Program/Erasure Speed, Endurance, Retention, and Disturbance Characteristics of Single-Poly Embedded Flash Cells

Seung-Hwan Song, Jongyeon Kim, and Chris H. Kim
Department of Electrical and Computer Engineering, University of Minnesota
200 Union Street SE, Minneapolis, MN 55455, USA (Email: songx278@umn.edu)

Abstract—N-channel and P-channel single-poly embedded flash (eflash) memory cells were implemented in a standard CMOS logic process. Among the different configurations based on standard I/O devices, the N-channel cell with a PMOS-NMOS-NMOS combo and the P-channel cell with an NMOS-NMOS-PMOS combo were found to be most attractive in terms of program/erase performance, while the cell with a coupling device having P+ poly showed longer retention characteristic than the cells with a coupling device having N+ poly. Negligible program disturbance and floating gate coupling were observed in all cell types.

Keywords: Flash Program/Erase, Flash Reliability, Embedded Flash, Single-Poly Embedded Flash Cell

I. INTRODUCTION

Embedded flash (eflash) memory serves as an essential building block in system-on-chip applications providing a secure non-volatile storage for program, code, and system parameters during periods when the chip is not powered [1]. Eflash also plays an important role for mitigating circuit variation and reliability issues which often resort to the programmable or tunable digital and analog circuit architectures. Despite the growing need for moderate amounts of nonvolatile storage, existing eflash technologies such as dual-poly or split-gate eflash (Table I) require considerable process overhead to build Floating Gate (FG) and high voltage (>14V) transistors [2, 3]. Single-poly eflash [4-7] on the other hand has no process overhead as it utilizes standard I/O devices readily available in a logic CMOS process (Fig. 1). So far, various cell transistor and FG doping types have been proposed to enhance the electron injection efficiency of single-poly eflash cells [5-7]; however, cell characteristics have not been fully compared yet. On the other hand, the prior eflash cell in [4] is extremely large (~700µm²) due to the dedicated high voltage switch, sense amplifier, SRAM in each cell to avoid program disturbance, while other prior eflash cells in [5, 6] have the over-stress issues of the unselected cells due to the high writing voltage levels. In this paper, we compare various single poly 5T eflash configurations in terms of program/erase speed, endurance, and retention. Additionally, we show that single poly 5T eflash has minimal program disturbance and FG coupling, which verifies that self-boosting [8, 9] in conjunction with a tight BL pitch can be utilized effectively without causing significant disturbance issues.

II. SINGLE-POLY 5T EFLASH CELL OPERATION

The basic operation principle of a 5T eflash cell is shown in Fig. 2. The operation principle of an N-channel 5T eflash was elaborately described in [7]. For P-channel 5T eflash, the negatively boosted high voltage (i.e. -7.6V) is applied to the selected WWL during erase operation, while PWL and WWL are both pulled down to a negative voltage during program operation. A high electric field is generated in the M2 gate oxide during erase operation and the M3 gate oxide of a selected BL (i.e. BL=0V for an N-channel 5T eflash or BL=1.2V for a P-channel 5T eflash) during program operation. The difference in the charge stored in FG for the erased ('E') and programmed ('P') states results in different BL discharging or charging rates during read operations.

### TABLE I. EFLASH MEMORY OPTIONS

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Schematic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BL</td>
<td>CG</td>
<td>FG</td>
</tr>
<tr>
<td>Process Overhead</td>
<td>Floating Gate</td>
<td>Nanocrystal</td>
<td>None</td>
</tr>
<tr>
<td>Min. Tuner/Code</td>
<td>CHE Injection</td>
<td>SS Injection</td>
<td>FN Tunneling</td>
</tr>
<tr>
<td>Erase Method</td>
<td>FN Tunneling</td>
<td>FN Tunneling</td>
<td>FN Tunneling</td>
</tr>
<tr>
<td>Cell Size</td>
<td>0.4µm²/90nm</td>
<td>N.A.</td>
<td>0.4µm²/65nm</td>
</tr>
</tbody>
</table>

Fig. 1. Various single-poly eflash cell cross sections.
III. PROGRAM/ERASE SPEED, ENDURANCE, RETENTION

A. Program and Erase Speed Comparison

Table II shows various configurations of N-channel (N-ch.-1, 2, 3) and P-channel (P-ch.-1, 2, 3) 5T eflash cells having different combinations of M1-M3. All cell types can be implemented in a standard CMOS process. The device sizing ratio \( \frac{W_{M1}}{W_{M2,3}} \) is set as 8 for all designs. Cross section of M1-M3 for the different P-channel 5T eflash configurations and the corresponding energy band diagrams of M2 and M3 transistors during erase and program operations are shown in Figs. 3, 4, respectively. The voltages applied to the tunnel oxide during erase and program operations are shown in Figs. 3, 4, and summarized in Table III. During erase operation, the erase speed of P-ch.-3 is faster than that of P-ch.-2 as the coupling transistor M1 of P-ch.-3 (=PCAP) operates in a depletion mode, producing a high \( V_{OX,E} \) in M3, while M1 of P-ch.-3 (=PCAP) operates in a depletion mode, producing a low \( V_{OX,P} \) in M3. As a result, P-ch.-1 and P-ch.-2 show faster program performance at the same PWL/WWL voltage levels of -7.6V as shown in Fig. 5. Similarly, among the three N-channel configurations in Table II, the highest \( V_{OX,E} \) and \( V_{OX,P} \) are expected for N-ch.-1. In summary, our results show that N-ch.-1 and P-ch.-1 are the best 5T eflash cell choices in terms of the erase and program speed among the 6 configurations shown in Table II. Note that an N-channel cell has fast program performance while a P-channel cell has fast erase performance.

**Table II. Configurations of Various 5T Eflash Cells**

<table>
<thead>
<tr>
<th>Configuration</th>
<th>( M_1 )</th>
<th>( M_2 )</th>
<th>( M_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-ch. 5T Eflash</td>
<td>PMOS</td>
<td>PMOS</td>
<td>NMOS</td>
</tr>
<tr>
<td>P-ch. 5T Eflash</td>
<td>NCAP*</td>
<td>NCAP*</td>
<td>NMOS</td>
</tr>
<tr>
<td>PCAP**</td>
<td>PCAP**</td>
<td>PCAP**</td>
<td></td>
</tr>
</tbody>
</table>

**Table III. Program/Erase Speed of Various 5T Eflash Cells**

<table>
<thead>
<tr>
<th>Configuration</th>
<th>( V_{OX,E}(V) )</th>
<th>( V_{OX,P}(V) )</th>
<th>( T_E(\text{ms}) )</th>
<th>( T_P(\text{ms}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-ch. 5T Eflash</td>
<td>7.06</td>
<td>7.24</td>
<td>1</td>
<td>0.002</td>
</tr>
<tr>
<td>P-ch. 5T Eflash</td>
<td>6.90</td>
<td>7.37</td>
<td>0.004</td>
<td>75</td>
</tr>
<tr>
<td>N-ch. 3</td>
<td>N.A.</td>
<td>6.43</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
<tr>
<td>P-ch. 3</td>
<td>N.A.</td>
<td>7.35</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>P-ch. 3</td>
<td>N.A.</td>
<td>6.80</td>
<td>3</td>
<td>2000</td>
</tr>
</tbody>
</table>

* ** PCAP: capacitor having P-type body and P-type gate

* ** PCAP: capacitor having P-type body and P-type gate

---

**Fig. 2. Bias conditions and read timing of (left) N-channel [7] and (right) P-channel 5T eflash cells. Boosted electron and hole channel voltages (\( V_{CHN} \) and \( V_{CHP} \)), and boosted channel sub-threshold and junction leakages (\( I_{SUB} \) and \( I_J \)) are illustrated in program inhibited BL cells.**
Fig. 3. M₁-M₂-M₃ cross sections of various P-channel 5T eflash memory cells.

Fig. 4. Energy band diagrams of M₂ and M₃ transistors in various P-channel 5T eflash cells during erase and program operations.

Fig. 5. Measured erase and program speeds of various P-channel 5T eflash memory cells.
B. Endurance and Retention Characteristics

Fig. 6 (top) shows the measured endurance characteristic of the different P-channel 5T eflash configurations where a negative cell $V_{TH}$ shift is observed. P-ch.-1 shows the least amount of $V_{TH}$ shift. Similarly, N-channel cells show a cell $V_{TH}$ shift in the positive direction [7]. Some of the activated carriers can be trapped in the $M_2$ gate oxide of 5T eflash. Fig. 6 (bottom) shows the larger variations in the erased cell $V_{TH}$ for P-ch.-2 and P-ch.-3, further reducing the sensing margin for higher P/E cycles. Those large variations of the erased cell $V_{TH}$ for P-ch.-2 and P-ch.-3 can be attributed to the large variation in the depletion capacitances of $M_2$ during erase operations. The measured retention result of various P-channel 5T eflash memories in Fig. 7 (top) shows the least cell $V_{TH}$ shift for P-ch.-3. One of the possible reasons is that the coupling device having p+ poly dominantly reduces the gate leakage current. This is because the Fermi level in the p+ poly silicon is close to the valence band edge which in turn reduces the number of conduction band electrons participating in the charge loss process as shown in Fig. 7 (bottom) [10].

![Fig. 6](image_url)

**Fig. 6.** Measured endurance and cell $V_{TH}$ distributions of various P-channel 5T eflash memory cells.

![Fig. 7](image_url)

**Fig. 7.** (top) Measured retention result of three different 5T eflash cells. (bottom) Energy band diagram explaining least charge loss of P-ch.-3 type 5T eflash.

IV. Disturbance and Coupling Effect

A. Program Disturbance Effect

From the bias conditions of a program inhibited cell shown in Fig. 2, the boosted channel voltage ($V_{CHN}$ and $V_{CHP}$) should be kept high with suppressed sub-threshold and junction leakage currents to prevent program disturbance in an unselected cell [11]. The select transistor ($S_1$) utilizes a longer channel length to minimize the sub-threshold leakage of the boosted channels. A program voltage margin of $-4V$ and a negligible cell $V_{TH}$ disturbance up to $\sim 1s$ program pulse were measured for 10k pre-cycled cells as shown in Fig. 8. This confirms the effectiveness of the self-boosting technique in a standard logic technology, which allows the row-by-row program/erase array architecture without making disturbance issue of the unselected WL cells.

B. Floating Gate Coupling Effect

The tighter BL pitch and shorter FG coupling distance compared to other single poly eflash [4, 6] may increase the parasitic inter-FG coupling effect for the 5T eflash cells. For
characterizing this coupling effect, the even BL’s were programmed first and subsequently the odd BL’s were programmed thereby making the even BL cells victims affected by the floating gate coupling when the odd BL’s are programmed (Fig. 9). The measured result shows a modest change in the mean and standard deviation of the cell V_{TH} distribution (17mV and 3.7% respectively) due to the FG coupling effect.

C. Multi-Level Cell Feasibility

Feasibility of a Multi-Level Cell (MLC) operation was also investigated. For this purpose, a group of cells is programmed to a P3 state using a single program pulse, and then the other groups are sequentially programmed to P2 and P1 states using a balanced Incremental Step Pulse Programming (ISPP) scheme [8] with a 0.1V step increment as illustrated in Fig. 10 (top). The minimal program disturbance and FG coupling effect enables precise programming for P1 and P2 states without affecting the cell V_{TH} values of P3 state as shown in Fig. 10 (middle). The final programmed distribution and retention characteristics of N-ch.-1 type 5T eflash cells in Fig. 10 (bottom) shows four distinct states having a good sensing margin of 0.4V for a 150°C baking temperature and 100 P/E pre-cycle case.

Fig. 8. Measured program disturbance characteristics of 5T eflash.

Fig. 9. 5T eflash array layout [7] and measured FG coupling.

V. CONCLUSION

Moderate-density single-poly eflash memory is an attractive eNVM candidate for SoC designs where a dedicated eflash process is not available. This type of eflash can be built using standard I/O devices that are readily available in a generic logic process. In this work, various single-poly eflash cell topologies were fabricated and characterized. Measured data shows that N-channel cells with a PMOS-PMOS-NMOS...
topology (i.e. N-ch.-1) provides the optimal balance between program/erase speed, endurance, and retention characteristics, while supporting self-boosting with minimal program disturbance and FG coupling issues. Two 65nm test chips (Fig. 11) were fabricated as part of this study.

Fig. 11. Die microphotographs and feature summary of the two 65nm eflash test chips fabricated as part of this work.

REFERENCES