Program/Erase 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)

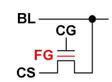
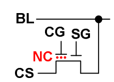
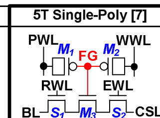
***Abstract*— N-channel and P-channel single-poly embedded flash (eflash) memory cells were implemented in a standard CMOS logic process. Among the different configurations based on standard I/O devices, the N-channel cell with a PMOS-PMOS-NMOS combo and the P-channel cell with an NMOS-NMOS-PMOS combo 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 1) 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 (~700μm^2) 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.

TABLE I. EFLASH MEMORY OPTIONS

Cell Type	1T Dual-Poly [2]	1.5T Split-Gate [3]	5T Single-Poly [7]
Cell Schematic			
Process Overhead	Floating Gate	Nanocrystal	None
Min. Tunnel Oxide	10nm (Dedicated)	5nm (Dedicated)	5nm (Standard I/O)
Program Method	CHE Injection	SS Injection	FN Tunneling
Erase Method	FN Tunneling	FN Tunneling	FN Tunneling
Cell Size	0.44μm^2 in 90nm	N. A.	8.62μm^2 in 65nm

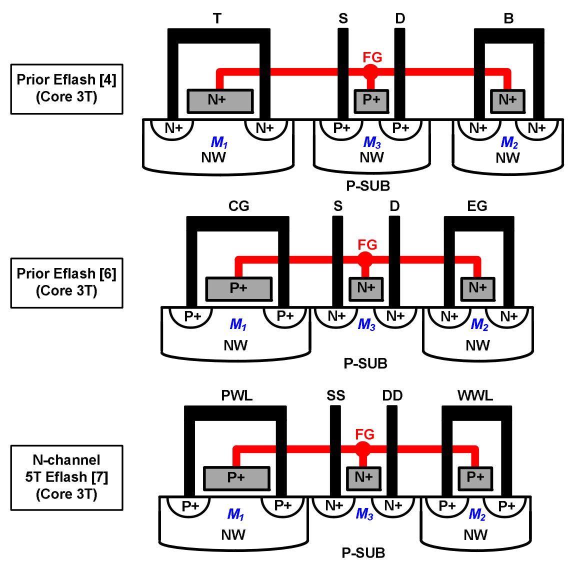


Fig. 1. Various single-poly eflash cell cross sections.

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 M_2 gate oxide during erase operation and the M_3 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.

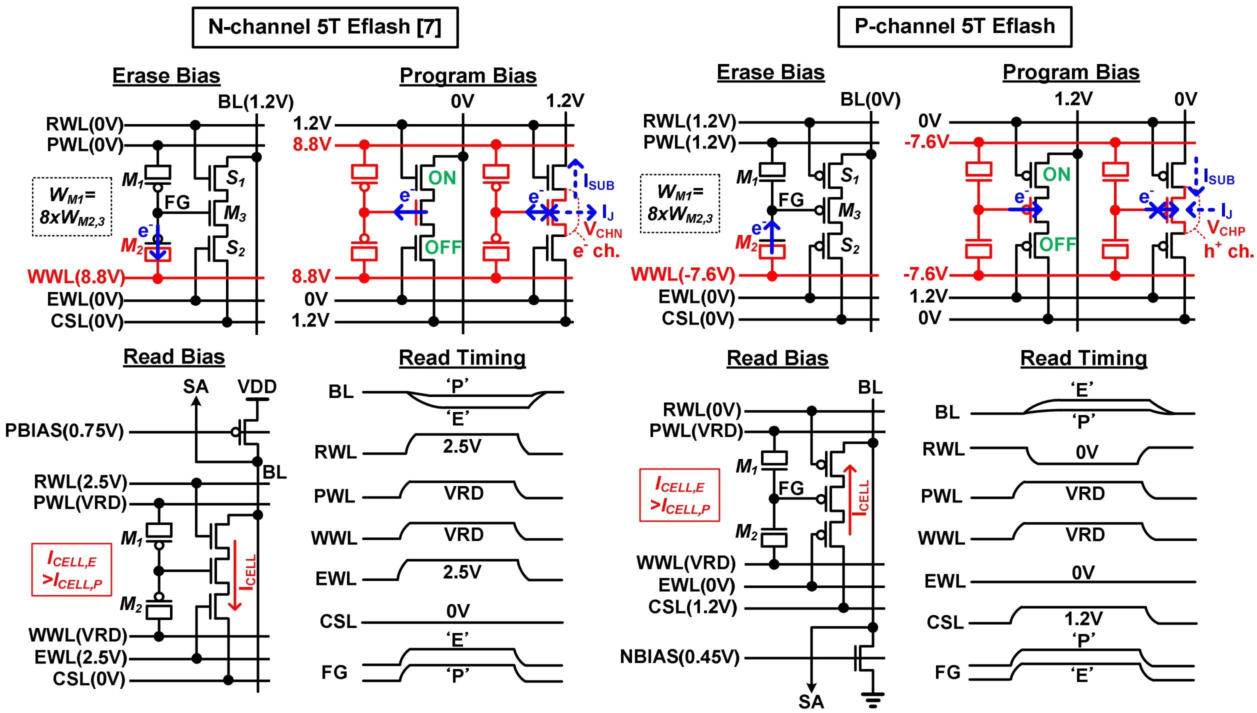


Fig. 2. Bias conditions and read timing of (left) N-channel [7] and (right) P-channel 5T eflash cells. Boosted electron and hole channel voltages (V_{CHN} and V_{CHP}), and boosted channel sub-threshold and junction leakages (I_{SUB} and I_J) are illustrated in program inhibited BL cells.

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 M_1-M_3. All cell types can be implemented in a standard CMOS process. The device sizing ratio $W_{M1}/W_{M2,3}$ is set as 8 for all designs. Cross section of M_1-M_3 for the different P-channel 5T eflash configurations and the corresponding energy band diagrams of M_2 and M_3 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 (V_{OX_E} and V_{OX_P}) were simulated with typical device parameters and writing voltages. The erase and program speeds (T_E and T_P) were measured from the fabricated 65nm test chips for a target cell V_{TH} window of 1V and summarized in Table III. During erase operation, the erase device M_2 of P-ch.-1 operates in an inversion mode, thereby producing a high V_{OX_E} in M_2. On the other hand, M_2 of P-ch.-2 and P-ch.-3 operates in a depletion mode, producing a low V_{OX_E} in M_2 as considerable portion of the WWL voltage is applied to the depletion region underneath the tunnel oxide. As a result, the erase speeds of P-ch.-2 and P-ch.-3 are ~1000× slower than P-ch.-1 at the same WWL voltage level of -7.6V as shown in Fig. 5 (top). The erase speed of P-ch.-3 is faster than that of P-ch.-2 as the coupling transistor M_1 of P-ch.-3 (=PCAP) falls quickly into the accumulation mode, producing a higher coupling effect between PWL and FG. During program operation, the coupling device M_1 of P-ch.-1 and P-ch.-2 (=NMOS) operates in an inversion mode, producing a high V_{OX_P} in M_3, while M_1 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 (bottom). 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

Configuration		M_1	M_2	M_3
N-ch. 5T Eflash	N-ch.-1	PMOS	PMOS	NMOS
	N-ch.-2	PMOS	NCAP*	NMOS
	N-ch.-3	NCAP*	NCAP*	NMOS
P-ch. 5T Eflash	P-ch.-1	NMOS	NMOS	PMOS
	P-ch.-2	NMOS	PCAP**	PMOS
	P-ch.-3	PCAP**	PCAP**	PMOS

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

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

Configuration		V_{OX_E}(V)*	V_{OX_P}(V)**	T_E(ms)***	T_P(ms)***
N-ch. 5T Eflash	N-ch.-1	7.06	7.24	1	0.002
	N-ch.-2	N. A.	7.21	N. A.	N. A.
	N-ch.-3	N. A.	6.43	N. A.	N. A.
P-ch. 5T Eflash	P-ch.-1	6.90	7.37	0.004	75
	P-ch.-2	N. A.	7.35	5	60
	P-ch.-3	N. A.	6.80	3	2000

* WWL=8.8V, PWL=0V for N-ch. 5T Eflash
* WWL=-7.6V, PWL=1.2V for P-ch. 5T Eflash
** WWL=PWL=8.8V, BL(selected)=0V for N-ch. 5T Eflash
** WWL=PWL=-7.6V, BL(selected)=1.2V for P-ch. 5T Eflash
*** Measured erase and program speeds for ΔV_{TH}=1V

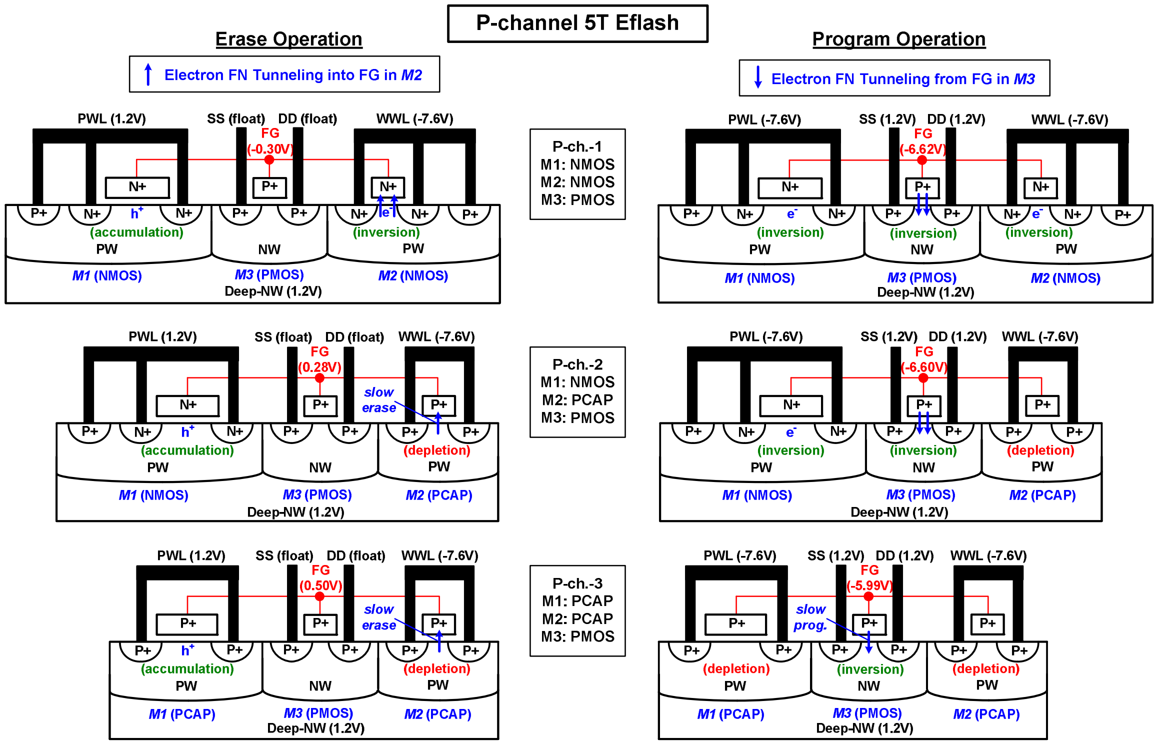


Fig. 3. M_1-M_2-M_3 cross sections of various P-channel 5T eflash memory cells.

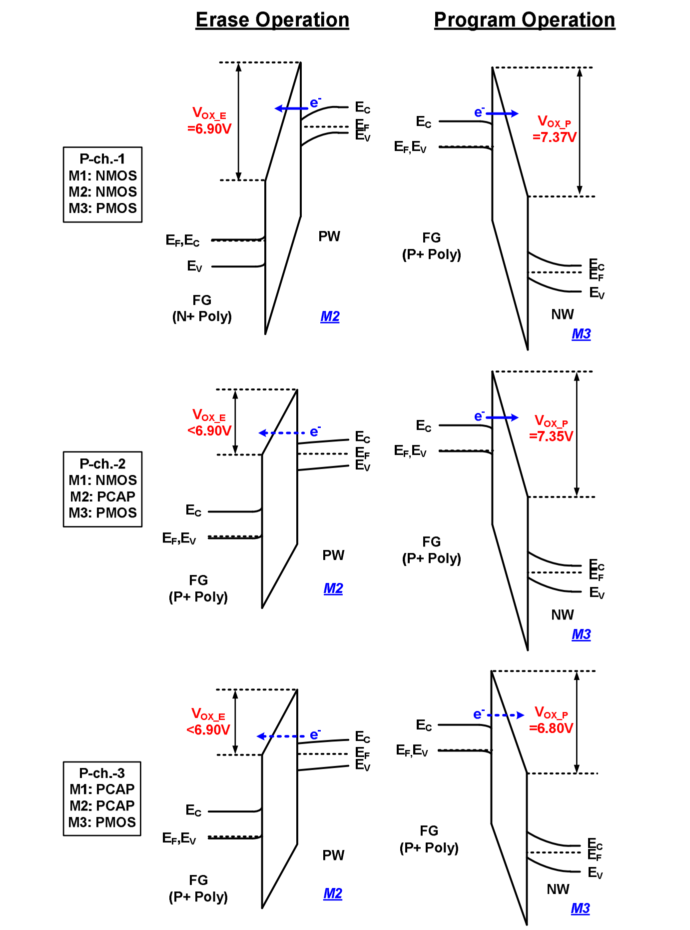


Fig. 4. Energy band diagrams of M_2 and M_3 transistors in various P-channel 5T eflash cells during erase and program operations.

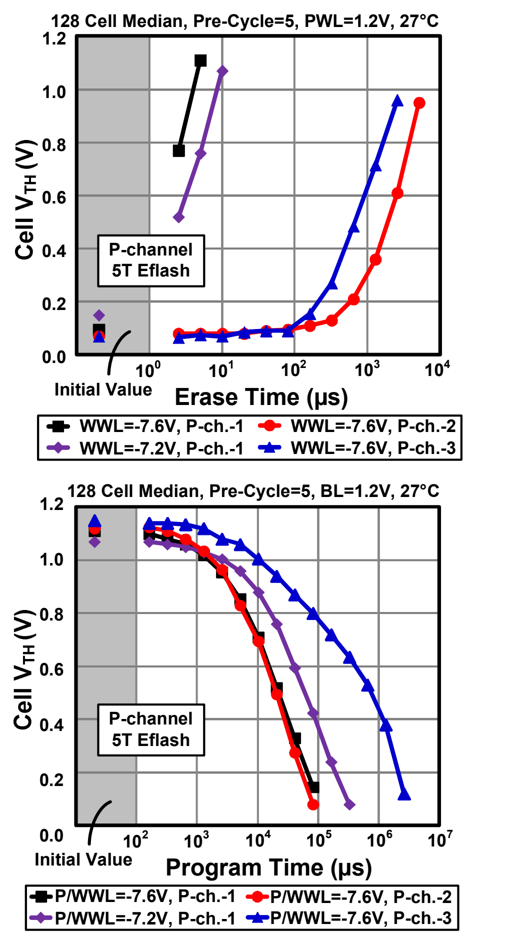


Fig. 5. Measured erase and program speeds of various P-channel 5T eflash memory cells.

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_3 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].

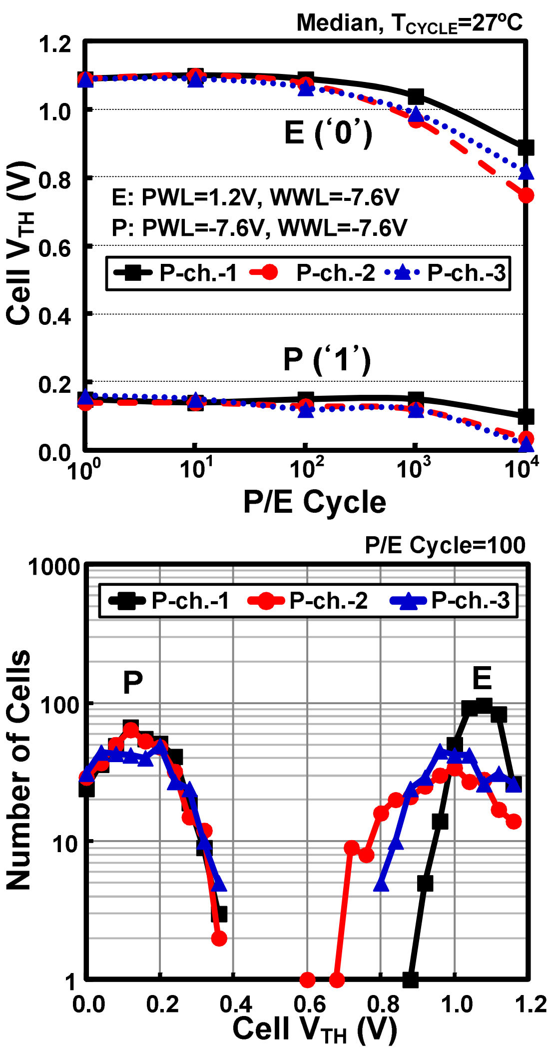


Fig. 6. Measured endurance and cell V_{TH} distributions of various P-channel 5T eflash memory cells.

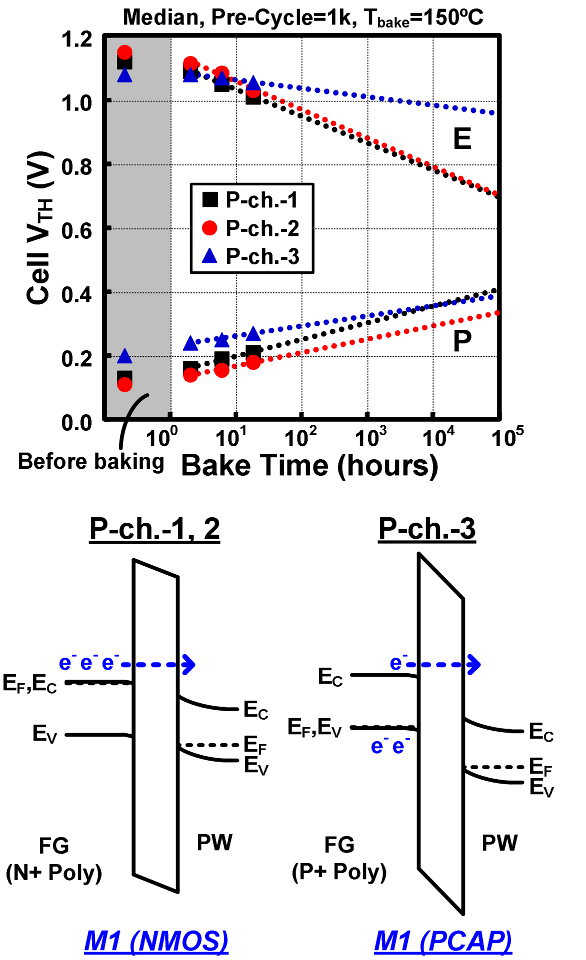


Fig. 7. (top) Measured retention result of three different 5T eflash cells. (bottom) Energy band diagram explaining least charge loss of P-ch.-3 type 5T eflash.

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 a ~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.

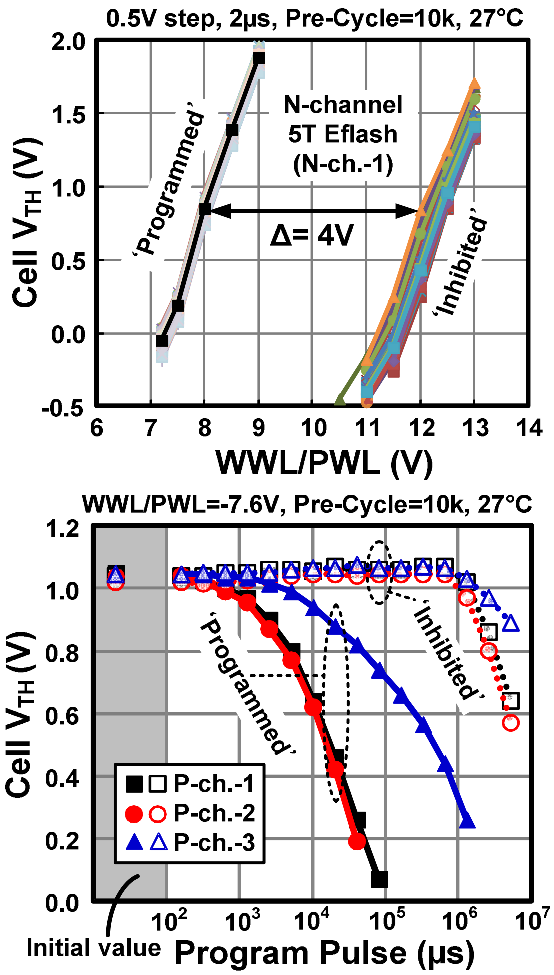


Fig. 8. Measured program disturbance characteristics of 5T eflash.

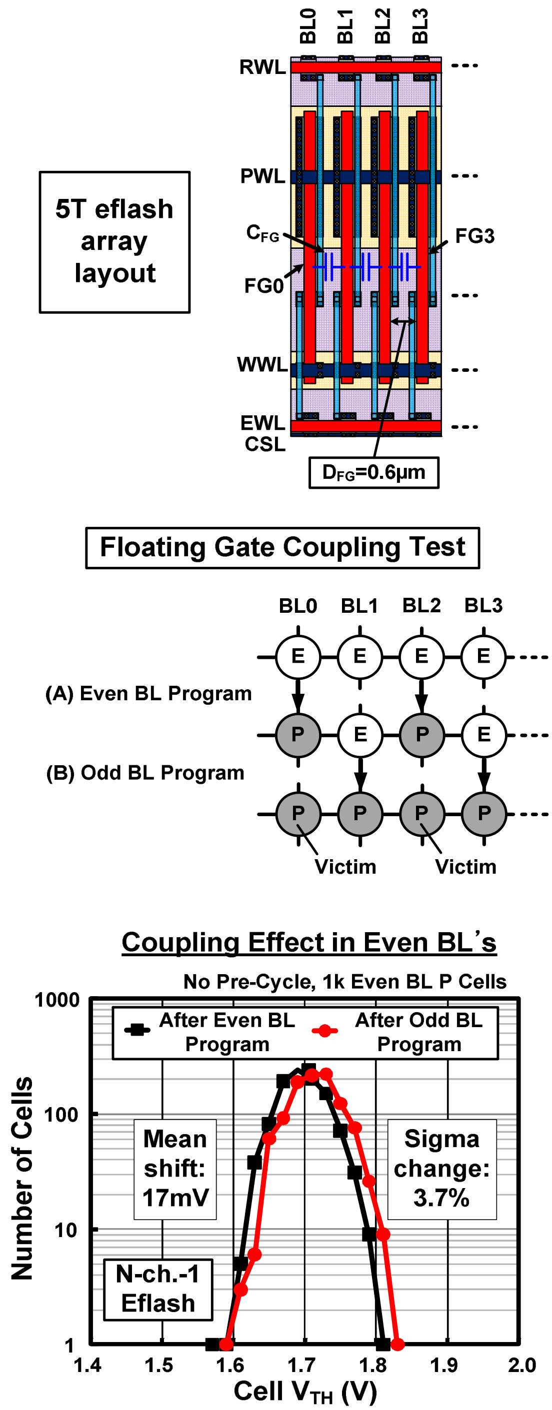


Fig. 9. 5T eflash array layout [7] and measured FG coupling.

V. CONCOLUSION

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.

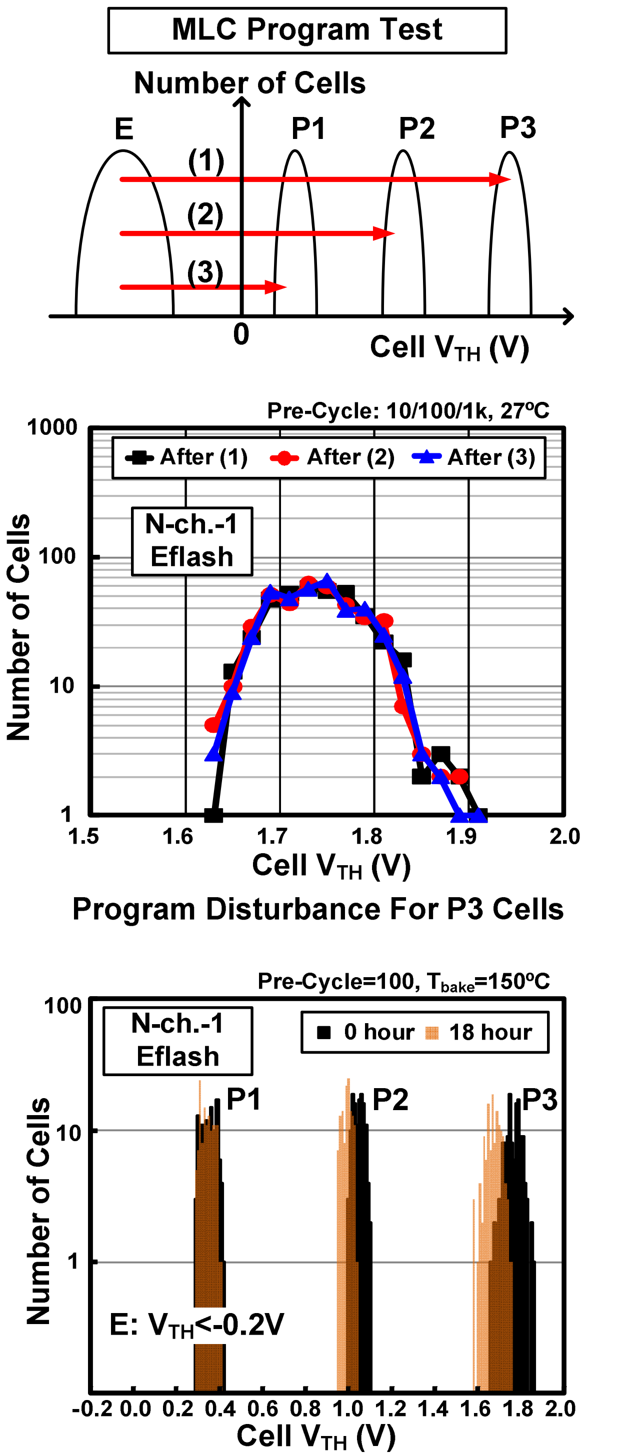
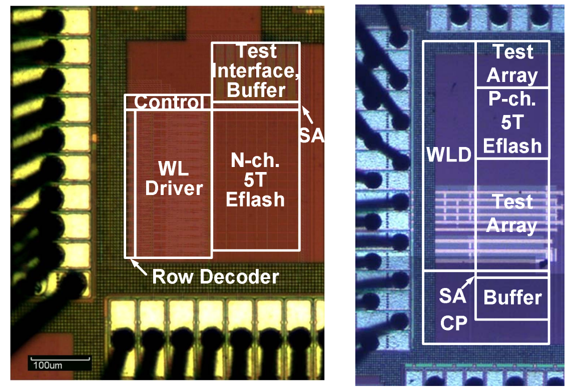


Fig. 10. (top) MLC program test sequence, (middle) MLC program test results showing negligible disturbance/coupling issues. (bottom) MLC retention characteristics of n-ch.-1 5T eflash cells.



Process	65nm LP CMOS
Features	N-ch. 5T Eflash [7], P-ch. 5T Eflash
Supply Voltage	1.2V (Core), 2.5V (I/O)
Cell Transistor	2.5V I/O Device
Tunnel Oxide	5nm
Cell Size	8.62μm^2
Array Dimension	16 WL x 128 BL (N-ch.) 10 WL x 128 BL (P-ch.)

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

REFERENCES

[1] R. Strenz, "Embedded Flash Technologies and their Applications: Status & Outlook," *IEEE Int. Electron Devices Meeting (IEDM)*, pp. 211-214, 2011.

[2] H. Kojima, T. Ema, T. Anezaki, J. Ariyoshi, H. Ogawa, et al., "Embedded Flash on 90nm Logic Technology & Beyond for FPGAs," *IEEE Int. Electron Devices Meeting (IEDM)*, pp. 677-680, 2007.

[3] J. Yater, M. Suhail, S. Kang, J. Shen, C. Hong, et al., "16Mb Split Gate Flash Memory with Improved Process Window," *IEEE Int. Memory Workshop (IMW)*, pp. 1-2, 2009.

[4] J. Raszka, M. Advani, V. Tiwari, L. Varisco, N. Hacobian, et al., "Embedded Flash Memory for Security Applications in a 0.13μm CMOS Logic Process," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 46-47, 2004.

[5] B. Wang, H. Nguyen, Y. Ma, R. Paulsen, "Highly Reliable 90-nm Logic Multitime Programmable NVM Cells Using Novel Work-Function-Engineered Tunneling Devices," *IEEE Trans. on Electron Devices*, vol. 54, no. 9, pp. 2526-2530, September 2007.

[6] Y. Yamamoto, M. Shirahama, T. Kawasaki, R. Nishihara, S. Sumi, et al., "A PND (PMOS-NMOS-Depletion MOS) Type Single Poly Gate Non-Volatile Memory Cell Design with a Differential Cell Architecture in a Pure CMOS Logic Process for a System LSI," *IEICE Trans. Electron.*, vol. E90-C, no. 5, pp. 1129-1137, May 2007.

[7] S. Song, K. Chun, C. H. Kim, "A Logic-Compatible Embedded Flash Memory Featuring a Multi-Story High Voltage Switch and a Selective Refresh Scheme," *IEEE Symp. on VLSI Circuits*, pp. 130-131, 2012.

[8] K. Suh, B. Suh, Y. Lim, J. Kim, Y. Choi, et al., "A 3.3V 32Mb NAND Flash Memory with Incremental Step Pulse Programming Scheme," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 128-129, 1995.

[9] T. Jung, Y. Choi, K. Suh, B. Suh, J. Kim, et al., "A 3.3V 128Mb Multi-Level NAND Flash Memory for Mass Storage Applications," *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pp. 32-33, 1996.

[10] Y. Shi, T. Ma, S. Prasad, S. Dhanda, "Polarity Dependent Gate Tunneling Currents in Dual-Gate CMOSFET's," *IEEE Trans. on Electron Devices*, vol. 45, no. 11, pp. 2355-2360, November 1998.

[11] S. Satoh, H. Hagiwara, T. Tanzawa, K. Takeuchi, R. Shirota, "A Novel Isolation-Scaling Technology for NAND EEPROMs," *IEEE Int. Electron Devices Meeting (IEDM)*, pp. 291-294, 1997.