th Annual Workshop on Future Electronics SAFE 2004 , pages 721-725, November 2004.
- [4] Wakita, N. et Shigyo, N. "Verification of overlap and fringing capacitance models of MOSFETs", Solid-State Electronics, Vol. 44, pages 1106-1109, 2000.
- [5] BSIM3 Manual, Chapter 4.
- [6] www.iue.tuwien.ac.at/software.0.html
- [7] Shimoyama, N. et al. "Enhanced Hot-Carrier-Degradation in LDD MOSFET's Under Pulsed Stress", IEEE Transactions on Electron Devices, Vol 42, No. 9, 1995, pages 1600-1604.
- [8] Sources: a) Philips Semiconductors (0.35 μ m process),

b) <http://www.europractice.imec.be/europractice/on-line-docs/prototyping/sp/protprices2005.html> (AMS process C35B4), c) X-Fab (0.35 μ m process).

- [9] P. Vega-Castillo, Wolfgang H. Krautschneider. "Low voltage, low power, self-clocked memory read/program-verify circuitry with adjustable operating frequency", Proceedings of 8th Annual Workshop on Program for Research on Integrated Systems and Circuits (PRORISC), The Netherlands, November 17th-18th 2005.

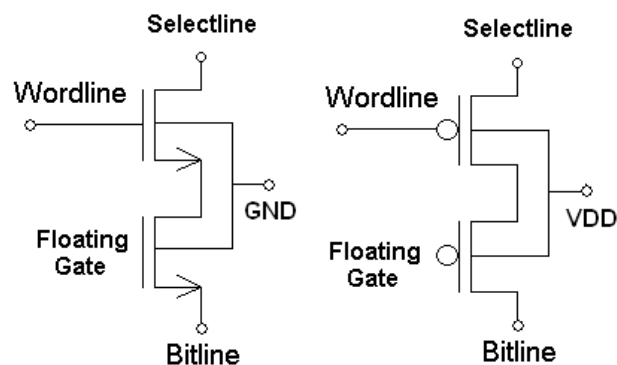


Figure 1. Schematics of the cells in the a) NMOS and b) PMOS version

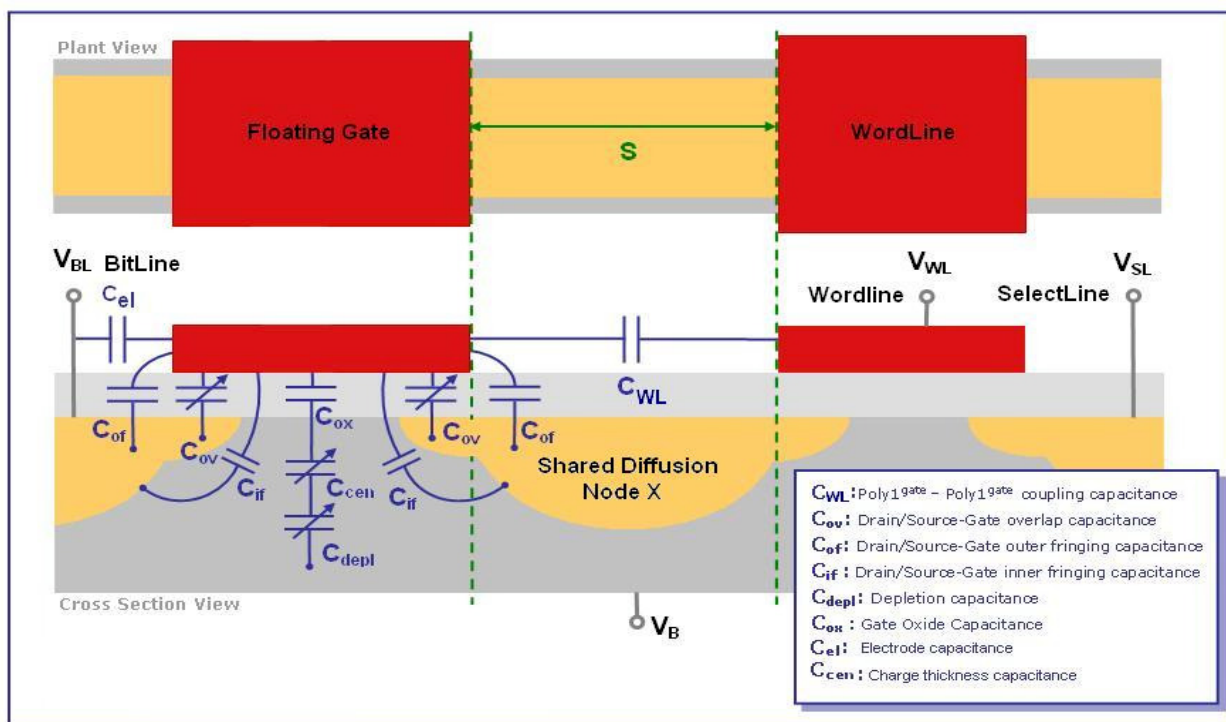


Figure 2. Top and cross sectional view of the cell showing the capacitances involved in the coupling characteristics. For simplicity, some of the intrinsic capacitances are not shown.

C_{WL} : Poly1 gate- Poly1 gate coupling capacitance, C_{ov} : Overlap capacitances, C_{of} : Drain/Source -Gate outer fringing capacitance, C_{if} : Drain/Source -Gate inner fringing capacitance, C_{depl} : Depletion capacitance, C_{ox} : Gate Oxide Capacitance, C_{e1} : Bitline electrode to Floating gate capacitance

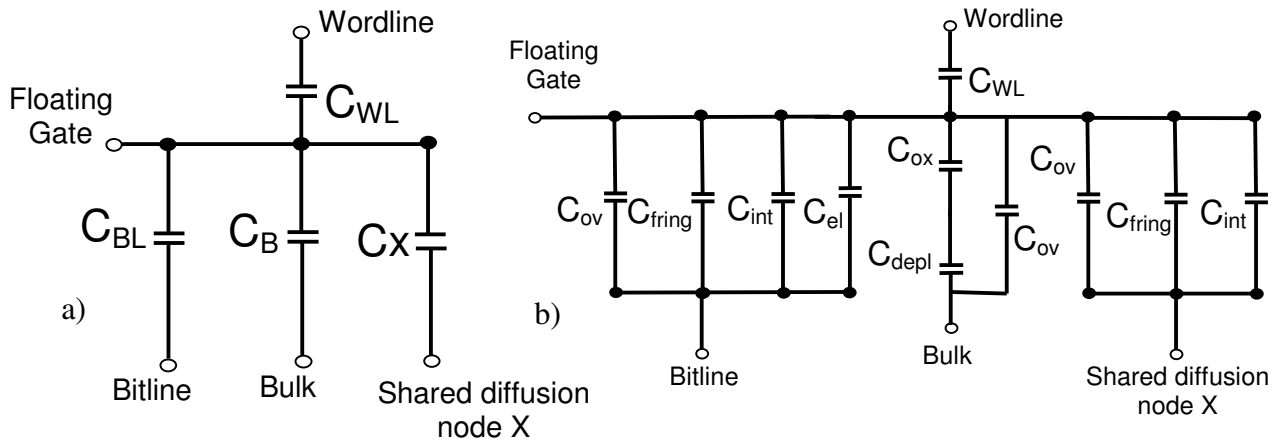


Figure 3. Circuit schematics for calculation of the coupling factors, a) showing the total equivalent capacitance between floating gate and the coupling nodes, b) showing the detail of the capacitances involved in the coupling mechanism. The C_{fring} includes the outer and inner fringing capacitances. C_{ov} includes the bias dependent and the bias independent overlap capacitances; C_{int} represents the intrinsic capacitance between the floating gate and the corresponding node.

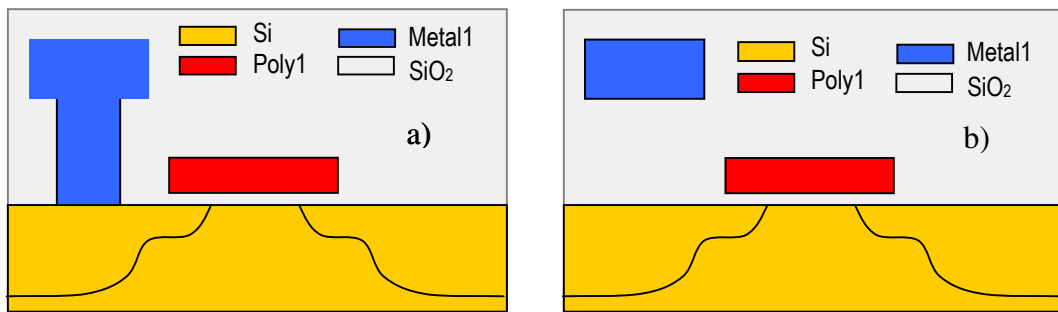


Figure 4. Structures defined for electrode capacitance extraction with SAP a) with contact, b) without contact

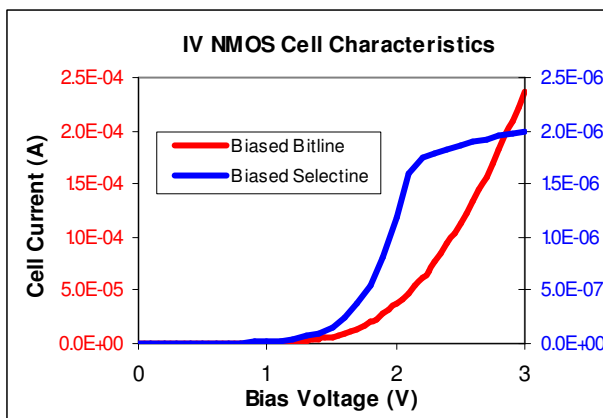


Figure 5. Measured IV characteristics of an NMOS Cell (Transistor width= $10\mu\text{m}$, transistor separation= $0.7\mu\text{m}$)

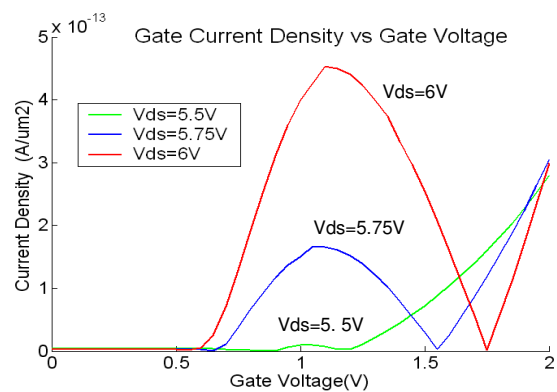


Figure 6. Gate current density vs gate voltage for an NMOS transistor $10\mu\text{m}/0.3\mu\text{m}$

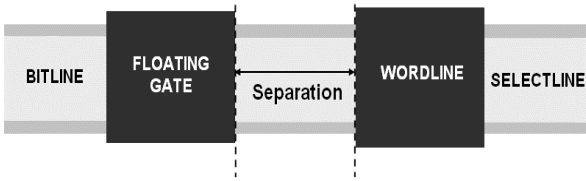


Figure 7. Top view of the cell, showing the transistor separation at the shared node X.

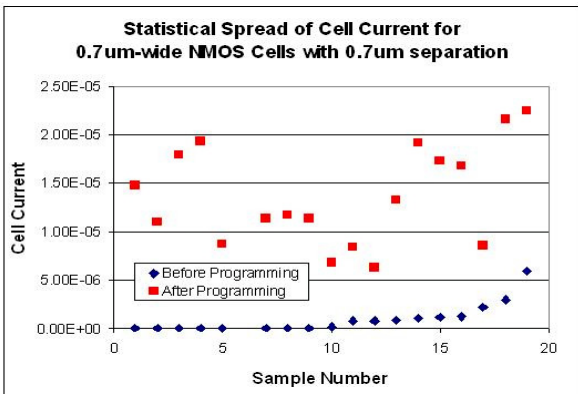


Figure 8. Statistical spread of cell characteristics before and after programming for $0.7\mu\text{m}$ cells with separation of $0.7\mu\text{m}$

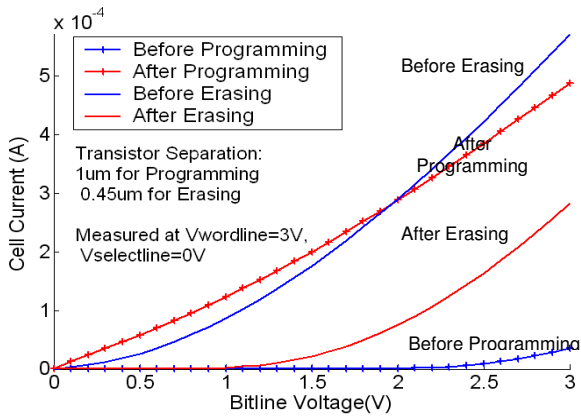


Figure 9. Best case of programming and erasing. Transistor separation: $1\mu\text{m}$ for programming, $0.45\mu\text{m}$ for erasing

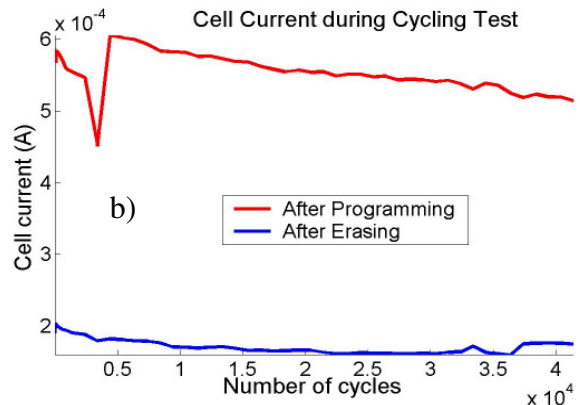
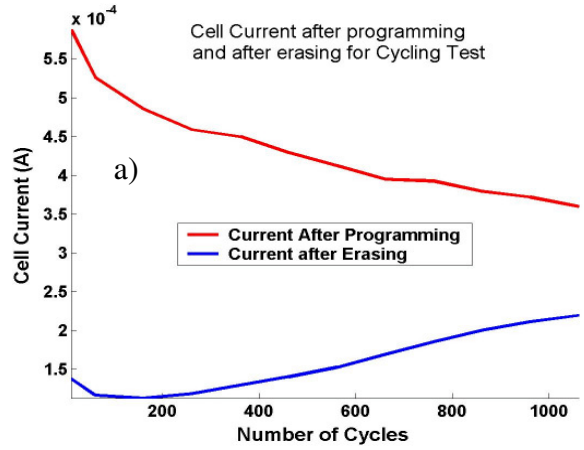


Figure 10. Maximum number of program and erase cycles for NMOS cells with $10\mu\text{m}$ transistor width: a) by simultaneous pulsing of wordline and bitline, b) by pulsing the bitline before the wordline