

Design of Logic-Compatible Embedded Flash Memories for Moderate Density On-Chip
Non-Volatile Memory Applications

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Seung-Hwan Song

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chris H. Kim, Hubert H. Lim

December 2013

© Seung-Hwan Song 2013

Acknowledgements

First, I thank Professor Chris H. Kim for making my PhD study very rewarding experience with this dissertation. It is my great fortune to have him as my PhD advisor, since this dissertation would not have been even started at all if I had not met him here in University of Minnesota. Indeed, I never expected that I would research on flash memory for my PhD dissertation, when I first came to Minnesota in Fall 2009. Moreover, I was so skeptical and frustrated about this research topic until Spring 2012. Thus, this dissertation would not have been finished if it had not been for his continuous encouragement and patience. Professor Chris H. Kim therefore should be appreciated for his support over more than last four years on this dissertation. I also thank him for allowing me to explore the interdisciplinary study and multiple internships at Broadcom, Qualcomm, and Seagate. It helped me to learn broad disciplines and have practical insights that would be beneficial for the advanced engineering research topics.

I also like to thank Professor Hubert H. Lim for serving as my co-advisor and giving me valuable feedback on my research. I would also want to express my gratitude to Professors Ramesh Harjani, Keshab Parhi, and Sachin Sapatnekar for serving as

committee members of my final and preliminary oral exams. I appreciate their time out of their busy schedules to review this dissertation and provide valuable comments.

I also wish to thank Professor Byung-Gook Park for me to enter the semiconductor research community through my master study with him at Seoul National University, which has been a strong basis for me to conduct research on semiconductor later in Samsung and here in University of Minnesota. I also thank Dr. Jun Jin Kong at Samsung who recommended me to study in Minnesota which was so wonderful place to me that could continue the professional career in a memory industry.

I also thank all my former and current colleagues in VLSI research lab for all their supports given to me in the lab. Especially, I thank Ki Chul Chun and Jongyeon Kim for their contribution to this dissertation. The technical discussion with them for years enabled me to finish this dissertation in this end. Also, I have to thank Tony Kim, John Keane, Dong Jiao, Wei Zhang, Pulkit Jain, Xiaofei Wang, Ayan Paul, Bongjin Kim, Wonho Choi, Weichao Xu, Saroj Satapathy, Qianying Tang, Somnath Kundu, Chen Zhou, and Hoonki Kim for all their contributions they created in the lab.

I also would like to thank my English teachers Tom, Polly, and Caroline for their warm care to me. My research presentations would not have been appreciated by audiences without their face-to-face advice on my English pronunciation. I also have to mention Yejin who had introduced Tom to me. I thank her with my best wish for her and her family. Also, my life in Minnesota would not have been delightful without many unlisted Korean friends here. Thank them all so much.

I also thank my parents for their encouragement of my PhD studying in US. Now, I am likely to know that they had to sacrifice their lives with their only child for his future. I have nothing to say I am so sorry about that. I was very happy that they could come to see my first PhD presentation in Hawaii. Whenever I attend Symposium on VLSI Circuits as a world-wide researcher later, I will recall them. Thank them so much. I also want to thank my parents-in-law for their continuous prays on my success of the PhD research. Without their prays, all the achievements happened in the end of my PhD study would not have occurred. I also thank my sister-in-law and brother-in-law. Their encouragements have cheered up me and my wife over the years.

Finally, I thank my lovely wife Ji Hye Hur for her continuous support to my study. This dissertation would not have been made if she had not sacrificed her professional career at Samsung Electronics. I also feel so sorry that our life has been always like exploring the wonder world, since we married each other in 2008. I hope she likes our next journey with our adorable soon-to-be-born twins, Teun-Teun and Tan-Tan. I wish they would be grown to be healthy and smart, and later could find this lovely twin city campus again together where their parents spent their early thirties making wonderful memories.

Abstract

An on-chip embedded NVM (eNVM) enables a zero-standby power system-on-a-chip with a smaller form factor, faster access speed, lower access power, and higher security than an off-chip NVM. Differently from the high density eNVM technologies such as dual-poly eflash, FeRAM, STT-MRAM, and RRAM that typically require process overhead beyond standard logic process steps, the moderate density eNVM technologies such as e-fuse, anti-fuse, and single-poly embedded flash (eflash) can be fabricated in a standard logic process with no process overhead. Among them, a single-poly eflash is a unique multiple-time programmable moderate density eNVM, while it is expected to play a key role in mitigating variability and reliability issues of the future VLSI technologies; however, the challenges such as a high voltage disturbance, an implementation of logic compatible High Voltage Switch (HVS), and a limited cell sensing margin are required to be solved for its implementation using a standard I/O device. This thesis focuses on alleviating such challenges of the single-poly eflash memory with three single-poly eflash designs proposed in a generic logic process for moderate density eNVM applications.

Firstly, the proposed 5T eflash features a WL-by-WL accessible architecture with no disturbance issue of the unselected WL cells, an overstress-free multi-story HVS expanding the cell sensing margin, and a selective WL refresh scheme for the higher cell endurance. The most favorable eflash cell configuration is also studied when the performance, endurance, retention, and disturbance characteristics are all considered. Secondly, the proposed 6T eflash features the bit-by-bit re-write capability for the higher overall cell endurance, while not disturbing the unselected WL cells. The logic compatible on-chip charge pump to provide the appropriate high voltages for the proposed eflash operations is also discussed. Finally, the proposed 10T eflash features a multi-configurable HVS that does not require the boosted read supplies, and a differential cell architecture with improved retention time. All these proposed eflash memories were implemented in a 65nm standard logic process, and the test chip measurement results confirmed the functionality of the proposed designs with a reasonable retention margin, showing the competitiveness of the proposed eflash memories compared to the other moderate density eNVM candidates.

Table of Contents

List of Tables	X
List of Figures	xi
Chapter 1 Introduction	1
1.1 Embedded Non-Volatile Memory	2
1.2 Single-Poly Embedded Flash Memory	6
1.3 Summary of Thesis Contributions	10
Chapter 2 A Logic-Compatible 5T Embedded Flash Featuring a Multi- Story High Voltage Switch and a Selective WL Refresh Scheme ...	11
2.1 Overview of the Embedded Flash Memory Candidates	12
2.2 Proposed Logic-Compatible 5T Eflash.....	16
2.2.1 Overall Eflash Memory Architecture.....	16
2.2.2 Proposed 5T Eflash Cell Operation	17
2.2.3 Proposed Multi-Story High-Voltage Switch.....	21
2.2.4 Selective WL Refresh Scheme.....	26

2.3 Test Chip Measurement Results	30
2.3.1 Initial Cell V_{TH} , and Writing Speed	30
2.3.2 Endurance and Retention	33
2.3.3 Effectiveness of Refresh Operation	37
2.4 Comparison With Other Single-Poly Eflash.....	39
2.5 Chapter Summary	41

Chapter 3 Study on the Optimal Configuration of Single-Poly

Embedded Flash Cell.....	43
3.1 Single-Poly Embedded Flash Cell Configurations	44
3.2 Electron Ejection and Injection Operations in Single-Poly Eflash Cells.....	48
3.3 Program/Erase Speed, Endurance, Retention, and Disturbance Characteristics of Eflash Cells.....	50
3.3.1 Single-Poly Eflash Cells Fabricated in a 65nm Standard Logic Process.....	50
3.3.2 Program and Erase Speed of Single-Poly Eflash Cells.....	51
3.3.3 Endurance and Retention Characteristics	56
3.3.4 Program Disturbance	60
3.3.5 Read Disturbance	62
3.3.6 Floating Gate Coupling Effect	64
3.4 Chapter Summary	66

Chapter 4 A Bit-by-Bit Re-Writable 6T Eflash in a Generic Logic

Process.....	68
---------------------	-----------

4.1 WL-by-WL and Bit-by-Bit Erasable Single-Poly Eflash Memories	69
4.2 Proposed Bit-by-Bit Re-Writable 6T Eflash Memory	71
4.2.1 Bit-by-Bit Floating Gate Boosting Scheme	71
4.2.2 Bit-by-Bit Re-Writable Eflash Memory Overview.....	73
4.2.3 Cell Operation of the Proposed 6T Eflash Memory	74
4.3 Negative High Voltage Switch and Charge Pump.....	78
4.3.1 Negative High Voltage Switch	78
4.3.2 Negative Charge Pump	80
4.3.3 Junction Breakdown Issue of the Designed Negative HVS and CP.....	81
4.4 Test Chip Measurement Results	82
4.5 Comparison with Other CMOS Logic Embedded NVM Options.....	89
4.6 Chapter Summary	91

Chapter 5 10T Differential Eflash Featuring Multi-Configurable High

Voltage Switch with No Boosted Read Supplies92

5.1 Issues in 5T and 6T Eflash Memory Designs	93
5.2 Proposed Solutions	94
5.2.1 Multi-Configurable HVS	94
5.2.2 Differential Cell Architecture	95
5.3 Proposed 10T Differential Eflash memory	97
5.3.1 Overall Architecture.....	97
5.3.2 Proposed Multi-Configurable High Voltage Switch.....	99
5.3.3 Proposed VPP Switch and Charge Pump.....	104

5.3.4 Operation of the Proposed 10T Differential Eflash Memory Cell.....	106
5.4 Test Chip Design and Discussion	111
5.5 Chapter Summary	114
Chapter 6 Conclusions	116
Bibliography	119

List of Tables

Table 2.1 Single-Poly Eflash Comparison.....	39
Table 3.1 Summary of the Study on the Optimal Configuration of Single-Poly Eflash Cell	66
Table 4.1 Logic Compatible Embedded NVM Comparison (OTP)	90
Table 4.2 Logic Compatible Embedded NVM Comparison (Eflash).....	90
Table 5.1 Logic Compatible Embedded Flash Memory Comparison	115

List of Figures

Fig. 1.1 Various nonvolatile solutions: (a) Battery-backed SRAM, (b) Off-chip nonvolatile memory, and (c) On-chip nonvolatile memory.	2
Fig. 1.2 (a) Various eNVM technologies and their wide range of applications. (b, c) High and moderate density eNVM examples; dual-poly eflash and anti-fuse.	4
Fig. 1.3 (a) Dual-poly eflash cell transistor where the floating and control gates are stacked. (b) Single-poly eflash cell basic structure consisting of coupling, read, and write devices.	5
Fig. 1.4 Comparison of the dual-poly and single-poly eflash bias conditions for (a) erase, (b) program, and (c) read operations.	7
Fig. 1.5 Challenges of the single-poly eflash memory: (a) high voltage disturbance, (b) implementation of logic compatible high voltage switch, and (c) limited cell sensing margin after long retention time.	9
Fig. 2.1 Dual-poly and split gate eflash memory cells attractive for high density eNVM applications. Due to the process overhead to build floating gate or split gate structures, they are not attractive for the moderate density eNVM applications.	13

Fig. 2.2 Single-poly eflash memory cells attractive for the moderate density eNVM applications owing to their logic compatibility and lower writing voltage level compared to dual-poly or split-gate eflash.	14
Fig. 2.3 Array architecture and unit cell layout of the proposed logic-compatible 5T eflash memory. Only standard I/O and core devices are used.	17
Fig. 2.4 Bias conditions for erase and program operations of the proposed 5T eflash cell. A single WL write operation ensures that unselected WL's are protected from the high voltage levels.	19
Fig. 2.5 (top) Bias condition and (bottom) simulated timing diagram during read operation. Waveforms are from 1k Monte Carlo runs using a post-layout extracted netlist.	20
Fig. 2.6 (top) The prior HVS having a limited output range with reliability and variability concerns is compared to (bottom) the proposed HVS increasing VOUT up to 10V and providing robust internal voltage levels by utilizing multi-story stacked latches and 4× I/O VDD driver with additional VPP supplies.	23
Fig. 2.7 Simulated voltage and current waveforms with the proposed HVS and a voltage doubler based on-chip charge pump [29, 61-63]. Low-to-high and high-to-low transitions of WWL during program operation are shown.	24
Fig. 2.8 Low-to-high transition of WWL for read operation.	25
Fig. 2.9 Measured waveforms of the proposed HVS for three different read and write voltage levels with off-chip supplies ($V_{RD}=0/0.6/1.2V$, $V_{PP4}=6/8/10V$, $V_{PP3}=0.75 \times V_{PP4}$, $V_{PP2}=0.5 \times V_{PP4}$, $V_{PP1}=0.25 \times V_{PP4}$).	25

Fig. 2.10 Physical model explaining endurance and retention characteristics considering oxide and interface trap creation, interface trap annihilation, and trap-assisted charge loss [47-53].	27
Fig. 2.11 Proposed selective WL refresh scheme. Only identified “Weak” WL’s are refreshed to avoid unnecessary P/E cycles in the good cells.	28
Fig. 2.12 Laboratory setup for test chip measurement.	30
Fig. 2.13 Measured initial cell V_{TH} distributions of four 2kb eflash memory chips show cell-to-cell and chip-to-chip variations. The initial cell distribution ranges from 0.44 to 0.80V with an average value of 0.61V.	31
Fig. 2.14 Measured cell V_{TH} for different P/E voltages and pulse durations. Note that WWL and PWL were supplied by an off-chip voltage source to eliminate any non-ideal effects.	32
Fig. 2.15 Measured endurance characteristic. (a) Programmed cell V_{TH} shows a positive shift for P/E cycles greater than 1k. (b) Erased cell V_{TH} shows a similar positive shift for P/E cycles greater than 1k. (c) Cell current measured from a single cell exhibits severe sub-threshold slope degradation beyond 1k P/E cycles, implying that a considerable number of interface traps have been generated [49].	34
Fig. 2.16 Measured retention characteristic. (a, b) Cell V_{TH} distributions for 1k and 10k P/E pre-cycled cells at a 150°C bake temperature. (c) Spatial bit maps showing the cell V_{TH} shift of erased and programmed cells after 100/1k/10k P/E pre-cycles. (d) Cell V_{TH} shift vs. initial cell V_{TH} level for 100/1k/10k pre-cycles....	36

Fig. 2.17 Measured retention characteristic as a function of baking temperature. Three chips were baked at 27/85/150°C, respectively. The effects of charge loss and interface trap annihilation are canceled out for the erased cells, while a negative cell V_{TH} shift is observed in the programmed cells [52].	37
Fig. 2.18 Retention characteristics before and after refresh: (a) distribution, (b) sensing margin.....	38
Fig. 2.19 Die photograph of 2kb 5T eflash test chip implemented in 65nm standard logic process.	41
Fig. 3.1 (a) Bird's eye view of the single-poly eflash memory cell core structure consisting of the three standard I/O devices (M_1 - M_3). (b) The cross section and circuit symbol of the four available standard I/O devices forming the single-poly eflash memory cell core structure.	46
Fig. 3.2 Electron ejection/injection, and (b) read operations of eight different single-poly eflash cell configurations (2.5V I/O devices, $V_{DD}=1.2V$). In all configurations, the coupling device (M_1) is upsized compared to the other two devices (M_2, M_3) for efficient electron ejection and injection operations.	47
Fig. 3.3 (a, b) Simulated coupling ratios of eight different eflash configurations (i.e. type I-VIII). (c) Summary of the eflash cell configurations and the simulation results when the width ratio ($=W_{M1}/W_{M2,3}$) is 8.	49
Fig. 3.4 Die microphotographs of the two single-poly eflash test chips fabricated in a 65nm standard logic process [54-56].	51

Fig. 3.5 (a) Measured electron ejection speed of various types of eflash memory cells. (b) The energy diagrams of the electron ejection device and the operation modes of the coupling device, which explain the fastest electron ejection speed with the type V configuration and the slowest one with type VIII configuration..... 53

Fig. 3.6 (a) Measured electron injection speed of various types of eflash memory cells. (b) The energy diagrams and operation modes of the electron injection device explaining the fastest electron injection speed with the type VI configuration. . 55

Fig. 3.7 (a) Measured endurance characteristic of the cells having three different eflash configurations (Type VI-VIII). (b) Energy diagram of the electron injection device where the oxide traps modifies the shape of the tunnel barrier. (c) Measured cell V_{TH} distributions of the cells having three different eflash configurations after 100 P/E cycles..... 57

Fig. 3.8 Measured retention result of the cells having three different configurations (Type VI-VIII) after (a) 1k and (b) 3 P/E pre-cycles. (c) Energy band diagrams of NMOS (i.e. Type VI, VII) and PCAP (i.e. Type VIII) coupling devices explaining that the cells with the coupling device (M_I) having n-type poly-silicon has more conduction band electrons causing more intrinsic charge loss from the floating gate..... 59

Fig. 3.9 (a, b) Program bias conditions of 5T eflash cells with type I and VI configurations. (c, d) Measured program disturbance characteristics of the cells with type I and type VI-VIII configurations. 61

Fig. 3.10 (a) Read stress condition to measure the voltage-accelerated life time. (b) Definition of the life time. The cell V_{TH} tail reaches the sensing limit by the read stress at the end of the life time. (c) Measured life time of the 5T eflash cells with type I configuration. 63

Fig. 3.11 (a) 5T eflash cell array layout. (b) Floating gate coupling test sequence. (c) Measured floating gate coupling effect of the 5T eflash cells with type I configuration. 65

Fig. 3.12 The preferred 5T single-poly eflash cell with the type II configuration when the performance, endurance, retention, and disturbance characteristics are all considered..... 67

Fig. 4.1 Disturbance issue comparison between WL-by-WL and bit-by-bit erasable single-poly eflash cells. The prior WL-by-WL erasable eflash requires all cells in the selected WL to be erased simultaneously, which results in unnecessary erase cycles for cells whose data remain unchanged. The prior bit-by-bit erasable eflash requires a boosted BL voltage (V_{PP2}) for erase inhibition of the unselected BL cell. This will cause high voltage disturbance issues in the unselected WL. In contrast, the proposed eflash enables bit-by-bit erase via a novel FG boosting scheme (See Fig. 4.2) minimizing disturbance issues. 70

Fig. 4.2 Bias conditions of the coupling TR compared between the prior WL-by-WL and the proposed bit-by-bit FG boosting schemes. The former boosts all the FG's in the selected WL irrespective of the BL levels while the later selectively boosts the FG depending on the write data (i.e. $-V_{PP} < -V_H < -V_L$). 72

Fig. 4.3 Test chip diagram of the proposed bit-by-bit re-writable eflash memory. The 6T eflash cell array, multi-story high voltage switch, and multi-stage charge pump are implemented using standard 2.5V I/O devices with a 5nm gate oxide. The sense amplifiers and BL drivers are implemented using 1.2V core devices.	73
Fig. 4.4 Bit-by-bit (a) write ‘0’ and (b) write ‘1’ phases of the proposed 6T eflash cell. The ‘0’ BL cell loses electrons from FG during a write ‘0’ phase, whereas the ‘1’ BL cell adds electrons in FG during a write ‘1’ phase via electron FN tunneling.	75
Fig. 4.5 (a) Read bias condition of the proposed 6T eflash cell and (b) Timing diagram for the full bit-by-bit update sequence.	77
Fig. 4.6 (a) Multi-story negative high voltage switch consists of a stacked latch stage and a driver stage which prevents gate overstress during read and write operation. (b) During read, WWL is driven to VRD through the PMOS string without changing the latch states. (c, d) During write, WWL is switched between VPP4 and GND by the SEL signal.	79
Fig. 4.7 A negative charge pump generating multiple boosted negative voltage levels (VPP1-VPP4) is implemented in a 65nm standard logic process by cascading four voltage doubler stages.....	81
Fig. 4.8 Illustration of the junction breakdown issue in the proposed negative HVS and CP. Junction breakdown of the PMOS devices in the HVS bottom latch and the CP final stage limits the maximum negative VPP4 level.....	82

Fig. 4.9 Measured boosted negative voltages (VPP1-VPP4) and output characteristics of the negative charge pump.....	83
Fig. 4.10 Measured waveforms of the charge pump and high voltage switch.....	83
Fig. 4.11 Measured bit-by-bit update result from pattern (0101) to pattern (1100).	84
Fig. 4.12 (top) Measured cell V_{TH} shift from the 6T eflash test chip. Note that multiple write pulses with a fixed pulse width of 10 μ s were applied for the bit-by-bit write '0', whereas a single write pulse was applied for the bit-by-bit write '1'. (bottom) Different test patterns give different coupling between adjacent cells.	85
Fig. 4.13 Measured cell endurance and retention characteristics.	86
Fig. 4.14 (top) Average number of stress cycles for different data transitions and (bottom) cell V_{TH} transition plot for a high-to-high transition in the 5T eflash discussed in chapter 2 [54, 55] and the proposed 6T eflash cells.	87
Fig. 4.15 Overall endurance estimated based on the average stress cycle count in Fig. 4.14. A smaller word size (i.e. larger column multiplexing ratio) improves the overall endurance for the proposed 6T eflash.	88
Fig. 4.16 Die photograph of 4kb eflash test chip implemented in a 65nm generic logic process.	88
Fig. 5.1 Issues in 5T and 6T eflash memory designs: non-reconfigurable HVS and limited single cell sensing margin.	94
Fig. 5.2 Non-reconfigurable HVS in 5T and 6T eflash designs is compared to the multi-configurable HVS in a proposed eflash.....	95

Fig. 5.3 Sensing margin of the single cell architecture is compared to that of the differential cell architecture.....	96
Fig. 5.4 The proposed eflash memory consists of 10T differential cell array, multi-configurable high voltage switches, VPP switch and charge pump, differential current sense amplifiers, and other logic blocks implemented using the standard core and I/O devices in a generic logic process.	98
Fig. 5.5 The proposed multi-configurable high voltage switch.	100
Fig. 5.6 Operation of the proposed multi-configurable HVS during read mode.	101
Fig. 5.7 Operation of the proposed multi-configurable HVS during write mode.....	101
Fig. 5.8 The simulated waveforms of the proposed multi-configurable high voltage switch operations during read mode (VDDE=2.5V, VRD=1.2V, No Boosted Supply, Temp.=25°C): (a) low to high transition of WWL, (b) high to low transition of WWL.....	102
Fig. 5.9 The simulated waveforms of the proposed multi-configurable high voltage switch operations during write mode (VDDE=2.5V, VRD=1.2V, VPP1=2.2V, VPP2=4.4V, VPP3=6.6V, VPP4=8.8V, Temp.=25°C): (a) low to high transition of WWL, (b) high to low transition of WWL.	103
Fig. 5.10 The proposed VPP switch and the voltage doubler based charge pump.....	104
Fig. 5.11 (a) The simulated node voltage waveforms of the proposed VPP switch and charge pump which change the operation mode (left) from read to write, and (right) from write to read (VDD=1.2V, VDDE=2.5V, VRD=1.2V, VPP1=2.2V, VPP2=4.4V, VPP3=6.6V, VPP4=8.8V, Temp.=25°C). (b) The node voltage and	

transistor operation conditions corresponding to time 0 or 1100ns (i.e. read mode), and time 500 or 1000ns (i.e. write mode).	106
Fig. 5.12 The proposed 10T differential eflash cell, and its erase/program/read operation bias conditions.....	107
Fig. 5.13 The differential current sense amplifier and the read timing diagram of the proposed 10T differential eflash cell.....	109
Fig. 5.14 1k Monte Carlo run waveforms during read operation of the proposed 10T differential eflash cell.....	110
Fig. 5.15 Test chip layout and feature summary.....	111
Fig. 5.16 (left) Measured retention characteristic of 5T eflash cell at 150°C. (right) Measured single cell and emulated differential cell sensing margins from the worst case 100, 1k, 10k P/E pre-cycled 5T eflash cells.	112
Fig. 5.17 Illustrations of the proposed eflash operations and power consumption during each operation mode.....	113
Fig. 5.18 Illustration of the reliability aware system-on-chip architecture where the embedded flash memory can be used to store the reliability information that is changed infrequently during the entire device life-time.	114

Chapter 1 Introduction

Low power electronic devices typically employ standby mode to minimize leakage current [1, 2]. One important requirement during this standby mode is to retain critical information through a nonvolatile solution. Fig. 1.1 illustrates three different nonvolatile solutions. First solution is to use a battery-backed SRAM where a power gated SRAM is put in a data retention mode. This approach is based on readily available technology (e.g. SRAM and battery); however, the on-board backup battery increases the system complexity and cost while the SRAM still consumes leakage power even under a data retention voltage. Second or third solution is to use an off-chip or on-chip Non-Volatile Memory (NVM) which allows the system to completely shut down without losing the critical information as long as it is stored in the NVM, thereby achieving zero-standby power dissipation. The off-chip NVM can be fabricated in a dedicated process technology with higher capacity and reliability than the on-chip NVM; however, the off-chip NVM has a larger form factor, slower access speed and larger power dissipation for the off-chip communication, and more security concern than the on-chip NVM.

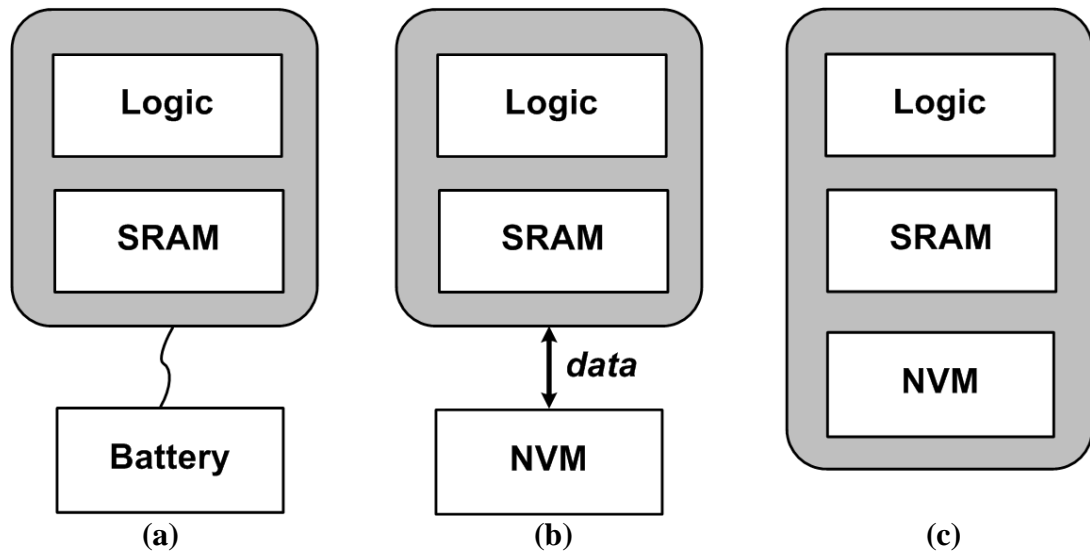


Fig. 1.1 Various nonvolatile solutions: (a) Battery-backed SRAM, (b) Off-chip nonvolatile memory, and (c) On-chip nonvolatile memory.

1.1 Embedded Non-Volatile Memory

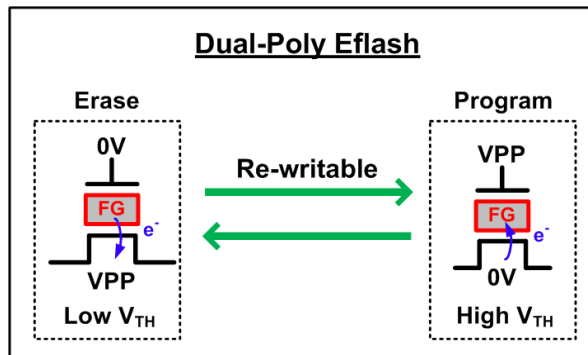
A number of on-chip embedded Non-Volatile Memory (eNVM) technologies have been developed for the high density (~Mb) and moderate density (~kb) applications as shown in Fig. 1.2. The dual-poly embedded flash (eflash) and anti-fuse are illustrated as high and moderate density eNVM examples, too. High density eNVM such as dual-poly, charge-trap or split-gate eflash memory, FeRAM, STT-MRAM, and RRAM technologies have such applications as code and data storage, and cache memories [3-16]; however, they typically require process overhead beyond a generic logic technology. For example, the dual-poly eflash supports multiple program and erase operations by changing the number of electrons stored in Floating Gate (FG) via electron tunneling into or out of the FG which is isolated from the other conducting nodes; however, this FG structure

requires a process overhead beyond the logic technology. The One-Time Programmable (OTP) memories such as e-fuse (electronic-fuse) and anti-fuse, on the other hand, can be built in a generic logic technology, and have been widely adopted as a moderate density eNVM for the memory redundancy scheme, circuit trimming, digital calibration, and secure ID storage [17-25]; however, they do not allow multiple-time program capability, since they use the permanent electro-migration or breakdown mechanism for a program operation. For example, the anti-fuse uses irreversible gate-oxide breakdown mechanism for a program operation to change the cell status from high impedance (i.e. high-Z) state to low impedance (i.e. low-Z) state. A single-poly eflash, on the other hand, is multiple-time programmable eNVM that can be built in a generic logic process [26-40]. In Fig. 1.3, the typical dual-poly eflash cell consisting of one transistor where the FG and Control Gate (CG) are stacked is compared to the single-poly eflash cell basic structure typically consisting of three devices called coupling, read, and write devices of which the gates are back-to-back connected to form the FG node; therefore, the single-poly eflash has no process overhead beyond generic logic technology. On the other hand, it uses the reversible writing mechanism, electron tunneling, which enables the multiple program and erase operations, making it attractive as a cost-effective moderate density eNVM. It is also expected to play a key role in mitigating variability and reliability issues of the future VLSI technologies that the traditional OTP memories cannot easily have solved. For example, adaptive self-healing techniques or on-line repair schemes can be applied to improve the device life time based on the reliability diagnostic data stored in the single-

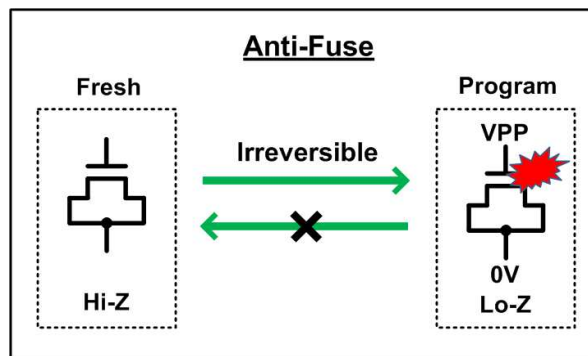
poly eflash memory, which is especially believed to be a critical technique in many future biomedical healthcare devices having more stringent reliability requirements [41-46].

	High Density eNVM (~Mb)	Moderate Density eNVM (~kb)
Technologies	Dual-Poly Eflash, FeRAM, STT-MRAM, RRAM	E-Fuse, Anti-Fuse, Single-Poly Eflash
Applications	Code Storage, Data Storage, Cache Memory	Redundancy Scheme, Circuit Trimming, Digital Calibration, Secure ID Storage

(a)



(b)



(c)

Fig. 1.2 (a) Various eNVM technologies and their wide range of applications. (b, c) High and moderate density eNVM examples; dual-poly eflash and anti-fuse.

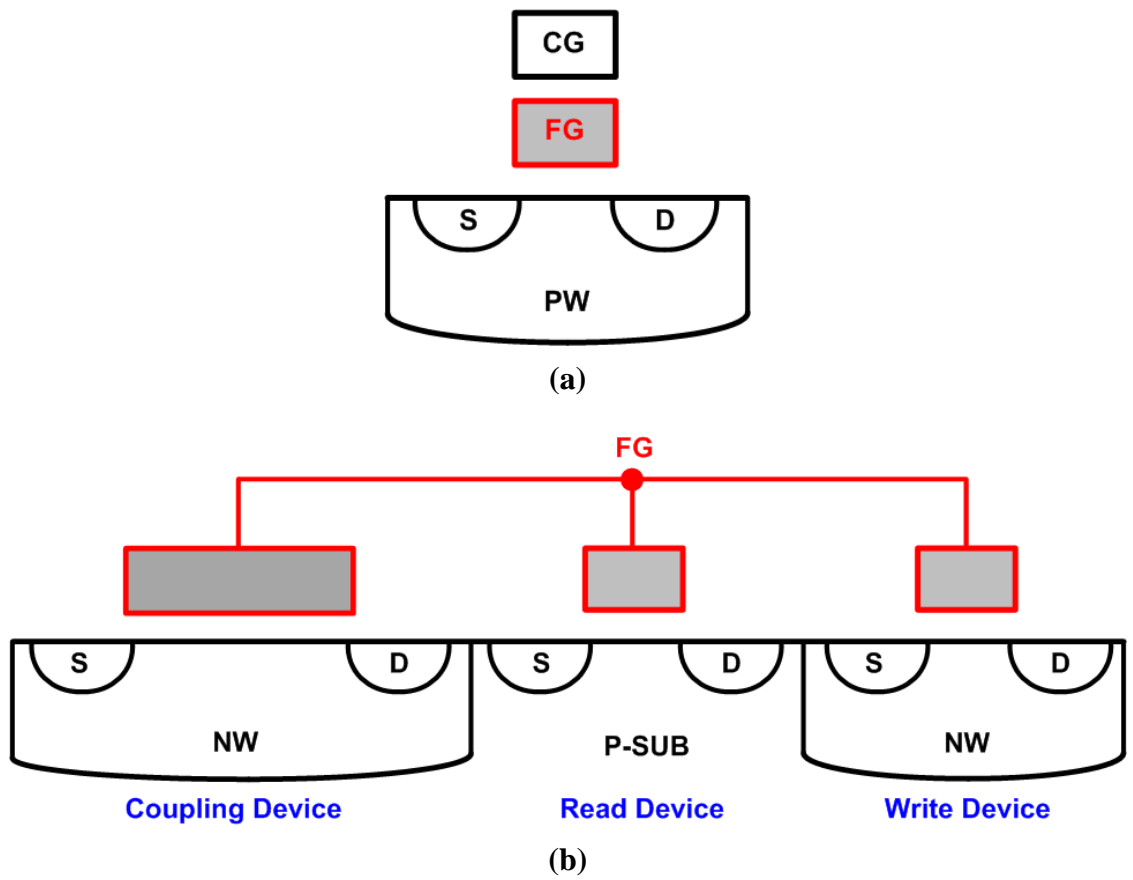


Fig. 1.3 (a) Dual-poly eflash cell transistor where the floating and control gates are stacked. (b) Single-poly eflash cell basic structure consisting of coupling, read, and write devices.

1.2 Single-Poly Embedded Flash Memory

In Fig. 1.4, the bias conditions of the dual-poly and single-poly eflash cells are compared for erase, program, and read operations with the coupling ratio of ~ 0.5 for dual-poly eflash cell, and ~ 1 for single-poly eflash cell. The n-well of the coupling device of the single-poly eflash cell functions like the CG of the dual-poly eflash cell; therefore the high coupling ratio between the n-well of the coupling device and FG can be readily achieved in the single-poly eflash cell by simply upsizing the coupling transistor, whereas the coupling ratio between CG and FG in dual-poly eflash cell is difficult to be more than 0.5 without a significant process overhead beyond the logic technology. During erase and program operations, the high electric field is generated in the gate oxide, which enables the electrons to be ejected from FG during erased operation and to be injected into FG during program operation, changing the number of electrons stored in FG. During read operation, the cell current (i.e. I_{CELL}) is sensed differently between the erased and programmed states because of the different number of electrons stored in FG resulting in the different conductance for the erased and programmed states. During all operations, the required n-well voltage levels of the single-poly eflash cell are smaller than the required CG voltage levels of the dual-poly eflash cell, because of the higher coupling ratio of the single-poly eflash. These lower voltage levels for erase, program, and read operations of the single-poly eflash compared to the dual-poly eflash reduce the power consumption and simplify the high voltage circuitry significantly, which is a highly attractive feature, potentially for the single-poly eflash to be embedded in low power electronic devices built in a standard logic technology.

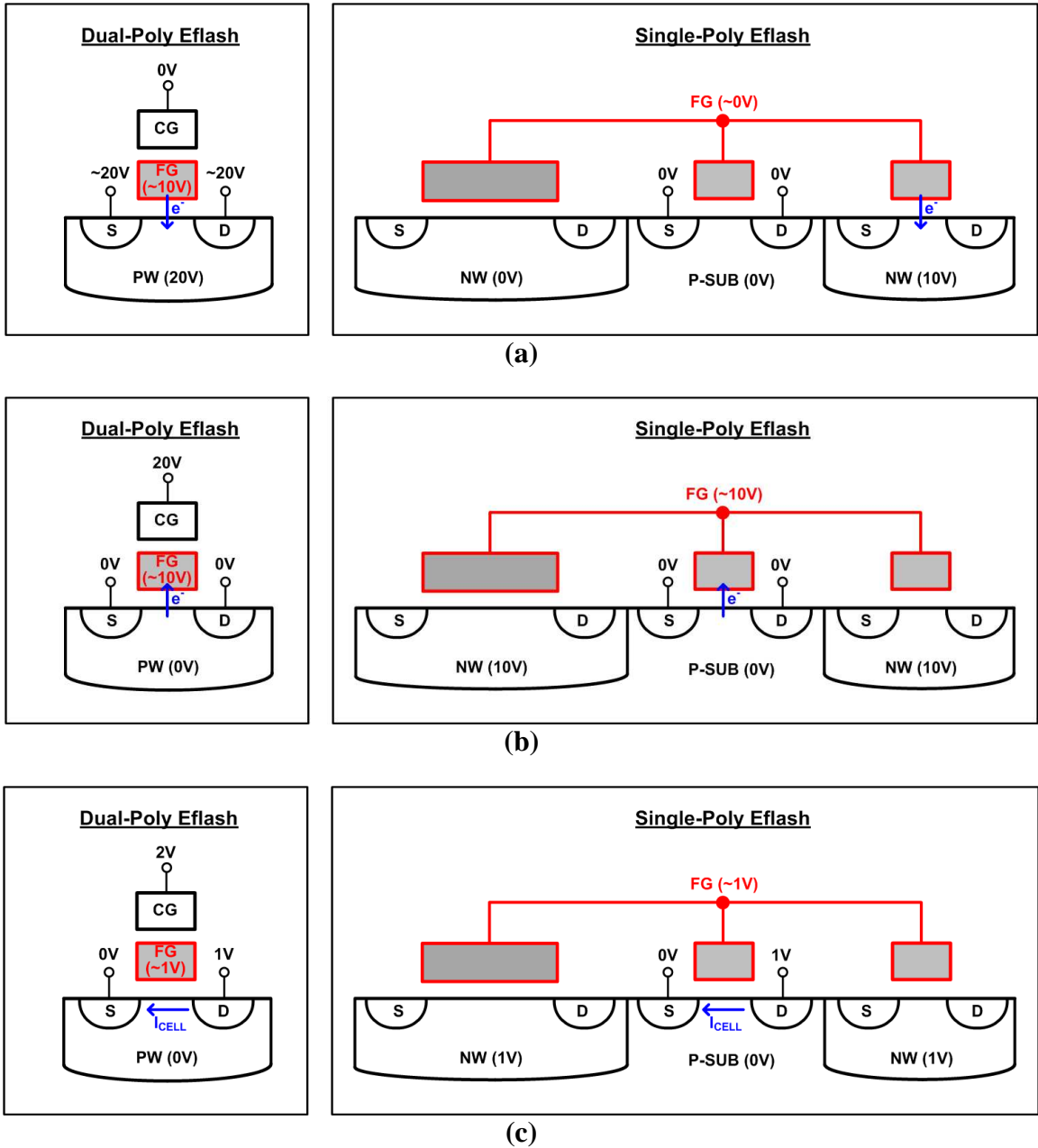


Fig. 1.4 Comparison of the dual-poly and single-poly eflash bias conditions for (a) erase, (b) program, and (c) read operations.

The various design challenges of the single-poly eflash memory, however, need to be addressed as illustrated in Fig. 1.5, too. Firstly, the unselected cells can be unintentionally distorted due to the high voltage (HV1-3) operation. When a high voltage (HV3) is applied to the selected WL for the erase or program operation, another high voltages (HV1 and HV2) need to be applied to the unselected WL and the unselected BL in order to minimize the high voltage stress driven to the unselected cells; however, since the single-poly eflash memory cell is implemented using a standard I/O device of which the supply voltage is much smaller than these high voltage levels (HV1-3), the unselected cells are prone to be affected by those high voltage levels driven for the selected cell operations, which is called high voltage disturbance. Secondly, it's difficult to implement the logic compatible High Voltage Switch (HVS) without a reliability concern. As mentioned earlier, the selected cell needs to be driven to such a high voltage as 10V during erase or program operation, whereas it may need to be discharged to as low voltage as 0V during idle mode or read operation. Thus, the eflash memory needs to have the HVS circuit that outputs such a high or low voltage level selectively depending on the input I/O supply signal; however, this is not a simple task, since an HVS circuit has to be implemented using the standard logic devices (ex. 2.5V I/O device), too. Thirdly, the cell sensing margin after long retention time is limited, since the single-poly eflash cell implemented using standard I/O devices typically has thinner tunnel oxide than that of the dual-poly eflash. The typical tunnel oxide thickness (T_{OX}) of the standard I/O device is around 5nm for 2.5V device, which is not believed to be thick enough to meet the typical eflash retention target (ex. 10 year after programmed) due to the electron leakage

out of or into FG changing the number of electrons stored in FG and the cell threshold voltage (V_{TH}). This cell V_{TH} instability is further accelerated by defects created after large number of program and erase cycles [47-53], limiting the available erase and program cycle counts of the single-poly eflash built using 2.5V I/O devices within 1k cycles.

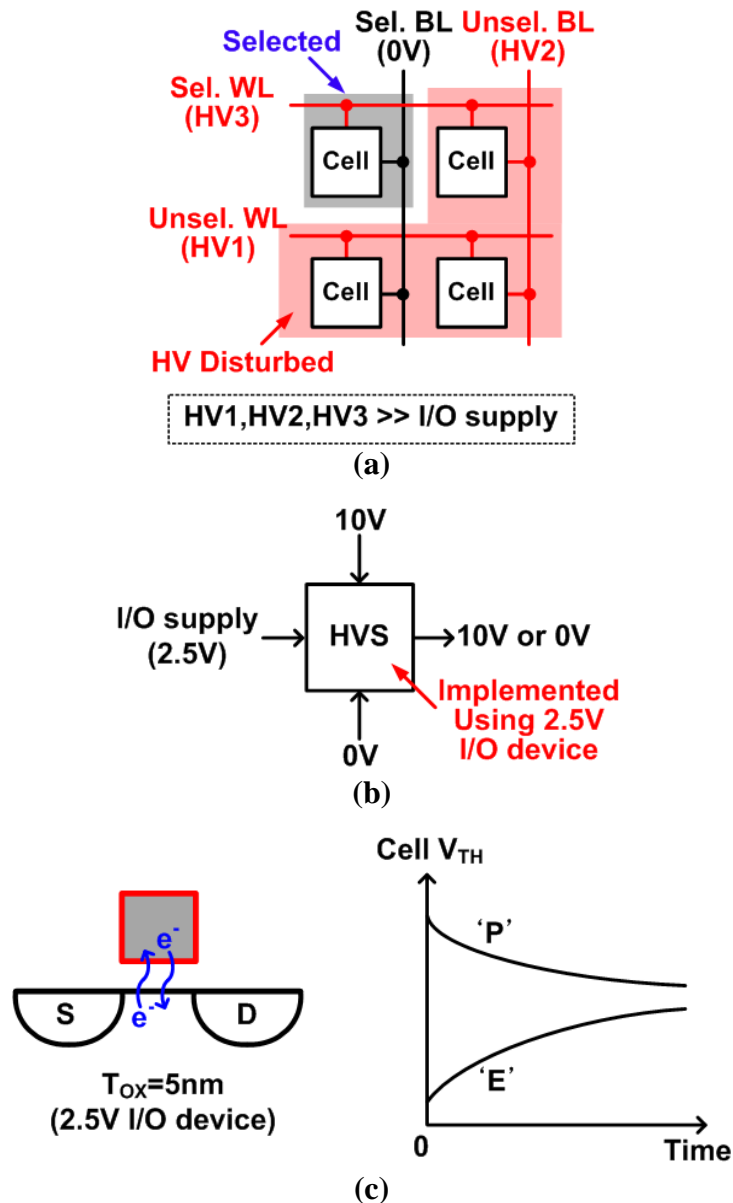


Fig. 1.5 Challenges of the single-poly eflash memory: (a) high voltage disturbance, (b) implementation of logic compatible high voltage switch, and (c) limited cell sensing margin after long retention time.

1.3 Summary of Thesis Contributions

The remainder of this work will explore the benefits of three eflash designs proposed to alleviate the aforementioned design challenges of the single-poly eflash memory for moderate density eNVM applications [54-57]. The 5T eflash memory proposed in chapter 2 features the WL-by-WL accessible architecture with negligible high voltage disturbance, the overstress-free multi-story HVS expanding the cell sensing margin, and a selective WL refresh scheme for the cell endurance to more than 10k P/E cycles [54, 55]. The various types of eflash structures are studied in chapter 3 to find the most favorable eflash configuration when the performance, endurance, retention, and disturbance characteristics are all considered [56]. The 6T eflash memory proposed in chapter 4 features the bit-by-bit re-write capability without disturbing the unselected WL cells for improving the overall eflash endurance [57]. The negative HVS and charge pump circuits implemented in a generic logic process are also illustrated. The 10T eflash memory proposed in chapter 5 features the multi-configurable HVS not requiring the boosted supplies during read operation, and the differential cell architecture improving the eflash cell retention time. Then, this work is concluded in chapter 6.

Chapter 2 A Logic-Compatible 5T Embedded Flash Featuring a Multi-Story High Voltage Switch and a Selective WL Refresh Scheme

There has been numerous device and circuit level research on high-density non-volatile memories such as dual-poly or split-gate eflash, STT-MRAM, PRAM, and RRAM. However, only few attempts have been made to develop a cost effective moderate-density non-volatile solution using standard I/O devices. In this chapter, a logic-compatible eflash memory that uses no special devices other than standard core and I/O transistors is demonstrated in a generic logic process having a 5nm tunnel oxide. An overstress-free high voltage switch and a selective WL refresh scheme are employed for improved cell threshold voltage window and higher endurance cycles.

2.1 Overview of the Embedded Flash Memory Candidates

We give a brief overview of high-density and moderate-density eflash memory candidates in this section prior to discussing the proposed eflash in later sections. Dual-poly and split-gate eflash memory cells for high density eNVM applications are illustrated in Fig. 2.1. A 1T dual-poly eflash utilizes Channel Hot Electron (CHE) injection method for program operation dissipating a high program power [4], whereas 2T and 3T dual-poly eflash memories utilizes Fowler-Nordheim (FN) tunneling method for program operation reducing program power dissipation with the increased program voltage for an efficient FN tunneling due to the low coupling ratio between CG and FG [6, 7]. On the other hand, a charge trap based eflash technology is demonstrated in various literatures, as it can reduce the additional mask count beyond logic [8, 9]. Later, the Split-Gate (SG) eflash reducing the write voltage level as well as enhancing the electron injection efficiency was proposed [10, 11], but it utilizes the Source Side Injection (SSI) method still requiring considerable current for an efficient program operation. Recently, SG-MONOS (Metal-SiO₂-SiN_x-SiO₂-Si) eflash technology was reported to have a higher reliability and better scalability due to its defect-resistance nature [12, 13]. All these candidates incur additional process steps beyond logic technology for forming the FG or Nano-Crystal (NC) and the dedicate thick tunnel oxide. Moreover, achieving a high coupling ratio between CG and FG (or NC) for effective program and erase operation involves process optimization well beyond what is needed for developing a standard

CMOS logic process; therefore, the program and erase voltage levels typically becomes greater than 10V, which requires significant modification to the standard logic process technology to build such high voltage circuits additionally. Thus, due to all these process overhead beyond logic technologies, the dual-poly or split gate eflash memories are not attractive for the cost-effective moderate density eNVM applications.

Eflash	1T Dual-Poly [4]	2T Dual-Poly [6]	3T Dual-Poly [7]	Split Gate [10]
Unit Cell Schematic				
Process	90nm Eflash	90nm Eflash	0.4μm Eflash	90nm Eflash
Process Overhead	Floating Gate	Floating Gate	Floating Gate	Split Gate
Tunnel Oxide	10nm (Dedicated)	7nm (Dedicated)	N. A.	N. A.
Program Method	CHE Injection	FN Tunneling	FN Tunneling	SS Injection
Erase Method	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Write Voltage	±10V	16V	22V	14V
Write Power	High	Low~Medium	Low	Medium
Erase Disturb	None	None	None	None
Measured Retention	>1k hours @ 250°C, EP 1k	>48 hours @ 250°C	N. A.	>1k hours @ 150°C, EP 10k
Cell Size	0.44μm ² (54F ²)	N. A.	4.36μm ² (27F ²)	N. A.

Fig. 2.1 Dual-poly and split gate eflash memory cells attractive for high density eNVM applications. Due to the process overhead to build floating gate or split gate structures, they are not attractive for the moderate density eNVM applications.

Single-poly eflash memory on the other hand does not have any process overhead compared to a generic logic process while a high coupling ratio can be easily obtained by upsizing the width of the coupling device as shown in Fig. 1.3. This feature helps reduce the required program and erase voltage levels resulting in a simpler high voltage circuitry. Hence, single-poly eflash is a promising candidate for moderate density (e.g. few kilobits) non-volatile storage in cases where a dedicated eflash process is not available [26-40]. Various single-poly eflash memory cells suitable for moderate density eNVM applications are illustrated in Fig. 2.2.

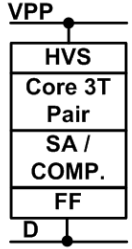
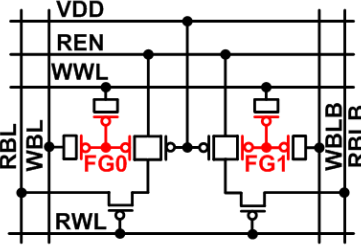
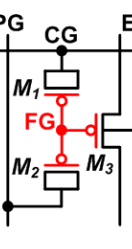
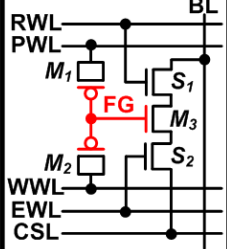
Eflash	Single-Poly [29]	10T Single-Poly [33, 34]	3T Single-Poly [38]	This Work [54, 55]
Unit Cell Schematic				
Process	0.13 μ m Logic	0.25 μ m/0.18 μ m/90nm/65nm Logic	65nm Logic	65nm Logic
Process Overhead	None	None	None	None
Tunnel Oxide	7nm (St. 3.3V I/O)	~5nm (Standard 2.5V I/O)	~5nm (St. 2.5V I/O)	~5nm (St. 2.5V I/O)
Program Method	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Erase Method	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Write Voltage	8V	8V	8V	5 to 10V
Write Power	Low	Low	Low	Low
Erase Disturb	None	(Write Disturb) Unselected WL's	Unsel. WL's	None
Measured Retention	>1k hours @ 150°C, EP 1k	>6500 hours @ 85°C, EP 100k	N. A.	>486 hours @ 27°C, EP 10k
Cell Size	700 μ m ² (est.) (41000F ²)	N. A. *Similar 10T Cell [36]: ~220 μ m ²	N. A.	8.62 μ m ² (2111F ²)

Fig. 2.2 Single-poly eflash memory cells attractive for the moderate density eNVM applications owing to their logic compatibility and lower writing voltage level compared to dual-poly or split-gate eflash.

Previously reported single-poly eflash memories, however, have write disturbance issues in the unselected WL's, as a write voltage greater than $2\times$ the nominal voltage has to be applied in both WL and BL directions. Furthermore, most of them temporarily overstress the High Voltage Switch (HVS) circuits which can result in oxide reliability issues. A dual cell architecture was reported in [29-36] to enhance the cell sensing margin at the expense of larger cell area, while several single cell architecture were proposed for a compact cell area [26-28, 37-40]. In this chapter, we present a 5T single-poly eflash memory that uses no special devices other than standard core and I/O transistors readily available in a standard logic CMOS technology based on a single cell architecture [54, 55]. The proposed row-by-row accessible array architecture alleviates the write disturbance issue in the unselected WL's. To achieve high reliability and good retention characteristics, the proposed 5T eflash memory employs an overstress-free multi-story HVS capable of expanding the cell V_{TH} window. A selective WL refresh scheme is also developed, which improves the overall cell endurance limit.

The remainder of this chapter is organized as follows. Section 2.2 describes the proposed single-poly 5T eflash architecture and cell operation, and the proposed high voltage switch and selective refresh scheme. Measurement results from test chips fabricated in a 65nm low power CMOS process are presented in section 2.3. Section 2.4 compares the proposed single-poly 5T eflash to the prior single-poly eflash, and a chapter summary is given in section 2.5.

2.2 Proposed Logic-Compatible 5T Eflash

2.2.1 Overall Eflash Memory Architecture

The array architecture and unit cell layout of the proposed logic-compatible 5T eflash are shown in Fig. 2.3. All five transistors (M_1 , M_2 , M_3 , S_1 , S_2) in the unit cell are implemented using standard 2.5V I/O transistors with a tunnel oxide thickness (T_{OX}) of 5nm. Here, M_1 is the coupling device, M_2 is the erase device, M_3 is the program/read device, and S_1 and S_2 are the selection devices for the program inhibition operation using self-boosting [58, 59]. The gate terminals of the three devices M_1 - M_3 are connected in a back-to-back fashion forming the FG node. The width of M_1 is made 8 times wider than that of both M_2 and M_3 achieving a high enough coupling ratio for effective erase and program operation. PMOS transistors biased in a non-depletion mode were utilized for M_1 and M_2 in order to achieve a high programming speed. The n-wells (i.e. PWL and WWL) are shared in the WL direction attaining a tight BL pitch. The column peripheral circuits such as the sense amplifier and BL driver are implemented using standard 1.2V thin T_{OX} core transistors while the high voltage switch in the WL driver is implemented using standard 2.5V I/O transistors. Details of the proposed 5T eflash cell operation and the multi-story high voltage WL driver circuits are given in the following sections.

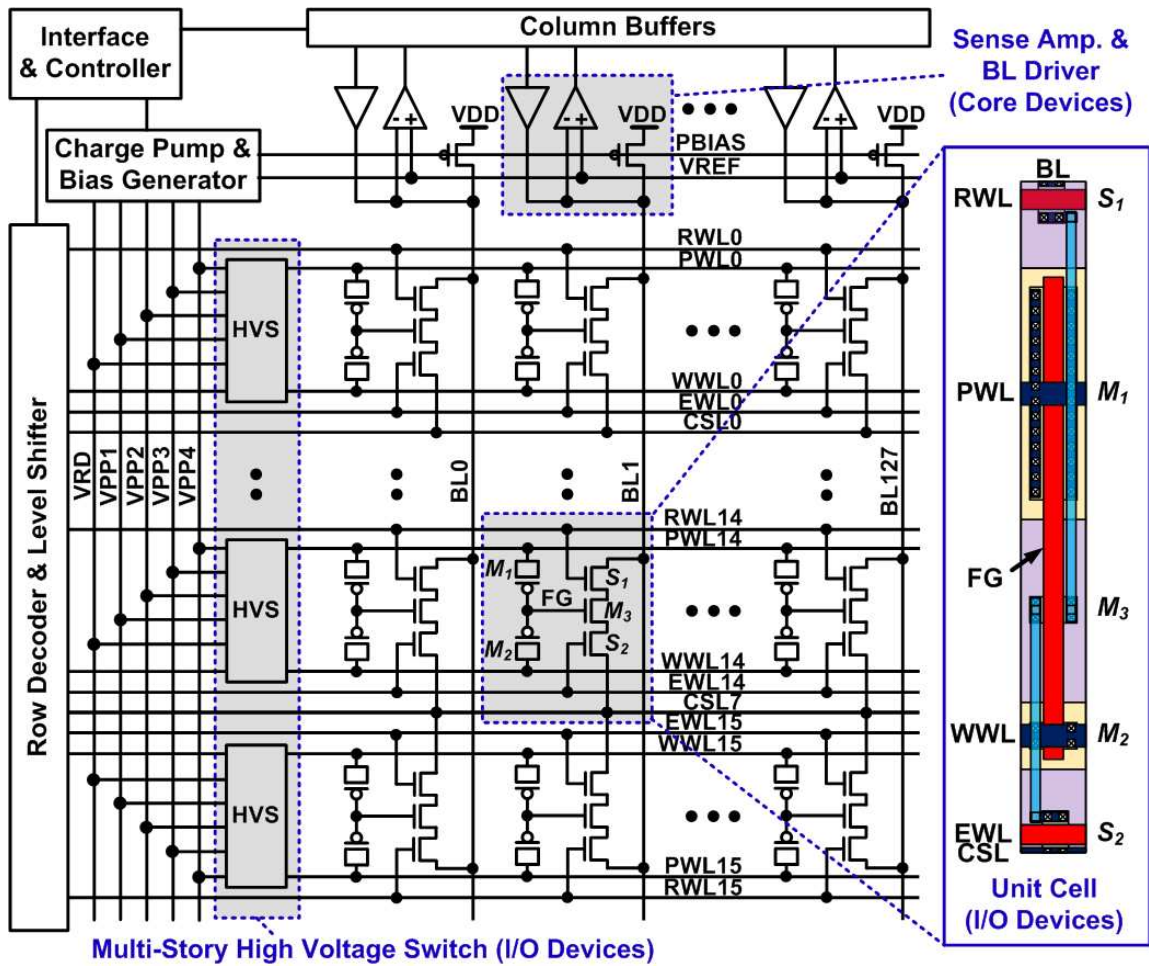


Fig. 2.3 Array architecture and unit cell layout of the proposed logic-compatible 5T eflash memory. Only standard I/O and core devices are used.

2.2.2 Proposed 5T Eflash Cell Operation

Cell bias conditions for erase and program operation of the proposed eflash cell are shown in Fig. 2.4. During erase operation, a high voltage pulse is applied to the selected Write-Word-Line (WWL) while Program-Word-Line (PWL) is biased at 0V. The large gate capacitance of the upsized M_1 generates a high electric field in the gate oxide of M_2 removing electrons from FG through FN tunneling. The coupling ratio of WWL to FG during erase operation is calculated as 0.13 regardless of BL voltage levels, as the upper

select transistor (S_1 in Fig. 2.3) is turned off. All the cells in the selected WL are erased simultaneously. Unlike the dual poly cell in [4], the proposed cell structure can support a single WL erase operation without requiring a negative boosted voltage and a complicated WL driver circuit. The n-well to substrate junction breakdown voltage of the process used in this work was measured to be greater than 13V so it can reliably support a 10V erase operation. During program operation, a high voltage pulse is applied to both the PWL and WWL of the selected WL, while self-boosting of the localized electron channel of the program/read device (M_3 in Fig. 2.3) prevents the cells of the unselected BL's from being programmed by turning off the two select transistors (S_1 and S_2 in Fig. 2.3) in the unselected BL's [58, 59]. The coupling ratio of PWL and WWL to FG during program operation is approximately 0.9. A WL-by-WL erase and program operation ensures that unselected WL's are protected from the high erase and program voltage levels while reducing the power consumption compared to prior single-poly eflash [30-40]. A separate erase device (M_2) in conjunction with the self-boosting technique allows the column peripheral circuits to be built using low voltage core devices without the need for high voltage protection circuits. This reduces the power consumption and improves read access time. The bias condition and simulated timing diagram for read operation are shown in Fig. 2.5. The extracted WL/BL parasitic capacitances and resistances are included in this Monte Carlo simulation. For the read operation, all BL's are pre-charged to the core supply voltage (i.e. 1.2V), while the selected PWL and WWL are pulled-up to the read reference level, VRD (i.e. 0.8V in this example). Once the pass transistors (S_1 , S_2) are activated, the BL voltage levels start to discharge at different rates depending on

whether it is a programmed or erased cell. Thereafter, the BL levels are compared to a reference voltage, V_{REF} , using voltage sense amplifiers to produce digital output signals SO_A and SO_B . V_{REF} of 0.8V is chosen to account for the variation in the FG node voltage of the erased cell (i.e. cell B) and for a better timing margin for SAEN signal. Sense amplifiers are located in each column, which enables parallel read operation to enhance data throughput during both normal and refresh operation (more details are given in Section 2.2.4). Note that the WL/BL lengths of the 2kb eflash array in this work are $120\mu\text{m}$ and $200\mu\text{m}$, respectively (Fig. 2.19), making the WL/BL parasitic elements small enough to achieve a 10ns read access time. For high density eflash memories having larger parasitic elements, however, a more sophisticated sensing scheme may have to be deployed to achieve such a fast read access time [60].

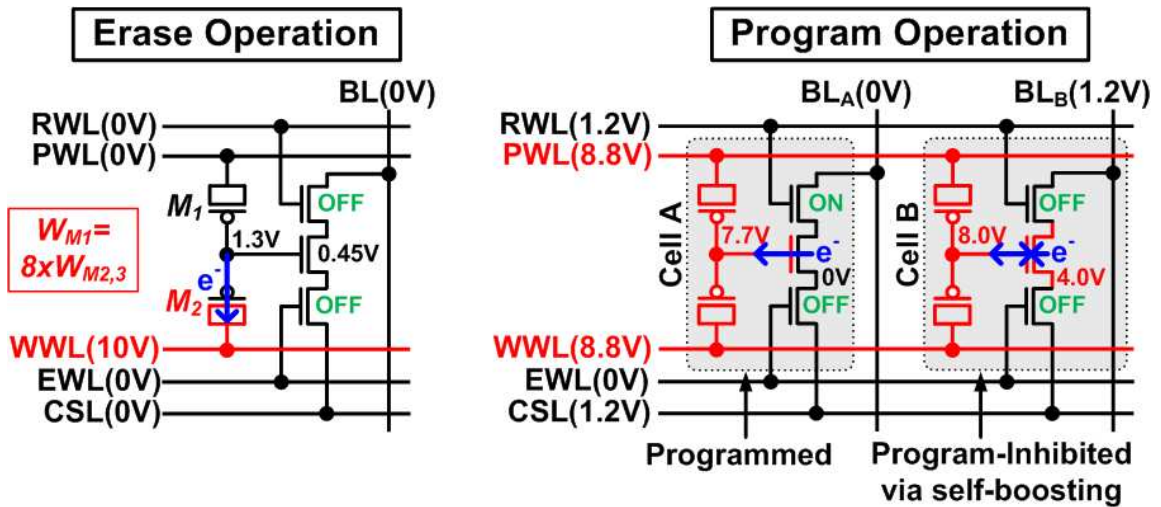


Fig. 2.4 Bias conditions for erase and program operations of the proposed 5T eflash cell. A single WL write operation ensures that unselected WL's are protected from the high voltage levels.

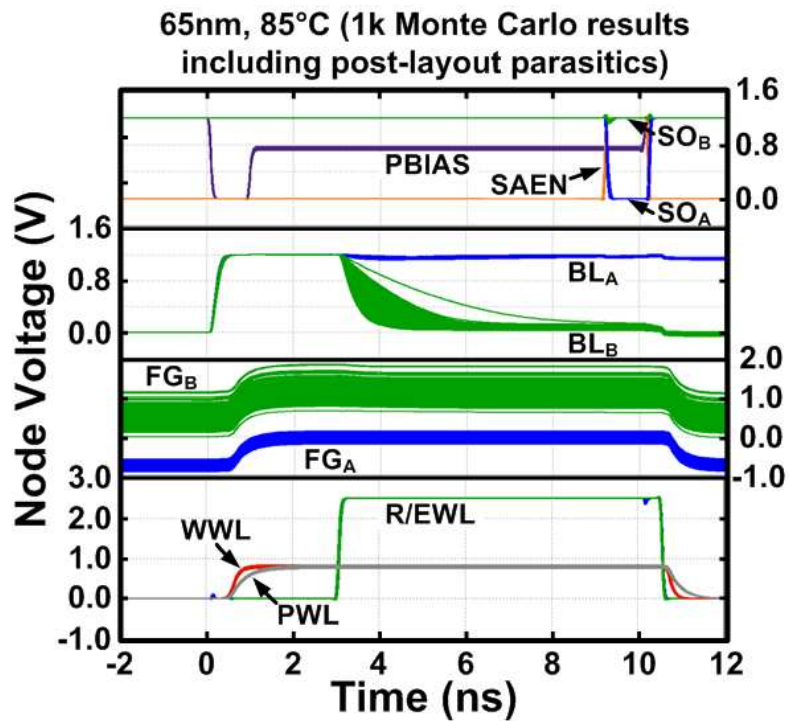
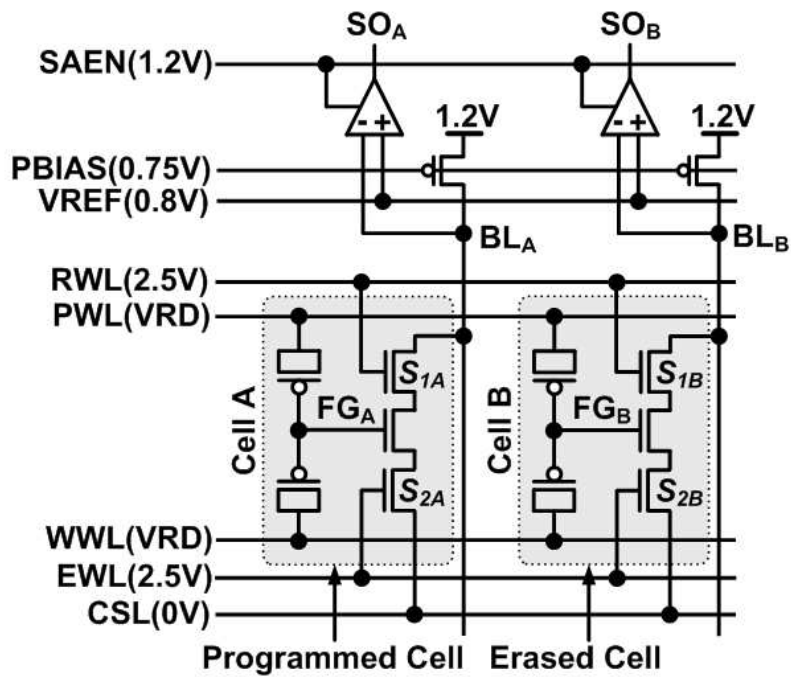


Fig. 2.5 (top) Bias condition and (bottom) simulated timing diagram during read operation. Waveforms are from 1k Monte Carlo runs using a post-layout extracted netlist.

2.2.3 Proposed Multi-Story High-Voltage Switch

Fig. 2.6 compares the prior HVS [29] to the proposed one. In the prior work, the maximum allowable program and erase voltages were limited to slightly higher than 2 times the nominal I/O voltage due to gate oxide reliability concerns. Thus, the prior HVS has a limited output voltage of 8V even with a 0.7V overstress voltage when 3.3V I/O devices are used. Another key issue was that the internal node voltage in the PMOS cascode is sensitive to the threshold voltage drop of the PMOS device, making the circuit susceptible to variation effects and limiting the output voltage range. The proposed HVS on the other hand has a maximum allowable program and erase voltages that are up to 4 times the nominal I/O voltage without any overstress voltage while providing robust output voltage (VOUT) levels by utilizing multi-story stacked latches and 4× I/O VDD driver with additional VPP supplies. The four boosted supply levels (VPP1-VPP4) can be generated from an on-chip charge pump [29, 61-63] with the highest voltage VPP4 being 3 to 4 times the nominal I/O voltage depending on the operating mode. All transistors in the multi-story HVS operate within the nominal voltage tolerance limit. Specialized Drain-Extended MOS (DE-MOS) devices were utilized in [29] to withstand the program and erase voltage levels of 8V, avoiding the junction breakdown limit. Instead, deep n-well layers are used sparingly in the proposed HVS design to minimize area overhead while keeping the drain to body voltages of all transistors to be less than 5V which is roughly half the junction breakdown voltage. When the input signal (VIN) switches, a level shifted selection signal (SEL) and an internally generated pulse activate the pull-down path for a short period which in turn changes the states of the 3 stacked latches. The

signal pulse width is kept short to minimize the static power consumption and current loading of the VPP levels, while the pull-down NMOS stacks in the latch stage are properly sized so that the latch states change within the short on-period of the signal pulse. NMOS transistors in the 3 stacked latches are up-sized to minimize the voltage undershoot that could cause oxide reliability issues.

Further details of the proposed multi-story HVS are provided in Figs. 2.7-2.8. During program operation, PWL/WWL pulses are applied to the selected row consisting of 128 cells (Fig. 2.7 (top)). Simulated current and voltage waveforms of the four boosted supplies are shown when PWL/WWL signal levels switch at $t=3\mu\text{s}$ and $t=5\mu\text{s}$ in Fig. 2.7 (middle, bottom). Parasitic capacitances were extracted from the array layout for accuracy. When SEL switches from low to high for a program operation (Fig. 2.7 (middle)), nodes A, B, D, and F are discharged to VPP3, VPP2, VPP1, and 0V, respectively, while node C and E are pulled up to VPP3 and VPP2. As a result, node M is connected to VPP3 and WWL is connected to VPP4 through the stacked PMOS transistors. The simulated waveform shows that the switching time is typically below 6ns which is significantly shorter than the usual program pulse width ($\sim 10\mu\text{s}$). When SEL switches from high to low (Fig. 2.7 (bottom)), the opposite transition occurs wherein WWL switches to 0V through the stacked NMOS transistors. When SRD is activated for read operation, the low-to-high transition of WWL occurs as illustrated in Fig. 2.8 where the bottom NMOS in the output stage turns on, making WWL switch to the read voltage VRD. The NMOS I/O transistor stack in the driver stage is properly up-sized to reduce the rise time of PWL and WWL for fast read operation that is simulated in Fig. 2.5.

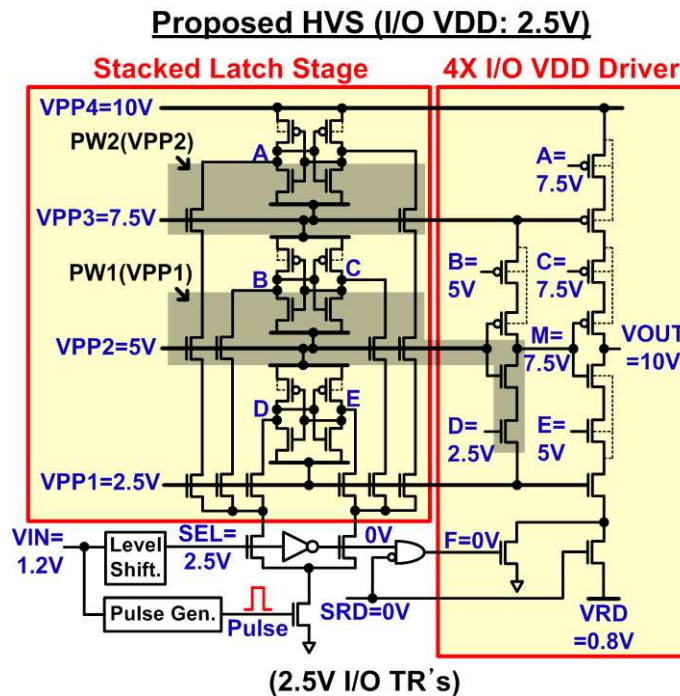
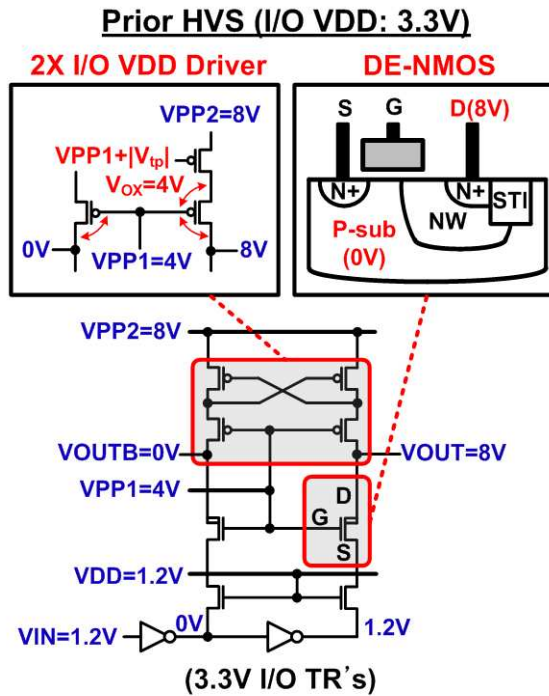


Fig. 2.6 (top) The prior HVS having a limited output range with reliability and variability concerns is compared to (bottom) the proposed HVS increasing VOUT up to 10V and providing robust internal voltage levels by utilizing multi-story stacked latches and 4× I/O VDD driver with additional VPP supplies.

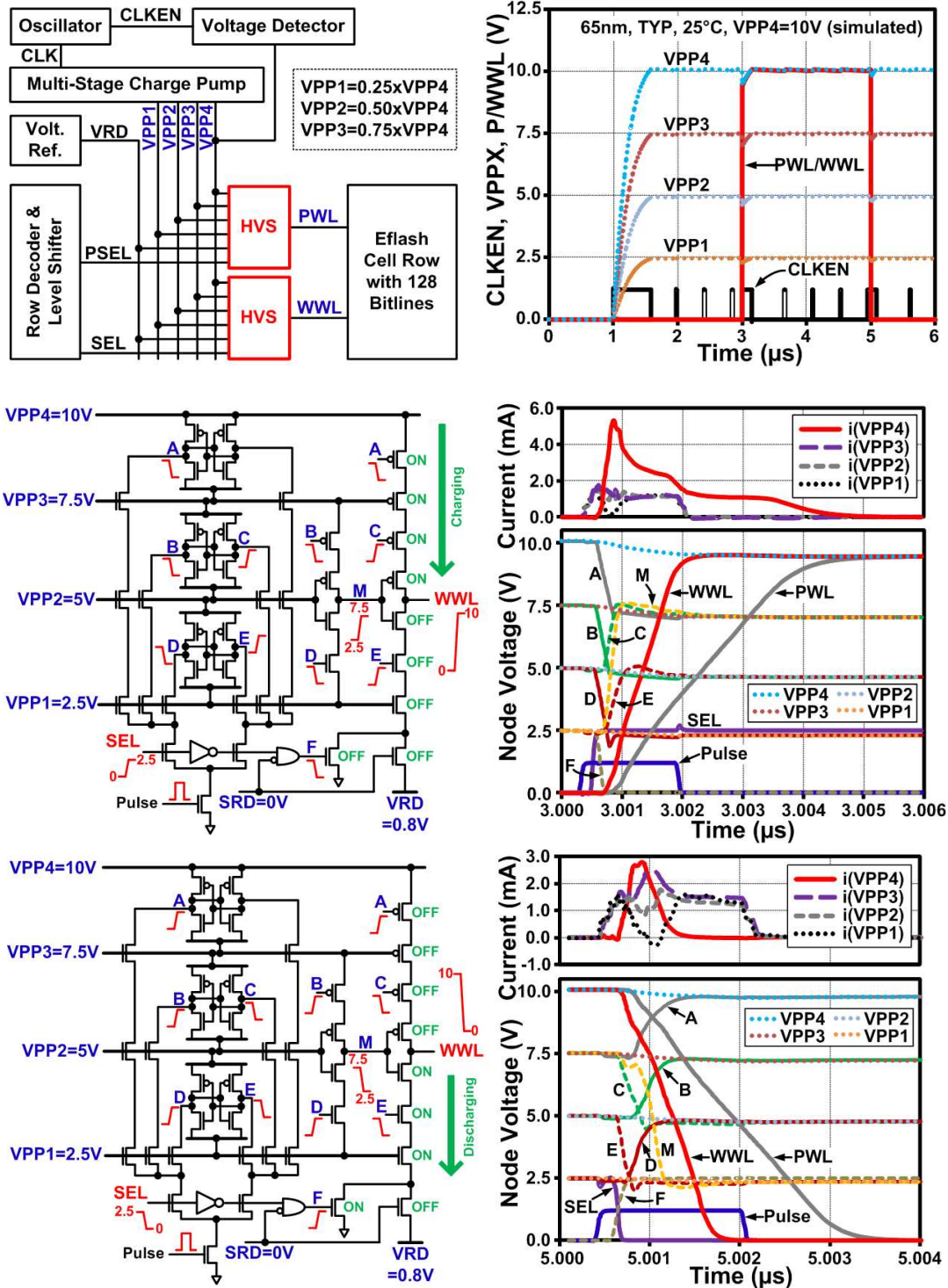


Fig. 2.7 Simulated voltage and current waveforms with the proposed HVS and a voltage doubler based on-chip charge pump [29, 61-63]. Low-to-high and high-to-low transitions of WWL during program operation are shown.

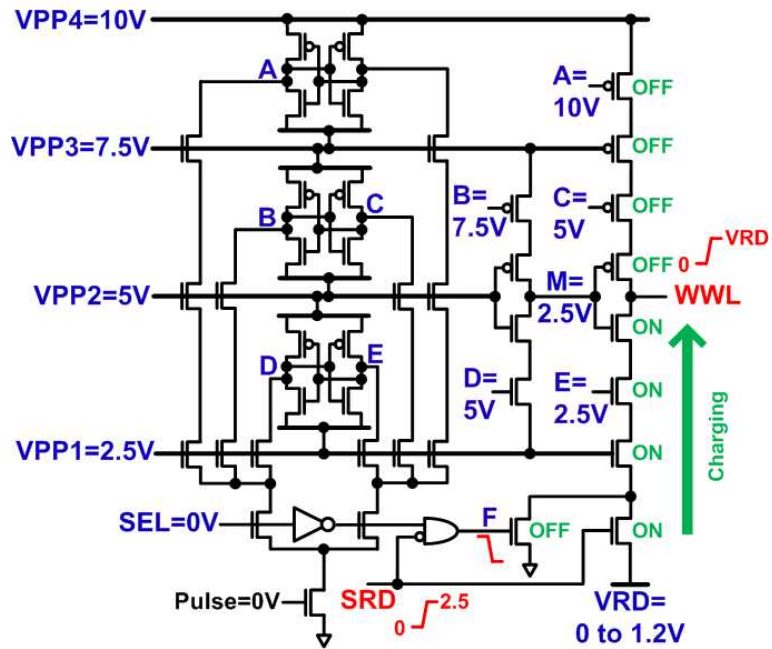


Fig. 2.8 Low-to-high transition of WWL for read operation.

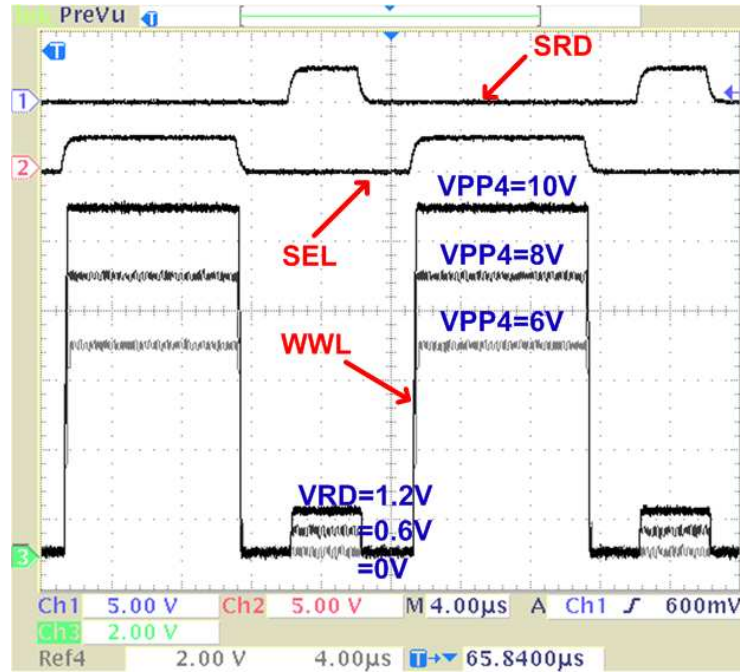


Fig. 2.9 Measured waveforms of the proposed HVS for three different read and write voltage levels with off-chip supplies ($VRD=0/0.6/1.2V$, $VPP4=6/8/10V$, $VPP3=0.75 \times VPP4$, $VPP2=0.5 \times VPP4$, $VPP1=0.25 \times VPP4$).

Measured waveforms of the proposed HVS for three different read and write voltage levels are shown in Fig. 2.9. Note that a charge pump circuit [29, 61-63] was implemented for obtaining the simulation results in Figs. 2.7-8, however, since we did not include the charge pump design in the test chip, external voltage supplies were used for the actual measurements.

By utilizing a higher erase voltage level (=10V) compared to prior designs (e.g. 8V in [29]), the cell V_{TH} window can be improved by more than 170% using the proposed HVS, as the V_{TH} of the eflash cell can be programmed to higher than 1.6V in 10 μ s or erased to lower than -0.3V in 1ms without resulting in oxide reliability problems in the HVS as verified in Fig. 2.15 (a). The proposed HVS operates reliably for a wide range of read voltages (from 0 to 1.6V) and write voltages (from 5 to 10V).

2.2.4 Selective WL Refresh Scheme

Previous literatures have reported that when a flash cell is programmed and erased repetitively, traps are created inside the tunnel oxide or oxide-silicon interface causing instability in the cell V_{TH} [47-53]. For example, the oxide and interface traps capture electrons during P/E cycling, resulting in positive cell V_{TH} shifts for both erased and programmed cells as illustrated in Fig. 2.10 (left). According to the interface trap annihilation model [48, 52], interface traps created as a result of the released hydrogen atoms during P/E cycling are partially restored during retention mode. As such, de-trapping of the electrons manifests negative cell V_{TH} shifts as illustrated in Fig. 2.10 (center). The oxide and interface traps are believed to facilitate the Trap-Assist-Tunneling

(TAT) phenomena [50-52], which in turn accelerates the charge loss from the silicon substrate and from the FG resulting in positive and negative cell V_{TH} shifts for the erased and programmed cells, respectively.

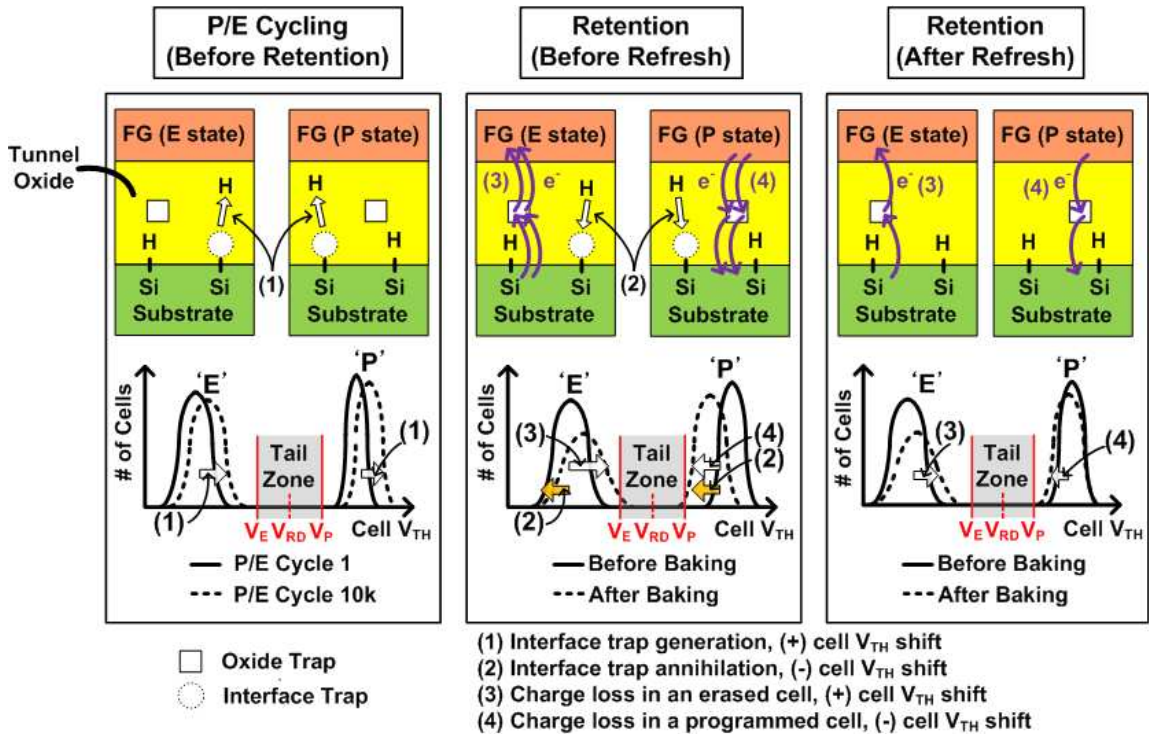


Fig. 2.10 Physical model explaining endurance and retention characteristics considering oxide and interface trap creation, interface trap annihilation, and trap-assisted charge loss [47-53].

Unlike high density flash memories where the charge loss effect can be minimized by optimizing the fabrication process, single-poly eflash memories are built using standard logic devices which are not necessarily optimized for good retention time. To improve the overall cell endurance in single-poly eflash, a refresh scheme is proposed in this work. Similar to Solid-State Drives (SSDs) where retention time can be traded off for

improved endurance and performance [64], an intermediate refresh is conceivable for eflash applications in case they have to support a high number of P/E cycles throughout the entire product lifetime. Since a considerable number of interface traps can be annihilated during retention mode before refresh [48, 52], the trap assisted charge loss becomes smaller after a refresh operation as described in Fig. 2.10 (right). This can enhance the sensing margin and retention time at the expense of additional erase and programming steps for refresh operation. Since the refresh is very infrequent (once a year at most), the impact on the endurance limit is quite negligible. On the other hand, the benefits of refresh (i.e. restored data window, improved post-refresh retention characteristic) are quite significant as demonstrated from our test chip. In fact, the enhanced post-refresh or post-reprogram cell retention characteristics and their potential for maximizing the overall SSD lifetime have been reported by other researchers [64-67].

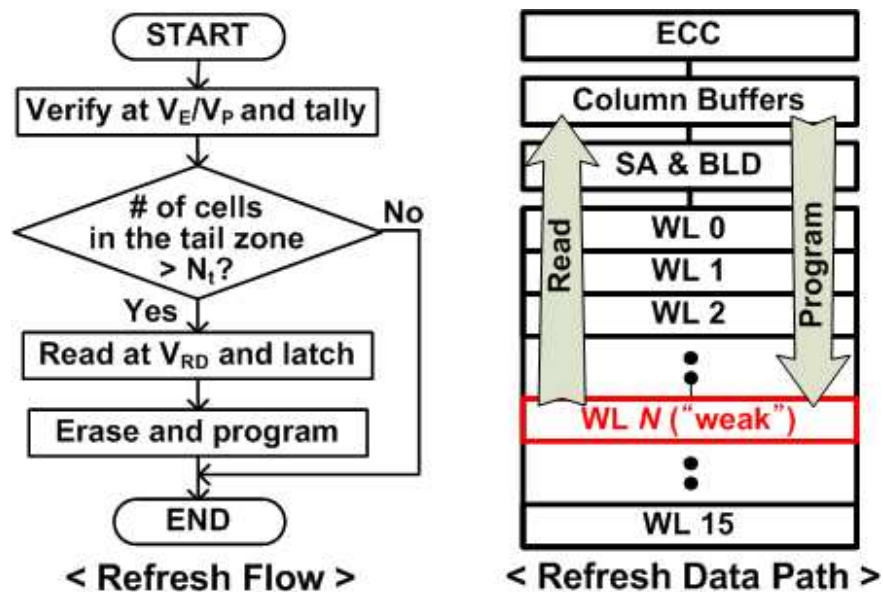


Fig. 2.11 Proposed selective WL refresh scheme. Only identified “Weak” WL’s are refreshed to avoid unnecessary P/E cycles in the good cells.

The proposed selective WL refresh scheme illustrated in Fig. 2.11 identifies “weak” WL’s by keeping track of the number of cells falling into the tail zone (Fig. 2.10). Two verify reference levels (V_E and V_P) are utilized to obtain the number of tail cells. Only the weak WL’s are refreshed, which prevents the “good” WL’s from being unnecessarily cycled. The refresh operation consists of the following two steps; first the original cell data in the weak WL is read and temporarily stored in the column buffer and then a single WL erase and program operation of the original data follows. Alternatively, one can also consider using Error Correction Codes (ECC) to achieve better eflash retention; however, this would require redundant bits in the cell array and will increase the read access time and power consumption. In contrast, a refresh scheme utilizes existing read/program/erase operations so the hardware overhead is low. Furthermore, the refresh is performed infrequently (once a year at most in this work) so the power overhead is also negligible. The main difference with the previous re-program scheme in [68] is that the proposed refresh includes an additional erase step to mitigate retention issues in the erased cells. It’s worth mentioning that the additional erase operation is critical in our design, since the eflash is built in a relatively thin oxide (5nm) logic process. This is in contrast to the previous design built using dedicated thick oxide floating gate devices [68] where the cell V_{TH} shift during retention mode was negligible. For any refresh scheme to work properly, a periodic wake-up of the eflash is necessary to ensure that the number of tail cells does not exceed a certain threshold before the next refresh operation occurs. Therefore, it is important to note that a refresh scheme is only applicable to eNVMs that are part of a system that is able to keep track of time and doesn’t have a case of no power

for longer than the worst case retention time (e.g. 1 year in this work). Test chip measurement results across a wide range of temperatures in Fig. 2.17 confirm that no cells are expected to cross the VRD threshold within 1 year of entering the tail zone. Therefore, an annual wake-up of the chip would be sufficient for the proposed selective refresh scheme to be effective.

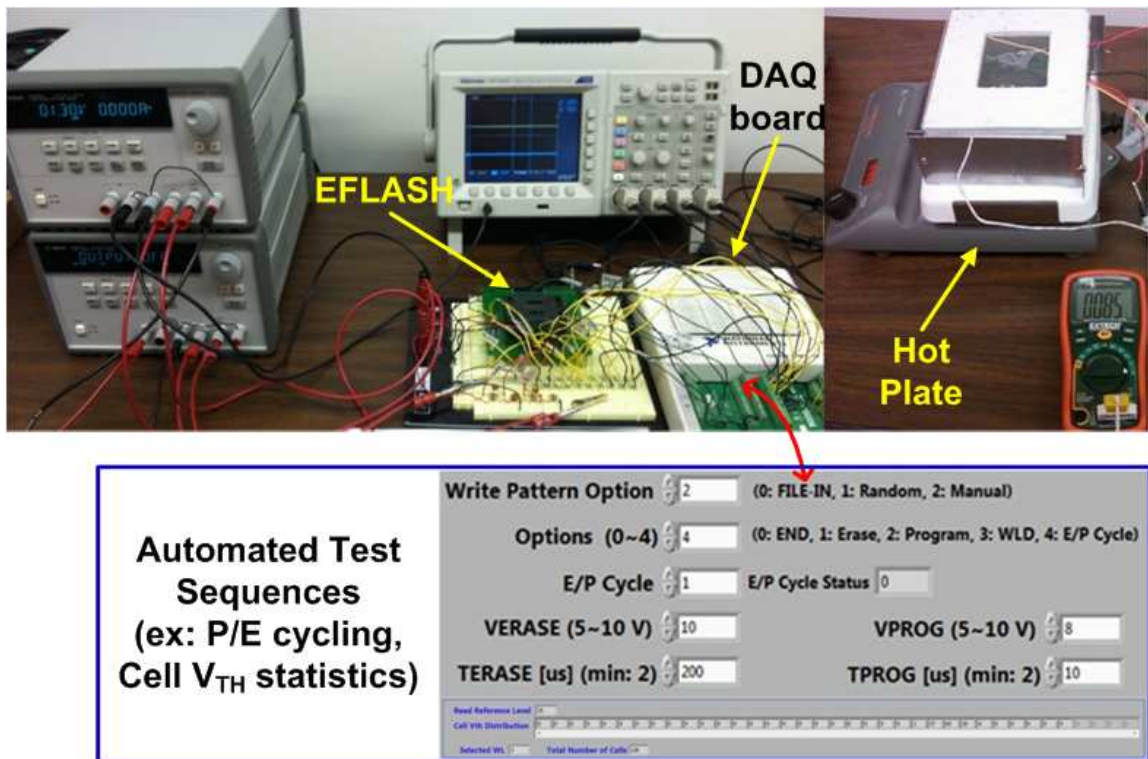


Fig. 2.12 Laboratory setup for test chip measurement.

2.3 Test Chip Measurement Results

2.3.1 Initial Cell V_{TH} and Writing Speed

To demonstrate the proposed circuit techniques, a 2kb eflash memory was implemented in a 65nm low power logic process. Fig. 2.12 shows the laboratory setup for

the test chip measurement. LabVIEW™ controlled data acquisition environment automated the test sequences such as program and erase cycling, and cell V_{TH} statistics calculation. A hot plate (Corning PC-400D) with the thermometer was utilized for baking the test chip up to 150°C to accelerate the retention test.

To measure the cell V_{TH} , we simultaneously swept the PWL and WWL voltage levels while checking whether the sensed data has flipped. Fig. 2.13 shows the initial cell V_{TH} distributions of four test chips which indicate the cell-to-cell and chip-to-chip variations prior to any program or erase operation. The initial cell V_{TH} distribution of four 2kb eflash memory cells has an average value of 0.61V and a 3-sigma value of 0.18V.

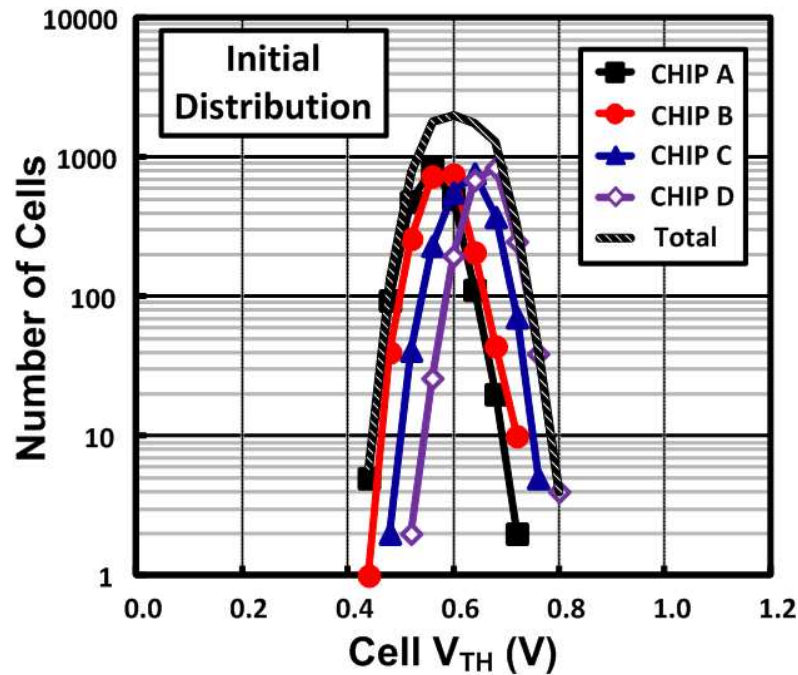


Fig. 2.13 Measured initial cell V_{TH} distributions of four 2kb eflash memory chips show cell-to-cell and chip-to-chip variations. The initial cell distribution ranges from 0.44 to 0.80V with an average value of 0.61V.

The measured erase and program speeds for different voltage levels are shown in Fig. 2-14. VPP1-VPP4 were supplied by an off-chip source to eliminate any non-ideal effects during the eflash memory cell characterization. The average and 3-sigma values of the cell V_{TH} distribution are plotted as a function of the erase and program pulse widths. The erase speed was found to be $\sim 1000\times$ slower than the program speed at similar voltage levels (Erase: 9V, Program: 8.8V). The cell V_{TH} spread increases with erase time and remains almost constant with program time as illustrated by the 3-sigma bars. Note that the program speed of a single cell configuration [26-28, 37-40] is roughly $1000\times$ faster than the write speed of a dual cell configuration [29-36] where the erase operation in one of the cells always limits the write performance according to our measurement results of erase and program speeds.

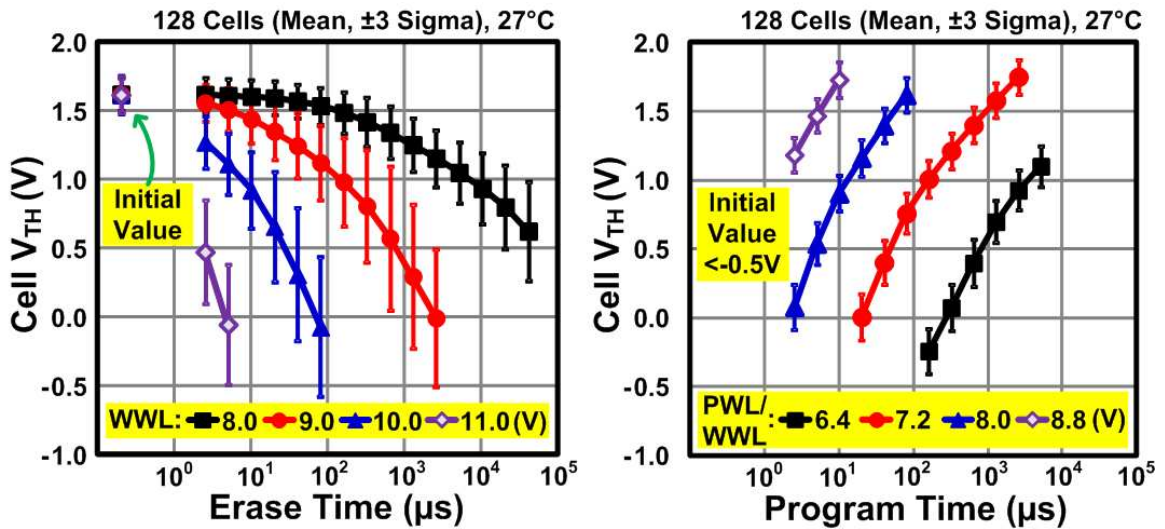


Fig. 2.14 Measured cell V_{TH} for different P/E voltages and pulse durations. Note that WWL and PWL were supplied by an off-chip voltage source to eliminate any non-ideal effects.

2.3.2 Endurance and Retention

Fig. 2.15 shows the measured endurance characteristics of the proposed eflash cells. P/E pre-cycling up to 10k cycles was performed at room temperature (27°C). Fig. 2.15 (a) shows results for an 8.8V/10 μ s program pulse and a 10V/1ms erase pulse, while Fig. 2.15 (b) and (c) are for 10V/100 μ s program and erase pulses. Note that all cells in the array experience the same program and erase stress during the pre-cycling period. For P/E cycles greater than 1k, the programmed cell V_{TH} starts to shift in the positive direction as shown in Fig. 2.15 (a). The erased cell V_{TH} shows a similar positive shift for P/E cycles greater than 1k as shown in Fig. 2.15 (b). The cell current measured from a single cell test structure shows a severe degradation in the sub-threshold slope with increased P/E cycles in Fig. 2.15 (c), implying that a considerable number of interface traps are generated. A linear relationship between the sub-threshold slope and the interface trap density was discussed in [49]. All the graphs show that a cell V_{TH} window greater than 1.9V is achieved for up to 10k P/E cycles.

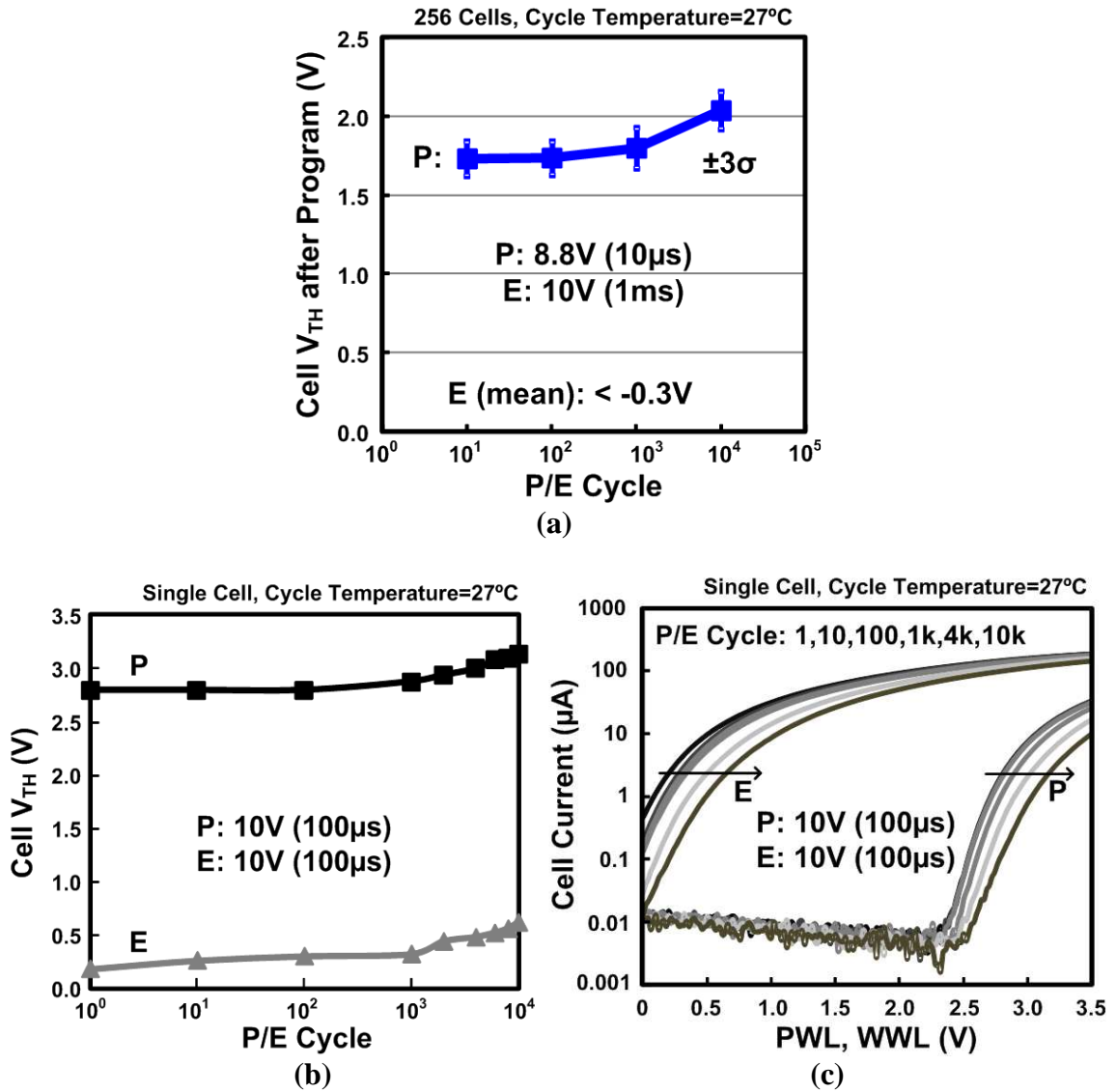


Fig. 2.15 Measured endurance characteristic. (a) Programmed cell V_{TH} shows a positive shift for P/E cycles greater than 1k. (b) Erased cell V_{TH} shows a similar positive shift for P/E cycles greater than 1k. (c) Cell current measured from a single cell exhibits severe sub-threshold slope degradation beyond 1k P/E cycles, implying that a considerable number of interface traps have been generated [49].

Fig. 2.16 shows the measured retention characteristic of the proposed eflash cells. Fig. 2.16 (a) and (b) show that a sufficient sensing margin is maintained at a 150°C bake temperature for the 1k and 10k pre-cycled cells, respectively. Fig. 2.16 (c) shows the cell V_{TH} shifts for the erased (upper) and programmed (lower) cells with P/E pre-cycling counts ranging from 100 to 10k. No apparent spatial correlation is observed within the same WL implying that the tail cells are randomly distributed. Similar data was shown in a prior work where abnormal tail cells during retention mode in a 16M flash memory array did not show any spatial correlation [50]. Fig. 2.16 (d) shows the relationship between the initial cell V_{TH} and cell V_{TH} shift for different P/E pre-cycles. As expected, cells with higher number of P/E pre-cycles typically exhibit a larger V_{TH} shift. The cell V_{TH} shifts for the equally P/E pre-cycled cells however do not show a strong correlation with the initial cell V_{TH} values. Fig. 2.17 shows the evolution of the tail cell V_{TH} for different baking temperatures. For P/E pre-cycling, an 8.8V/10 μ s program pulse and an 8.8V/10ms erase pulse were repetitively applied to three chips at room temperature (27°C). Then, the three programmed chips were baked at three different temperatures: 27°C, 85°C, and 150°C, respectively. The reason behind the sudden decrease in cell V_{TH} for the programmed cell baked at 150°C is because of the negative cell V_{TH} shift caused by the interface trap annihilation and charge loss as previously described in Fig. 2.10. For the erased cells on the other hand, the cell V_{TH} value is relatively constant because the interface trap annihilation and the charge loss has opposite effects on cell V_{TH} as explained in Fig. 2.10 [52].

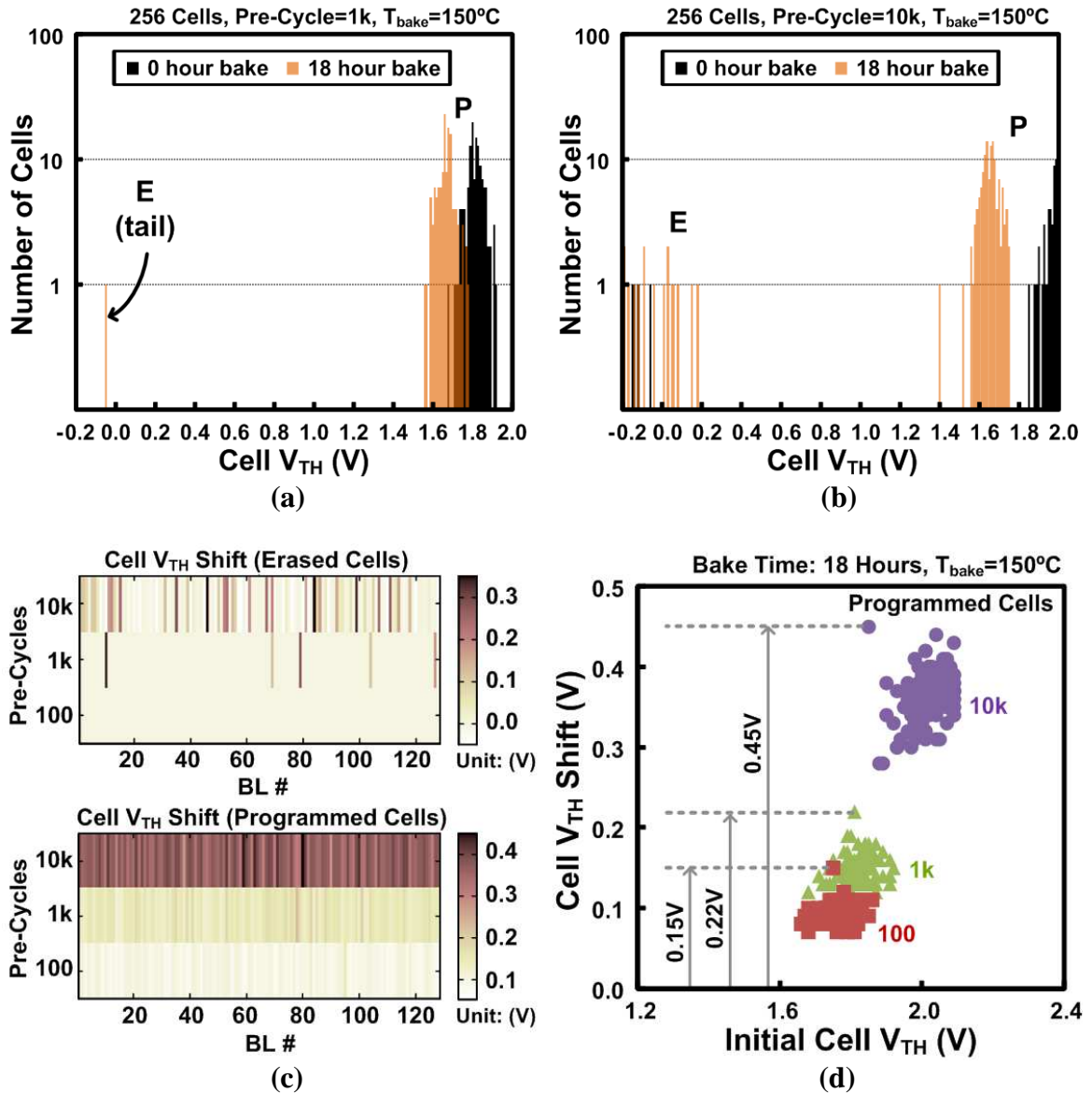


Fig. 2.16 Measured retention characteristic. (a, b) Cell V_{TH} distributions for 1k and 10k P/E pre-cycled cells at a 150°C bake temperature. (c) Spatial bit maps showing the cell V_{TH} shift of erased and programmed cells after 100/1k/10k P/E pre-cycles. (d) Cell V_{TH} shift vs. initial cell V_{TH} level for 100/1k/10k pre-cycles.

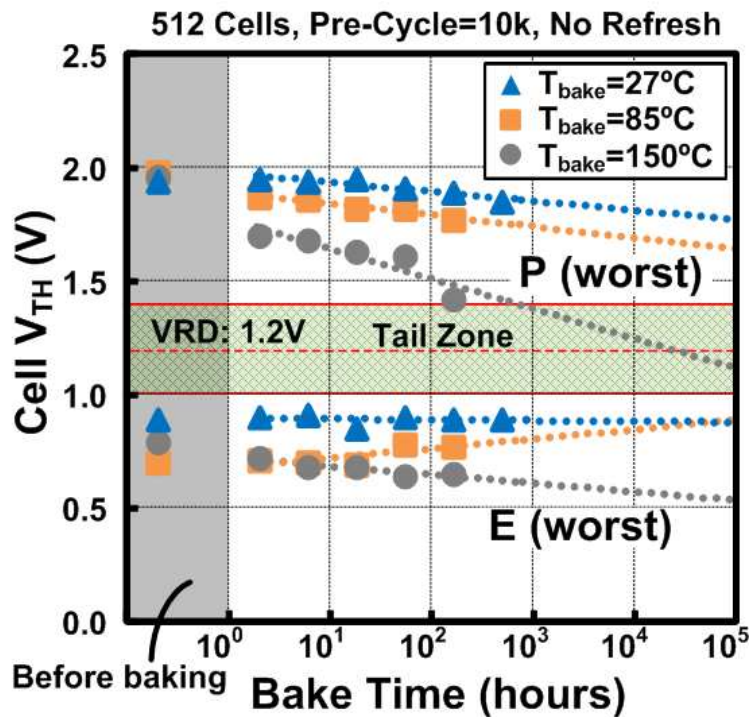
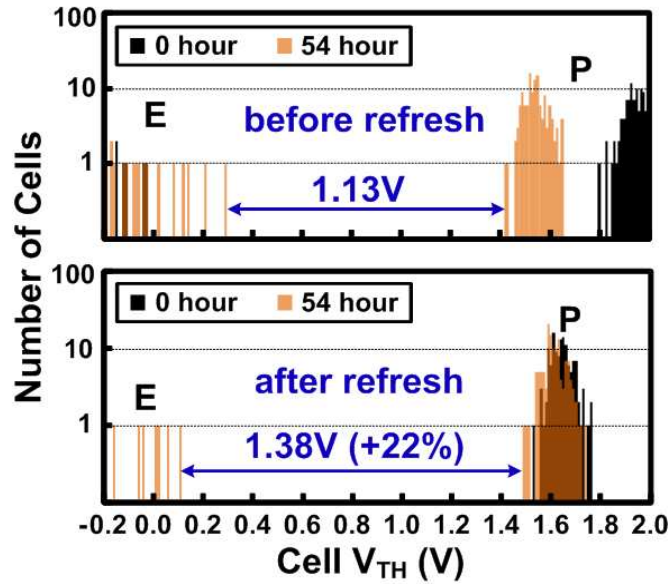


Fig. 2.17 Measured retention characteristic as a function of baking temperature. Three chips were baked at 27/85/150°C, respectively. The effects of charge loss and interface trap annihilation are canceled out for the erased cells, while a negative cell V_{TH} shift is observed in the programmed cells [52].

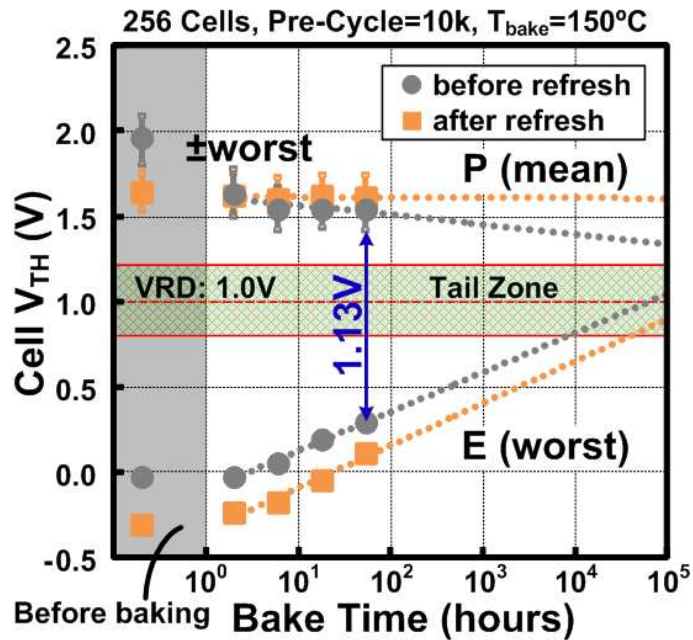
2.3.3 Effectiveness of Refresh Operation

Fig. 2.18 shows the cell retention characteristics of 256 cells with 10k P/E pre-cycles before and after the refresh operation for a baking temperature of 150°C. P/E pre-cycling was performed at room temperature (27°C) using an 8.8V/10 μ s program pulse and a 10V/1ms erase pulse. All cells experience the same program and erase pulses during pre-cycling. A read reference voltage of 1.0V was chosen to maintain a sufficient sensing margin for a wide range of bake times. The post-refresh 54 hour bake results show a 22% higher sensing window (1.13V \rightarrow 1.38V) compared to the pre-refresh 54 hour bake results

(Fig. 2.18 (a)). Projections based on the retention time of 10k pre-cycled cells (Fig. 2.18 (b)) suggests a ~5 times longer retention time which is primarily attributed to the slower cell V_{TH} shift as well as the reinforced erased cell V_{TH} after the refresh operation.



(a)



(b)

Fig. 2.18 Retention characteristics before and after refresh: (a) distribution, (b) sensing margin.

Table 2.1 Single-Poly Eflash Comparison

Single-Poly EFLASH	VLSI 2000 [26]	CICC 2001 [27]	ISSCC 2004 [29]	IEICE 2007 [30]	NVSMW 2008 [39]	This Work [54, 55]
Process	0.25 μ m Logic	0.14 μ m Logic	0.13 μ m Logic	N. A.	0.18 μ m Logic	65nm Logic
Cell Transistor	Thick Oxide TR	3.3V I/O Device	3.3V I/O Device	3.3V I/O Device	3.3V I/O Device	2.5V I/O Device
Tunnel Oxide	10nm	7nm	7nm	7.6nm	7nm	5nm
Writing Voltage	<10V	<6V	<8V	<8.5V	5, -5V	<10V
High Voltage Switch (overstress voltage)	No HVS	3.3V I/O Device (No overstress)	DE-NMOS, 3.3V I/O Device (0.7V)	3.3V I/O Device (1V)	3.3V I/O Device (No overstress)	2.5V I/O Device (No overstress)
Cell V_{TH} Window	3V	3V	0.7V	1.8V	3V	>1.9V
Cell Architecture	Single Cell	Single Cell	Dual Cell	Dual Cell	Single Cell	Single Cell
Erase Time (Unit)	1s (WL)	100ms (WL)	10ms (Block)	4ms (WL)	10ms (WL)	1ms (WL)
Program Time (Unit)	10ms (WL)	3ms (WL)	10ms (Block)	500ms (WL)	1ms (WL)	10 μ s (WL)
Read Time (Unit)	N. A. (WL)	N. A. (WL)	10 μ s (Block)	N. A.	N. A. (WL)	10ns (WL) *Simulated
Erase Method	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Program Method	CHE Injection	CHE Injection	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Cell Current (ON state)	>10 μ A	N. A.	N. A.	N. A.	N. A.	2.19 μ A (ave.)
Unit Cell Size	50 μ m ²	52 μ m ²	700 μ m ² (est.)	500 μ m ²	65 μ m ²	8.62 μ m ²
Capacity	4kb	35b	2kb	5kb	64kb	2kb

2.4 Comparison With Other Single-Poly Eflash

Table 2.1 compares various single-poly eflash memories. McPartland and Shukuri *et al.* presented single-poly eflashes based on a single cell architecture and CHE injection program method, respectively [26-28]. To utilize the higher power efficiency of the FN tunneling program method, Raszka *et al.* utilized 3.3V I/O devices with a 7nm tunnel oxide for the eflash cell [29]. The typical current for simultaneously programming 2kb cells via FN tunneling was reported as around 1 μ A [29]. The HVS in their work uses special DE-NMOS and cascoding stages. However, the cell V_{TH} window was limited to 0.7V even though devices in the HVS experience a voltage overstress. Yamamoto *et al.* also utilized 3.3V I/O devices with a 7.6nm tunnel oxide in the memory cell [30-31]. The cell V_{TH} window in this work was around 1.8V. These two prior eflash memories employ dual cell architectures to boost the sensing margin but this slows down the write speed by

~1000× as explained in Section 2.3.1 (Fig. 2.14). Moreover, each unit cell included a dedicated HVS to resolve the write disturbance issue in the unselected WL's (Fig. 2.2) [29-31]; however this significantly increases the memory footprint compared to other single-poly eflash memories shown in Table 2.1. Later, Roizin *et al.* presented a single-poly eflash with a single cell architecture and FN tunneling programming using 3.3V I/O devices having a 7nm tunnel oxide in a 0.18 μ m logic process [39].

In contrast, the proposed 5T eflash was successfully implemented in a 65nm low power standard CMOS logic process where the I/O devices have a 5nm tunnel oxide. Despite the HVS being overstress free, the proposed multi-story high voltage WL driver achieves a cell V_{TH} window greater than 1.9V as shown in Section 2.3.2 (Fig. 2.15) by allowing the WL voltage to be raised to 3 or 4 times the I/O voltage during erase and program operations. The proposed 5T eflash employs single cell architecture for fast program operation and all 128 cells connected to a single WL are accessed simultaneously improving overall throughput and simplifying the refresh sequence as shown in Section 2.2.4 (Fig. 2.11). Unselected WL's are protected from the high erase and program voltage through the WL-by-WL erase and program architecture. This feature helps improve overall memory endurance by not disturbing the unselected WL cells. By moving the HVS from inside the unit cell to the WL driver block and optimizing the cell layout, the proposed eflash memory achieves a 60 to 80 times smaller cell size compared to prior dual cell single-poly eflash memory implementations [29-31]. Compared to prior single cell implementations [26, 27, 39], our proposed cell size is still 6 to 7 times smaller making it a promising solution for cost-effective moderate-density

nonvolatile on-chip storage. Finally, the die microphotograph of the 65nm eflash test chip discussed in this chapter is shown in Fig. 2.19.

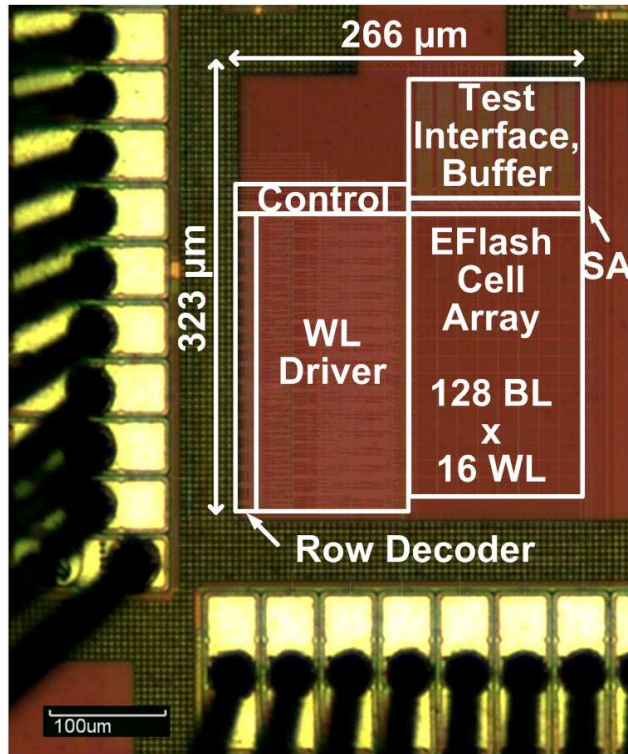


Fig. 2.19 Die photograph of 2kb 5T eflash test chip implemented in 65nm standard logic process.

2.5 Chapter Summary

Although single-poly eflash memories are not suitable for high-density NVM applications due to their large cell size, they can be useful in a wide range of moderate-density NVM applications where a few kilo bits of non-volatile storage need to be built in a generic logic process. These applications include zero standby power systems, adaptive

self-healing techniques, memory repair schemes targeted for time dependent failures, in-field on-line test, and so on. In this chapter, we proposed and experimentally demonstrated a logic-compatible eflash memory in a 65nm logic process targeted for the aforementioned applications. Our test chip features a new 5T eflash cell with negligible program disturbance, an overstress-free multi-story HVS for expanding the cell V_{TH} window, and a selective WL refresh scheme for improving the cell endurance to more than 10k P/E cycles.

Chapter 3 Study on the Optimal Configuration of Single-Poly Embedded Flash Cell

As discussed in prior chapters, a single-poly embedded flash (eflash) memory is a unique type of an embedded Non-Volatile Memory (eNVM) that can be built in a generic logic technology. So far, numerous single-poly eflash cells have been proposed for the cost-effective moderate density eNVM applications. But, the optimal configuration of a single-poly eflash cell has been still questionable. In this chapter, various single-poly embedded flash memory structures combining various standard I/O devices are theoretically analyzed with experimental data from the two eflash test chips fabricated in 65nm generic logic process having 5nm tunnel oxide. Based on our study, the cells with the non-depletion mode coupling devices and the electron ejection device having n-type poly-silicon are preferred for higher program and erase performance, while the cells with a coupling device having p-type poly-silicon is preferred for longer retention time than the cells with a coupling device having n-type poly-silicon. Additionally, the eflash cells having an NMOS read device are measured to support the self-boosting program

operation effectively without disturbance and coupling issues. In conclusion, 5T eflash cell structure combining PMOS coupling device, NCAP electron ejection device, and NMOS read device with two additional pass transistors supporting the self-boosting method is suggested as the most attractive single-poly eflash cell configuration for moderate density eNVM applications where a dedicated eflash process is not available.

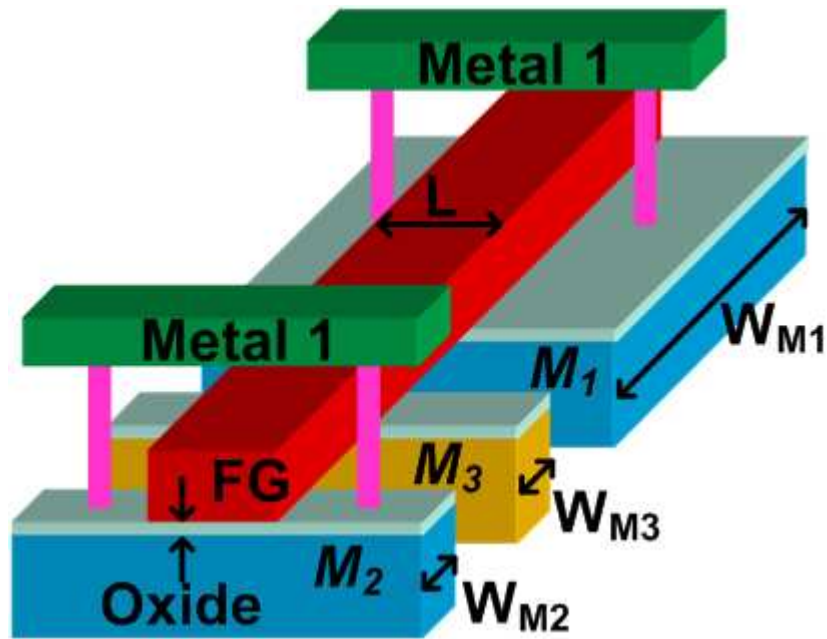
3.1 Single-Poly Embedded Flash Cell Configurations

Single-poly eflash is implemented using standard I/O devices readily available in the standard logic process and therefore has no process overhead beyond logic technology [26-40]. Typically, a single-poly eflash cell has the core structure consisting of the three standard I/O devices (M_1 - M_3) connected back-to-back to form an FG node as illustrated in Fig. 3.1 (a). The FG node is surrounded by the isolation layers including the gate oxide and functions as a non-volatile charge storage node. The coupling device (M_1) is upsized compared to the other ones (M_2 and M_3) for higher coupling from the control gate (i.e. the body of the coupling device, M_1) to the FG node. In a generic logic technology, n-type poly-silicon is combined with n-type source and drain, and p-type poly-silicon is combined with p-type source and drain. Thus, there are typically four available I/O device options (i.e. NCAP, PCAP, PMOS, NMOS) in a generic logic process as shown in a Fig. 3.1 (b).

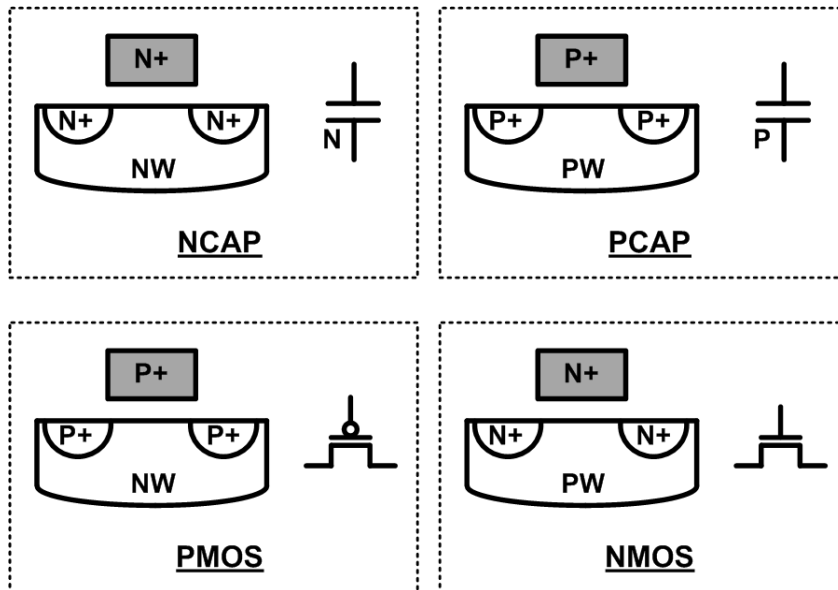
So far, numerous single-poly eflash memory cells using FN tunneling mechanism for program and erase operations have been proposed [29-40], where various device combinations have been adopted for the core three-device structures. For example, NCAP

(i.e. n-type poly and n-type body) is suggested for the coupling (M_1) and the tunneling (M_2) devices, and PMOS is suggested for the read device (M_3) in [29]. Later, PMOS, NCAP, and NMOS are adopted in the coupling (M_1), erase (M_2), and program/read (M_3) devices, respectively, for efficient coupling ratios during program and erase operations in [30]. Then, the work-function engineered n+ poly PMOS as a tunneling device (M_2) was reported to have higher electron ejection efficiency and reliability in [32]. But, an in-depth comparison on both performance and reliability of various single-poly eflash cell structures has not been done yet in these prior studies.

In this chapter, various single-poly eflash memory structures combining various standard I/O devices are theoretically analyzed with experimental data to find the optimal single-poly eflash cell configuration [56]. The remainder of this paper is organized as follows. Section 3.2 reviews various single-poly eflash cell configurations for electron ejection and injection operations. Section 3.3 shows program/erase speed, endurance, retention, and disturbance characteristics of various single-poly eflash cell configurations with the measurement results from two test chips fabricated in a 65 nm low power CMOS process having 5nm tunnel oxide. Finally, a chapter summary is given in Section 3.4.

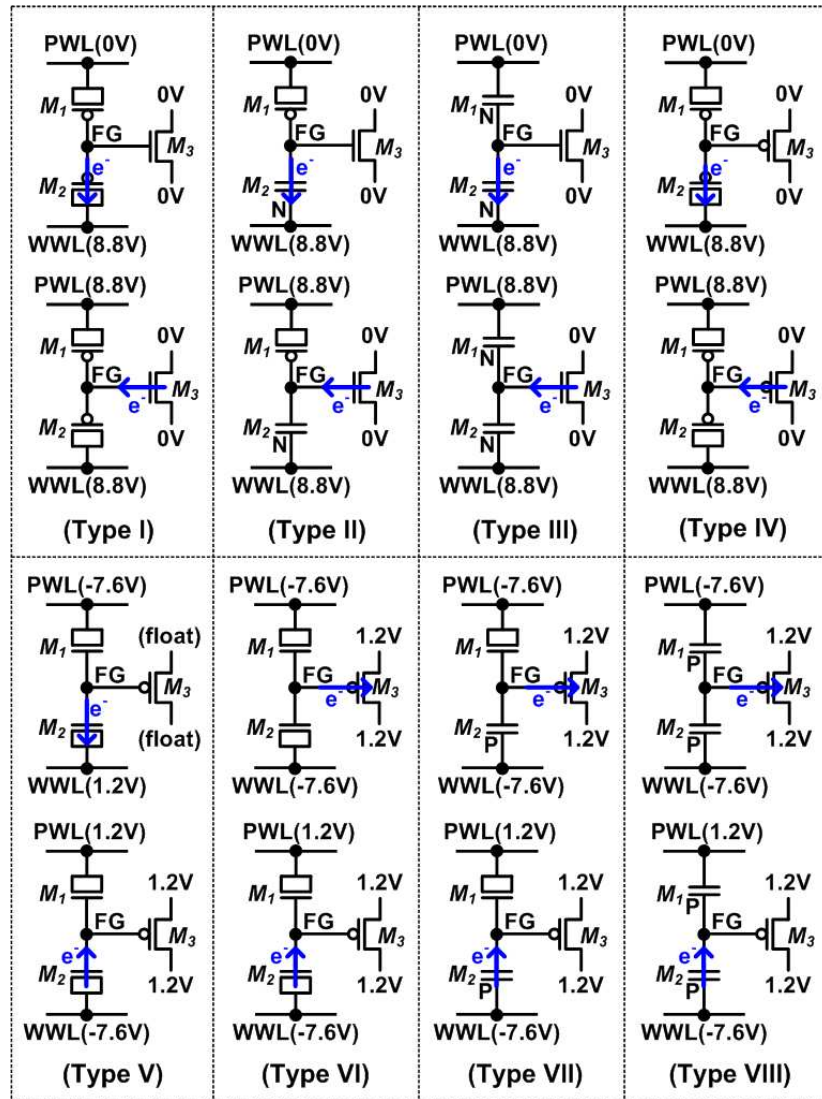


(a)

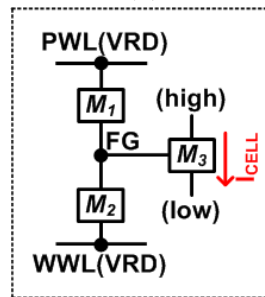


(b)

Fig. 3.1 (a) Bird's eye view of the single-poly eflash memory cell core structure consisting of the three standard I/O devices (M_1 - M_3). (b) The cross section and circuit symbol of the four available standard I/O devices forming the single-poly eflash memory cell core structure.



(a)

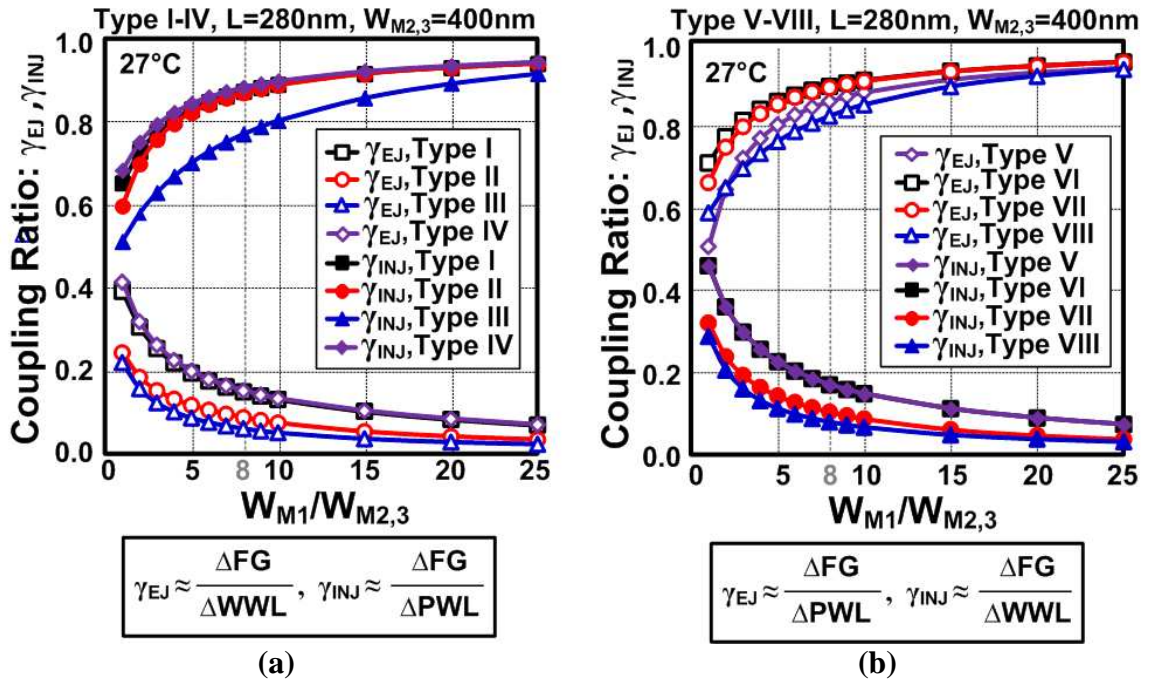


(b)

Fig. 3.2 (a) Electron ejection/injection, and (b) read operations of eight different single-poly eflash cell configurations (2.5V I/O devices, VDD=1.2V). In all configurations, the coupling device (M_1) is upsized compared to the other two devices (M_2, M_3) for efficient electron ejection and injection operations.

3.2 Electron Ejection and Injection Operations in Single-Poly Eflash Cells

The eight different configurations of single-poly eflash cells studied in this chapter are described in Fig. 3.2 with the typical bias conditions for the electron ejection/injection, and read operations. The positive boosted voltage (i.e. 8.8V) is used for the type I-IV configurations, whereas the negative boosted voltage (i.e. -7.6V) is used for the type V-VIII configurations. This allows the read device (M_1) to be controlled by the core supply level (i.e. 1.2V) without being directly connected to these boosted voltage levels. In all configurations, the coupling device (M_1) is upsized compared to the other two devices (M_2, M_3) so that the FG node voltage becomes close to the PWL voltage during electron ejection and injection operations, enabling electron FN tunneling in the ejection devices (M_2 for type I-V, and M_3 for type VI-VIII configurations) and the injection ones (M_3 for type I-IV, and M_2 for type V-VIII configurations). These electron FN tunneling phenomena change the number of the electrons stored in the FG and the cell V_{TH} . During read operation, the cell current is sensed differently between the electron ejected and injected states when the read reference voltage (VRD) is applied to the PWL and WWL. In this chapter, all single-poly eflash cells are implemented using 2.5V standard I/O devices, and the core supply voltage (VDD) is 1.2V. The typical VRD range is from 0V to 1.2V. The detailed operation principle of the single-poly 5T eflash memory cell having type I configuration was elaborately described in chapter 2.



Configuration	M_1	M_2	M_3	$V_{FG_EJ}(V)^*$	$V_{FG_INJ}(V)^*$	γ_{EJ}	γ_{INJ}
Type I	PMOS	PMOS	NMOS	1.3	7.7	0.15	0.87
Type II	PMOS	NCAP	NMOS	0.8	7.6	0.09	0.87
Type III	NCAP	NCAP	NMOS	0.5	6.8	0.06	0.77
Type IV	PMOS	PMOS	PMOS	1.3	7.8	0.15	0.88
Type V	NMOS	NMOS	PMOS	-6.3	-0.3	0.85	0.17
Type VI	NMOS	NMOS	PMOS	-6.6	-0.3	0.89	0.17
Type VII	NMOS	PCAP	PMOS	-6.6	0.3	0.89	0.10
Type VIII	PCAP	PCAP	PMOS	-6.0	0.5	0.82	0.08

*The zero initial charge in FG is assumed

(c)

Fig. 3.3 (a, b) Simulated coupling ratios of eight different eflash configurations (i.e. type I-VIII). (c) Summary of the eflash cell configurations and the simulation results when the width ratio ($=W_{M1}/W_{M2,3}$) is 8.

The coupling ratios of eight different eflash configurations having various M_1 - M_3 combinations (i.e. type I-VIII) and width ratios ($=W_{M1}/W_{M2,3}$) were simulated in Fig. 3.3 using 65nm CMOS technology models with the electron ejection and injection bias conditions illustrated in Fig. 3.2. Since the coupling ratios for electron ejection ($=\gamma_{EJ}$) and

injection ($=\gamma_{\text{INJ}}$) operations are defined as shown in Fig. 3.3 (a, b) for type I-IV and type V-VIII cells, respectively, they were calculated from the FG node voltages (i.e. $V_{\text{FG_EJ}}$ and $V_{\text{FG_INJ}}$) and the bias voltages applied to the cells, assuming the zero initial charge in FG and the negligible inter-poly coupling effect in an eflash array [69]. Lower γ_{EJ} and higher γ_{INJ} are preferred with type I-IV configurations, whereas higher γ_{EJ} and lower γ_{INJ} are preferred with type V-VIII configurations for the higher electron ejection and injection performance. This is because the oxide field can be increased with those coupling ratios, enhancing electron FN tunneling.

The simulation results show that the type II and VII configurations having non-depletion mode coupling device M_1 and depletion mode M_2 provide the well-balanced coupling ratios for both electron ejection and injection operations, while the width ratio of 8 minimizes the unit cell size with reasonable coupling ratios. The eflash cell configurations and the simulation results when the width ratio is 8 are summarized in Fig. 3.3 (c).

3.3 Program/Erase Speed, Endurance, Retention, and Disturbance Characteristics of Eflash Cells

3.3.1 Single-Poly Eflash Cells Fabricated in a 65nm Standard Logic Process

The two single-poly eflash test chips were fabricated in a 65nm standard logic process as a part of this work [54-56]. The die microphotographs of them are shown in Fig. 3.4. The cell with type I configuration and the cells with type V-VIII were included in these test chips, and implemented using 2.5V I/O devices having 5nm tunnel oxide. The

detailed operation principles of the wordline driver (WLD) and charge pump (CP) circuits, and another type of eflash designs included in the second test chip are discussed in chapter 4.

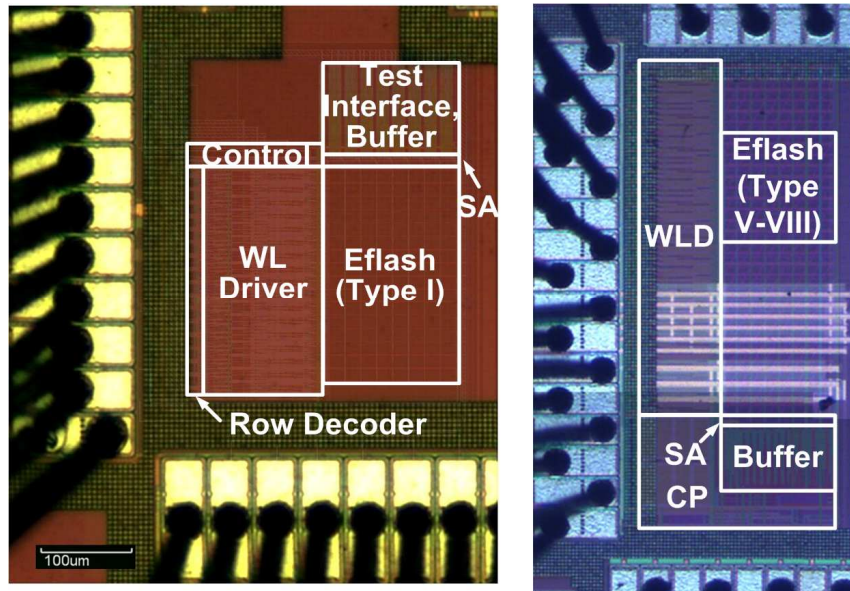


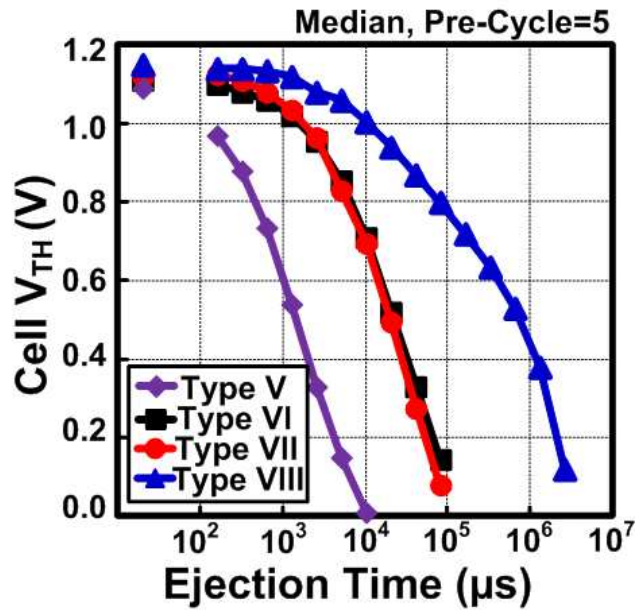
Fig. 3.4 Die microphotographs of the two single-poly eflash test chips fabricated in a 65nm standard logic process [54-56].

3.3.2 Program and Erase Speed of Single-Poly Eflash Cells

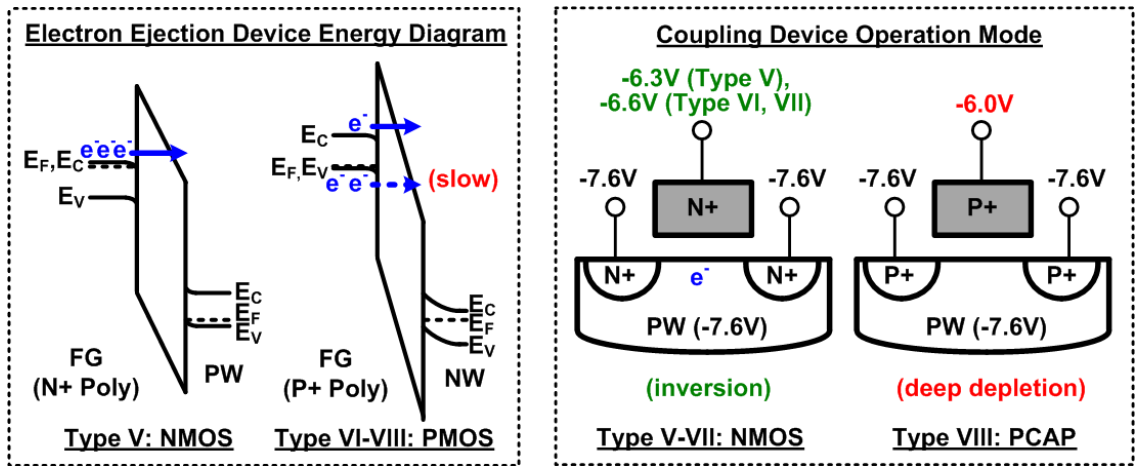
The measured electron ejection speeds of the cells with type V-VIII configurations are shown in Fig. 3.5 with the energy band diagrams of the electron ejection device and the cross sections of the coupling device indicating the operation modes. The bias conditions for the electron ejection operations illustrated in Fig. 3.2 (a) were applied to this measurement. The cell with the type V configuration shows the fastest electron ejection speed, which is because an NMOS electron ejection device in the type V configuration has more conduction band electrons than a PMOS electron ejection device in the type VI-

VIII configurations, as illustrated in Fig. 3.5 (b, left). Note that the conduction band electron tunneling in an electron ejection device with the type V configuration is order-of-magnitude faster than the valence band electron tunneling in an electron ejection device with the type VI-VIII configurations, though the oxide field applied to the electron ejection device is lower with the type V configuration [70]. On the other hand, the cell with the type VIII configuration shows the slowest electron ejection speed. This result is because the PCAP (i.e. p-type poly and p-type source/drain) coupling device in the type VIII configuration operates in a deep depletion mode during the electron ejection operation, reducing the coupling effect from the body of the coupling device M_I to the FG node, whereas the NMOS coupling device in the type V-VII configurations operates in an inversion mode, causing the larger coupling effect from the body of the coupling device M_I to the FG node as illustrated in Fig. 3.5 (b, right).

From the above observations, the electron ejection devices having n-type poly-silicon combined with non-depletion mode coupling device (i.e. the cells with configurations of type II and V) are expected to provide higher electron ejection performance than the others among all eight configurations listed in Fig. 3.2. In fact, the faster electron ejection speed of the cell with the configuration of type II compared to the cell with the configuration of type I was previously predicted by other researchers [30]. Similarly, higher electron ejection current from the n+ poly FG compared to the p+ poly FG was reported in [32], which is well matched to our measurement results shown in Fig. 3.5.



(a)

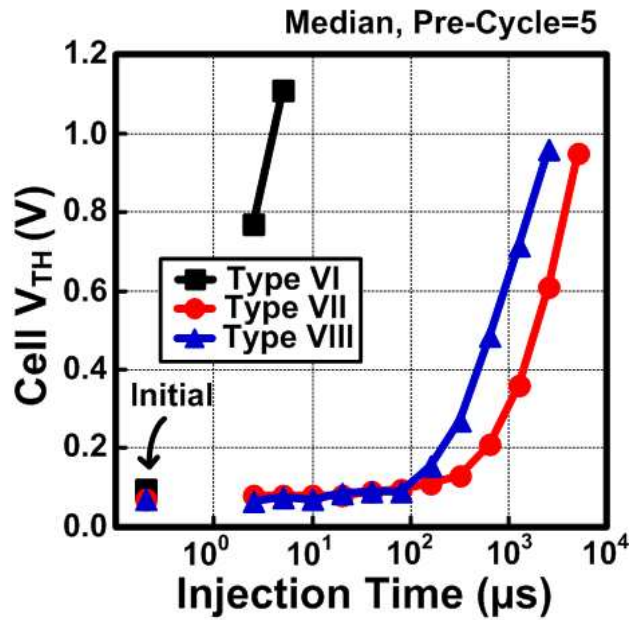


(b)

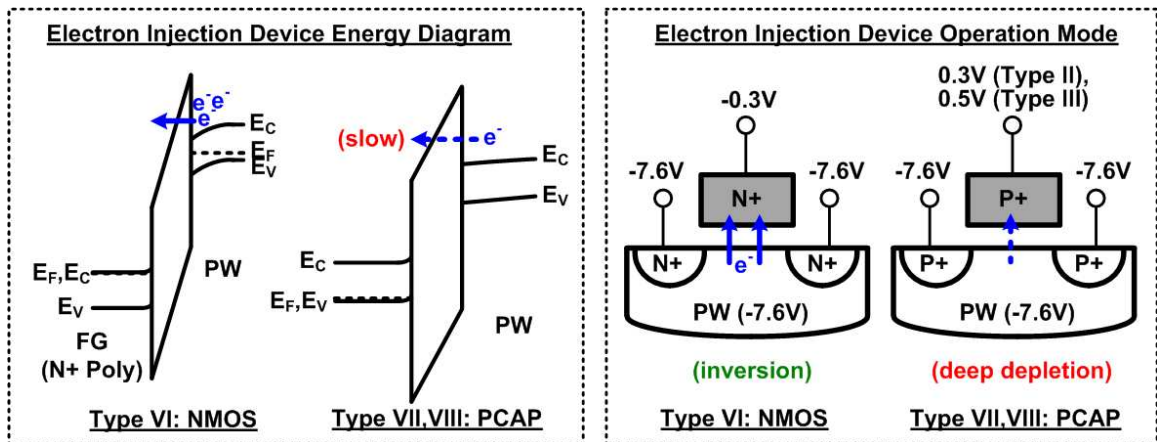
Fig. 3.5 (a) Measured electron ejection speed of various types of eflash memory cells. (b) The energy diagrams of the electron ejection device and the operation modes of the coupling device, which explain the fastest electron ejection speed with the type V configuration and the slowest one with type VIII configuration.

The measured electron injection speeds of the cells with type VI-VIII configurations are shown in Fig. 3.6 with the energy band diagrams and device cross sections of the electron injection device indicating the operation modes. The bias conditions for the electron injection operations given in Fig. 3.2 (a) were applied to this measurement. The cell with the type VI configuration shows the fastest electron injection speed, which is because the NMOS electron injection device in the type VI configuration operates in an inversion mode, whereas the PCAP electron injection device in the type VII, VIII configurations operate in a deep depletion mode reducing the number of available conduction band electrons for FN tunneling as illustrated in Fig. 3.6 (b). On the other hand, the electron injection speed of the cell with the type VIII configuration is faster than that of the cell with the type VII configuration, which is because the coupling device with the type VIII configuration (i.e. PCAP) falls earlier into the accumulation mode than the coupling device with type VII configuration (i.e. NMOS), producing a higher coupling effect from the body of the coupling device to the FG and a higher oxide field in the electron injection device; however, note that this marginal enhancement of the electron injection speed by the depletion mode coupling device is outweighed by the significant improvement of the electron ejection speed by the non-depletion mode coupling device as shown in Fig. 3.5 (a).

From the above observations, the cells with the non-depletion mode electron injection devices (i.e. the cells with configurations of type I-VI) are expected to provide higher electron injection performance than the others among all eight configurations listed in Fig. 3.2.



(a)



(b)

Fig. 3.6 (a) Measured electron injection speed of various types of eflash memory cells. (b) The energy diagrams and operation modes of the electron injection device explaining the fastest electron injection speed with the type VI configuration.

3.3.3 Endurance and Retention Characteristics

Fig. 3.7 (a) shows the measured endurance characteristic of the cells having three different eflash configurations. The negative cell V_{TH} shift after large number of cycling can be explained by the oxide traps generated in the electron injection devices (M_2) [52] modifying the shape of the tunnel barrier as illustrated in Fig. 3.7 (b), which in turn slows the electron injection speed. The cell with the type VI configuration shows the least amount of cell V_{TH} shift. The measured cell V_{TH} distributions after 100 P/E cycles in Fig. 7 (c) shows the larger cell V_{TH} variations of the electron injected state for the cells with the configuration of type VII and VIII, further reducing the sensing margin. These larger variations can be explained by the larger cell-to-cell variation of the deep depletion behavior of the PCAP electron injection device during the electron injection operation of the cells with the configuration of type VII and VIII. On the other hand, some holes from the channels of the read device (M_3) can be activated and trapped in the gate oxide of this read device (M_3) during electron ejection operations of the cells with the configuration of type VI-VIII, causing the similar cell V_{TH} negative shift trend shown in Fig. 3.7 (a), but the larger variation of the electron injected states of the cells with the configuration of type VII and VIII is not clearly explained with this anode hole injection theory [71, 72].

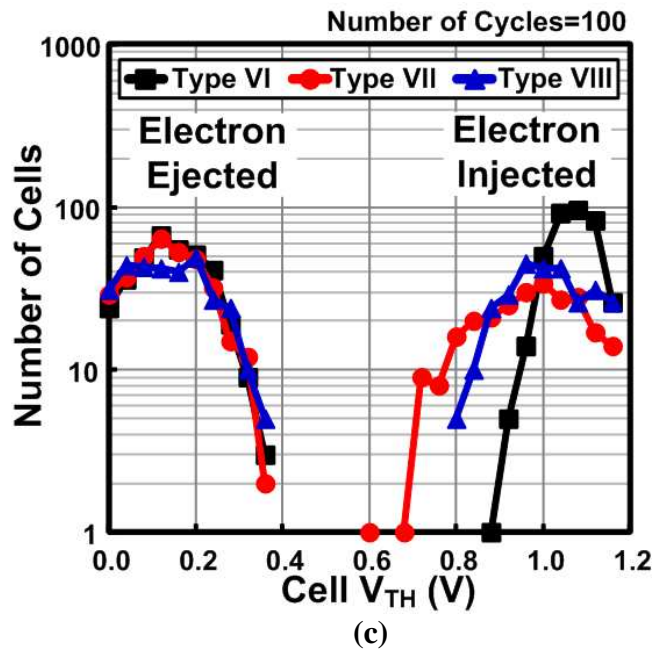
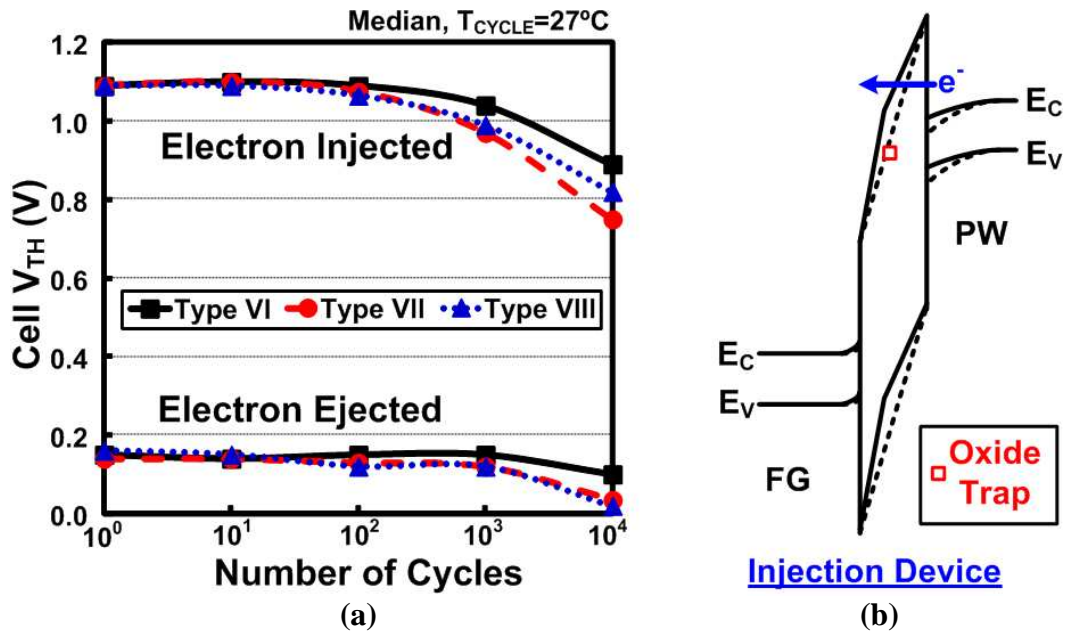


Fig. 3.7 (a) Measured endurance characteristic of the cells having three different eflash configurations (Type VI-VIII). (b) Energy diagram of the electron injection device where the oxide traps modifies the shape of the tunnel barrier. (c) Measured cell V_{TH} distributions of the cells having three different eflash configurations after 100 P/E cycles.

From the above discussions, the cells with the non-depletion mode electron injection devices (i.e. the cells with configurations of type I-VI) are expected to provide higher endurance than the others among all eight configurations listed in Fig. 3.2.

Fig. 3.8 shows the measured retention results of the 1k and 3 P/E pre-cycled cells having three different configurations of type VI-VIII. The smaller cell V_{TH} shifts for the cell with the configuration of type VIII were observed compared to the cells with the configurations of type VI and VII, which is explained by the superior intrinsic retention characteristic of the cell with the coupling device (i.e. M_1 in Fig. 3.1) having p+ poly silicon. 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 [70]. Thus, the cells with a coupling device having p-type poly-silicon (i.e. the cells with configurations of type I, II, IV, and VIII in Fig. 3.2) dominantly reduce the gate leakage current and are preferred for longer retention time, as the coupling device is designed to be 8 times wider than the other two devices (i.e. M_2 and M_3 in Fig. 3.1).

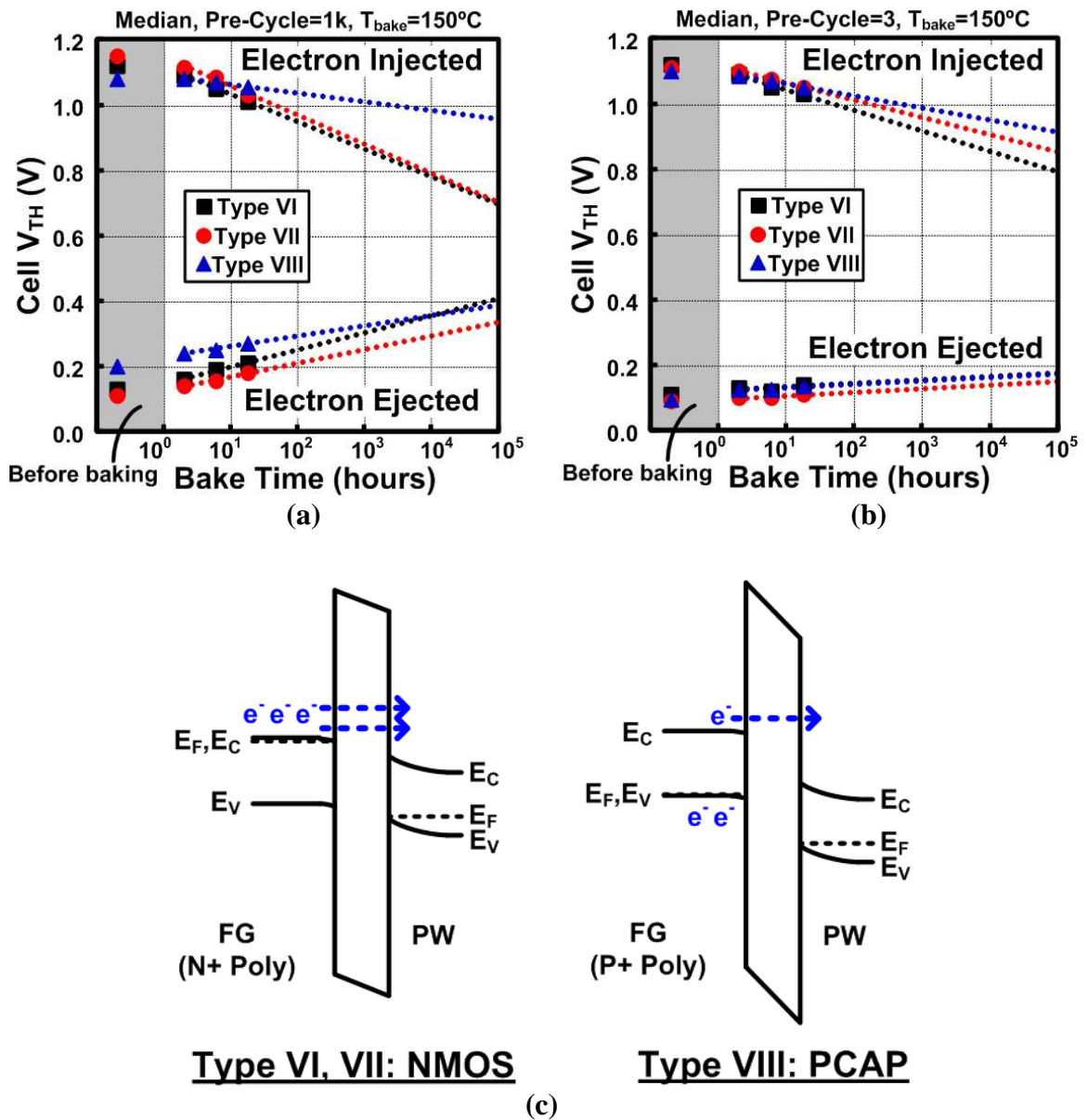


Fig. 3.8 Measured retention result of the cells having three different configurations (Type VI-VIII) after (a) 1k and (b) 3 P/E pre-cycles. (c) Energy band diagrams of NMOS (i.e. Type VI, VII) and PCAP (i.e. Type VIII) coupling devices explaining that the cells with the coupling device (M_I) having n-type poly-silicon has more conduction band electrons causing more intrinsic charge loss from the floating gate.

3.3.4 Program Disturbance

The program operation of the single-poly eflash cell can be achieved via self-boosting method [58, 59] with 5T cell structure, while not requiring the boosted BL voltage as discussed in chapter 2 [54, 55]. The program bias conditions of 5T eflash cells with type I and VI configurations are shown in Fig. 3.9 (a) and (b), respectively. For the inhibited BL cells, the channel voltages of read device (M_3) need to be boosted enough and the leakage currents from/to the boosted channel should be suppressed to minimize the program disturbance of the unselected BL cells [73]. The long channel pass transistor connecting the read device to BL and the short program pulse width are preferred to maintain the boosted channel voltages high enough. The measured program disturbance characteristics of the cells with type I and type VI-VIII configurations are shown in Fig. 3.9 (c) and (d), respectively. A voltage margin of $\sim 4V$ between the programmed and the inhibited cells was measured for 10k pre-cycled cells having type I configuration, and a negligible cell V_{TH} disturbance up to a $\sim 1s$ pulse was measured for 10k pre-cycled cells having type VI-VIII configurations. These results confirm the effectiveness of the self-boosting technique with the 5T cell structure shown in Fig. 3.9, which allows the WL-by-WL accessible array architecture with no high voltage disturbance issue of the unselected WL cells as addressed in chapter 2 [54, 55]. Though the 5T cells with all the type I-III and type VI-VIII configurations in Fig. 3.2 support this self-boosting technique, note that the 5T cells with the type I-III configurations are preferred, since the 5T cells with the type VI-VIII configurations require negative boosted supplies having the junction breakdown

concerns of the high voltage circuits (i.e. WLD and CP in Fig. 3.4) discussed further in chapter 4 [57].

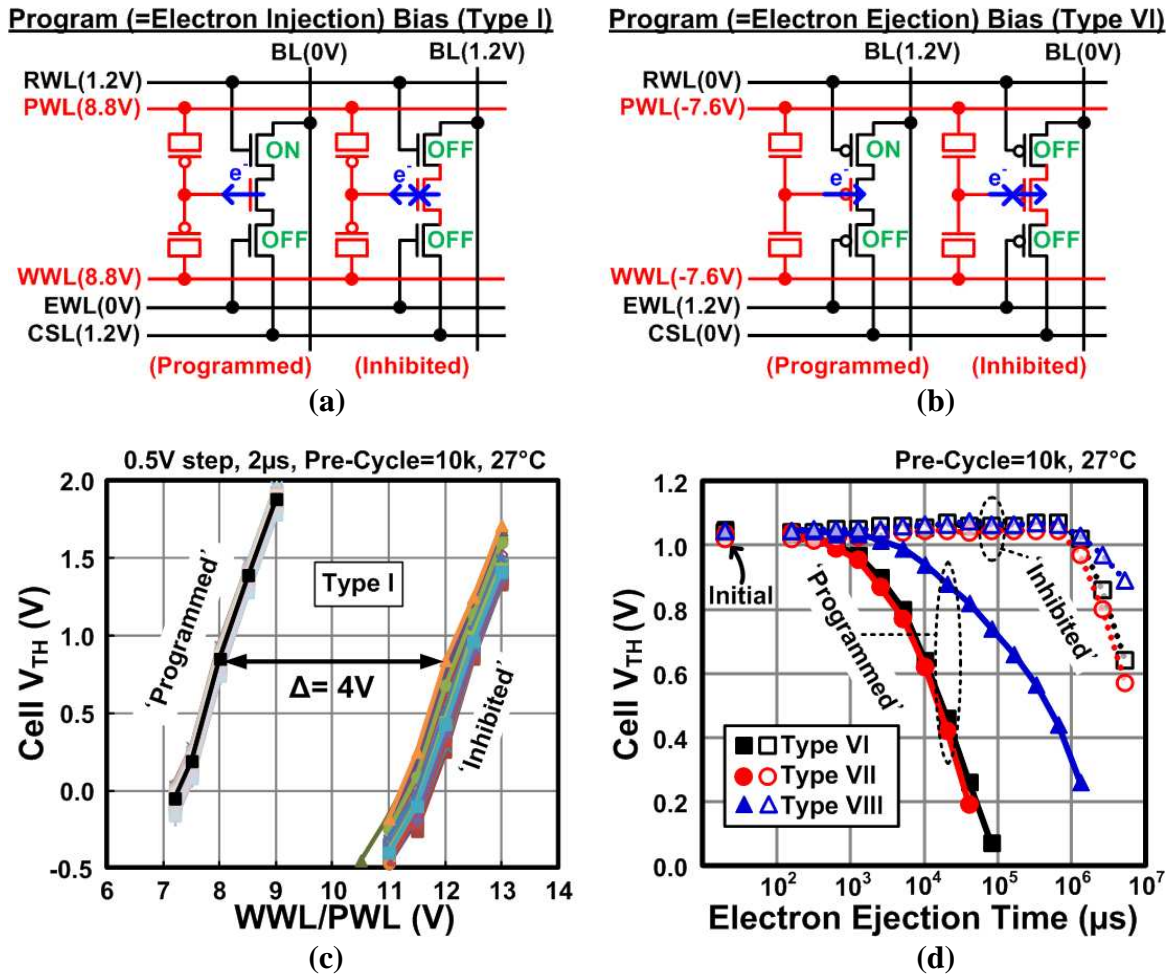


Fig. 3.9 (a, b) Program bias conditions of 5T eflash cells with type I and VI configurations. (c, d) Measured program disturbance characteristics of the cells with type I and type VI-VIII configurations.

3.3.5 Read Disturbance

The read disturbance of the 5T cell with the type I configuration is discussed in Fig. 3.10. The read stress voltage (VSTR) higher than the nominal read reference voltage is applied to PWL and WWL to measure the voltage-accelerated life time as illustrated in Fig. 3.10 (a). The life-time end under this read stress condition is defined as the moment when the cell V_{TH} tail reaches the sensing limit after which the cell is sensed incorrectly as illustrated in Fig. 3.10 (b). The read stress voltages tested in this work were 3.2, 4.0, 5.0, and 6.0V, and they were applied to PWL and WWL of the erased cells until the read fail occurred with a sensing limit of 0V. The measured read disturbance result of the 5T eflash cell with the type I configuration is shown in Fig. 3.10 (c).

The results predict almost negligible read disturbance up to the VSTR of 2V, considering the read access time of <10ns [55]; however, note that the disturbance can be significant when the higher VSTR such as 5V is applied to PWL and WWL, limiting the life time within 10ms for the 10k pre-cycled cells. Thus, this high voltage operation is only allowed up to ~1k times assuming the access time of 10ns with VSTR of 5V.

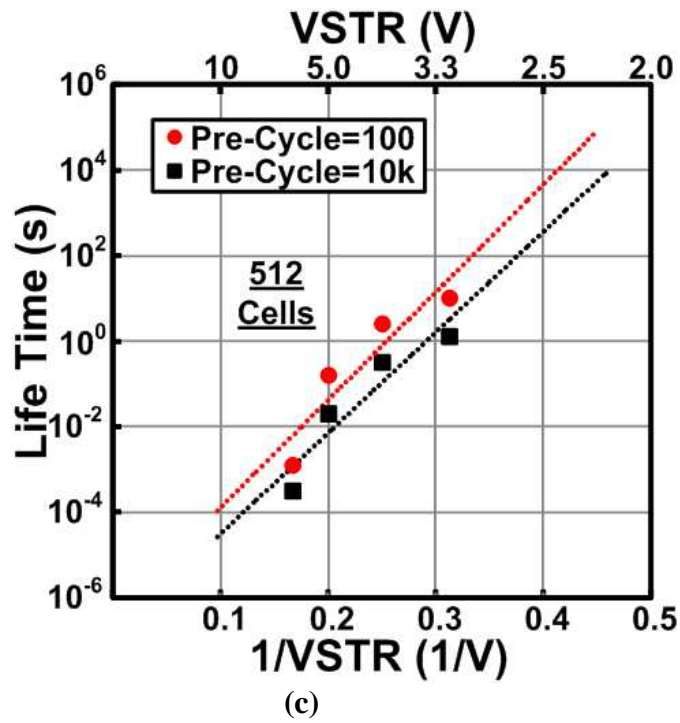
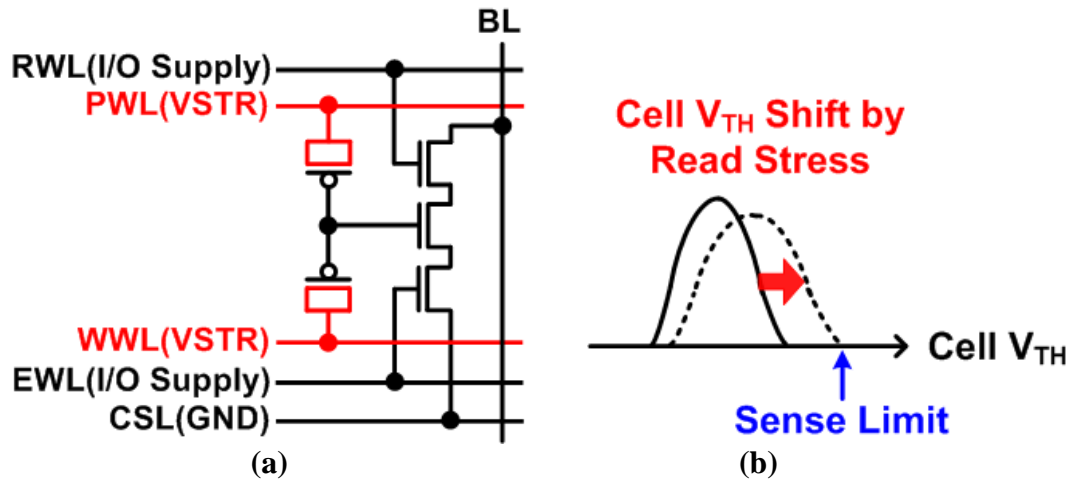


Fig. 3.10 (a) Read stress condition to measure the voltage-accelerated life time. (b) Definition of the life time. The cell V_{TH} tail reaches the sensing limit by the read stress at the end of the life time. (c) Measured life time of the 5T eflash cells with type I configuration.

3.3.6 Floating Gate Coupling Effect

The 5T cell layout is shown in Fig. 3.11 (a). The tighter BL pitch and shorter FG coupling distance of this type cell than other single-poly eflash cells [29, 30] may increase the parasitic inter-FG coupling effect. 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 as shown in Fig. 3.11 (b). The measured result from the cells with type I configuration in Fig. 3.11 (c) shows a modest change in the mean and standard deviation of the cell V_{TH} distribution (17mV and 3.7% respectively) due to this FG coupling effect, confirming no significant coupling issues are found in 5T cell having a tight BL pitch by sharing the n-wells between adjacent BL cells and adopting the self-boosting technique.

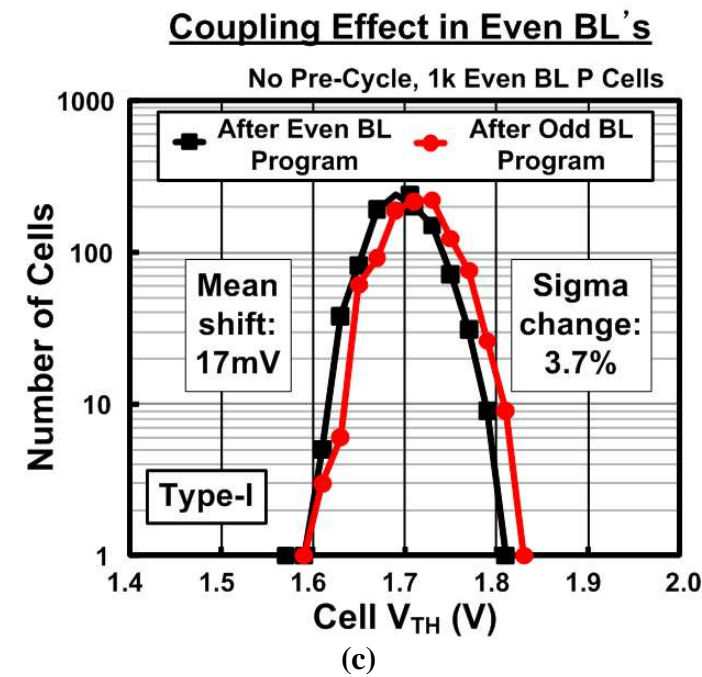
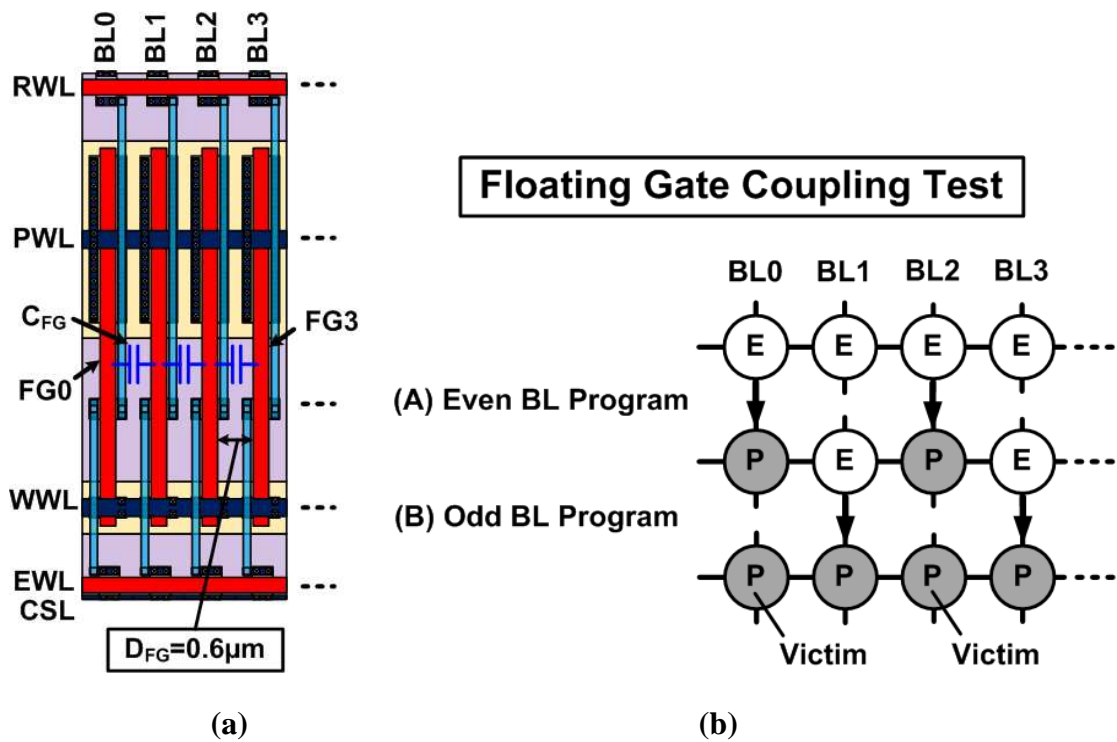


Fig. 3.11 (a) 5T eflash cell array layout. (b) Floating gate coupling test sequence. (c) Measured floating gate coupling effect of the 5T eflash cells with type I configuration.

3.4 Chapter Summary

In this chapter, various single-poly eflash cell topologies were discussed to find the optimal configuration. The results summarized in Table 3.1 show that the 5T eflash cell structure with the type II configuration having PMOS coupling device (M_1), NCAP electron ejection device (M_2) and NMOS electron injection (or read) device (M_3) is the most favorable configuration among all eight configurations listed in Fig. 3.2, when the performance, endurance, retention, and disturbance characteristics are all considered. This preferred 5T single-poly eflash cell with the type II configuration shown in Fig. 3.12 is built using standard I/O devices that are readily available in a generic logic process and therefore it is an attractive moderate density eNVM candidate where a dedicated eflash process is not available.

Table 3.1 Summary of the Study on the Optimal Configuration of Single-Poly Eflash Cell

Criteria	Coupling Ratios	Ejection Performance	Injection Performance / Endurance	Intrinsic Retention	High Volt. Disturbance	
Result	Fig. 3	Fig. 5	Figs. 6-7	Fig. 8	Figs. 9-11	
Preferred Char.	M_1 with non-depletion, M_2 with depletion	M_1 with non-depletion, M_2 having n-type poly	Injection device with non-depletion	M_1 having p-type poly	Self-boosting without disturbance/coupling issues	
Configurations	Type I	-	-	Yes	Yes	Yes
	Type II	Yes	Yes	Yes	Yes	Yes
	Type III	-	-	Yes	-	Yes
	Type IV	-	-	Yes	Yes	-
	Type V	-	Yes	Yes	-	-
	Type VI	-	-	Yes	-	Yes
	Type VII	Yes	-	-	-	Yes
	Type VIII	-	-	-	Yes	Yes

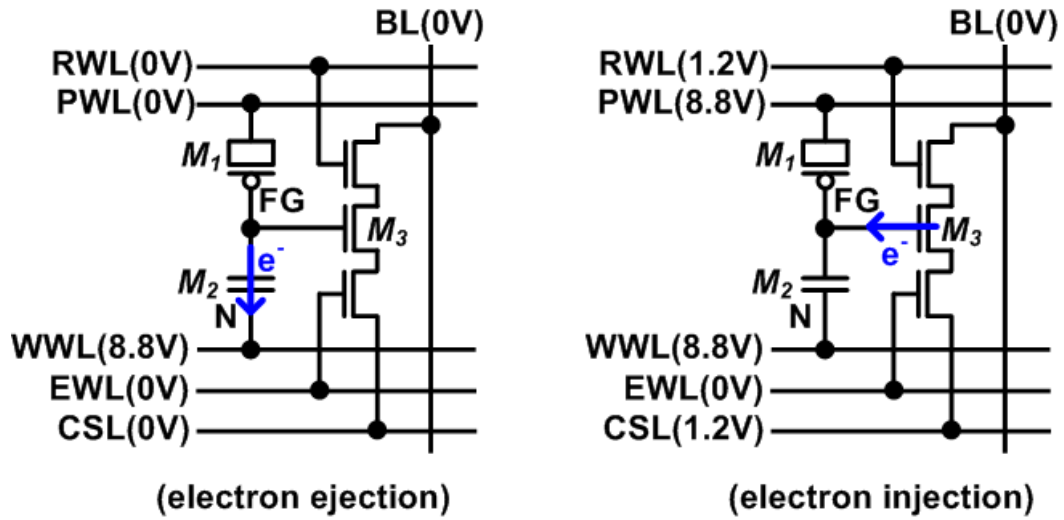


Fig. 3.12 The preferred 5T single-poly eflash cell with the type II configuration when the performance, endurance, retention, and disturbance characteristics are all considered.

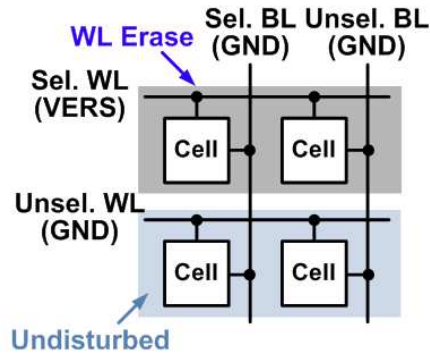
Chapter 4 A Bit-by-Bit Re-Writable 6T Eflash in a Generic Logic Process

Various eNVM technologies have been explored for high density applications including dual-poly embedded flash (eflash), FeRAM, STT-MRAM, and RRAM. On the other end of the spectrum, logic compatible eNVM such as e-fuse, anti-fuse, and single-poly eflash memories have been considered for moderate density low cost applications. In particular, single-poly eflash memory has been gaining momentum as it can be implemented in a generic logic process while supporting multiple program-erase cycles. One key challenge for single-poly flash is enabling bit-by-bit re-write operation without a boosted bitline voltage as this could cause disturbance issues in the unselected wordlines. In this chapter, we present details of a bit-by-bit re-writable embedded flash memory implemented in a generic 65nm logic process which addresses this key challenge. The proposed 6T eflash memory cell can improve the overall cell endurance by eliminating redundant program/erase cycles while preventing disturbance issues in the unselected wordlines. Details of the overstress-free high voltage switch and voltage doubler based charge pump circuit are also described.

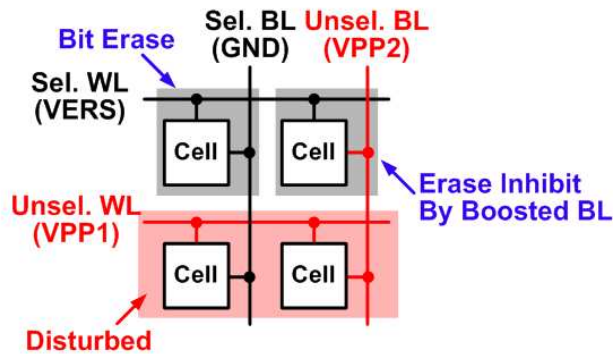
4.1 WL-by-WL and Bit-by-Bit Erasable Single-Poly Eflash Memories

One of the key endurance limiting factors for single-poly eflash is the large number of unnecessary erase cycles undergone by a cell when data is written to the entire wordline. To illustrate this issue, a high level comparison between WL-by-WL and bit-by-bit erasable single-poly eflash memories is shown in Fig. 4.1. In all three cases, a boosted voltage (VERS) is applied to the selected WL to induce FN tunneling in the cells to be erased. Prior WL-by-WL erasable eflash memories [38-40, 54-56] require the entire WL to be erased simultaneously prior to the program operation, which results in unnecessary erase cycles (Fig. 4.1, top). Prior bit-by-bit erasable eflash cells [36] on the other hand can erase the selected cells without incurring unnecessary erase cycles (Fig. 4.1, middle). However, a boosted BL voltage ($=VPP2$) has to be applied to the unselected BL's for erase inhibition. Since $VPP2$ in the BL direction will cause disturbance in the unselected WL cells, yet another boosted voltage ($=VPP1$) must be applied to the unselected WL's. The problem here is that for typical $VPP2$ and $VPP1$ voltage levels (e.g. 10V and 5V), the difference between the two is still greater than the nominal I/O VDD ($=2.5V$ or $3.3V$) so high voltage disturbance issues cannot be completely avoided in the unselected WL cells. The proposed eflash illustrated in Fig. 4.1 (bottom) achieves bit-by-bit erase using an FG boosting scheme that does not require a boosted BL voltage and thereby eliminating disturbance issues in the unselected WL cells. Note that bit-by-bit re-write and altering techniques [74, 75] proposed for stand-alone NAND flash memories cannot be directly applied to single-poly eflash memories as the implementation of the later must be done in a generic logic process.

Prior WL-by-WL Erasable Eflash



Prior Bit-by-Bit Erasable Eflash



Proposed Bit-by-Bit Erasable Eflash

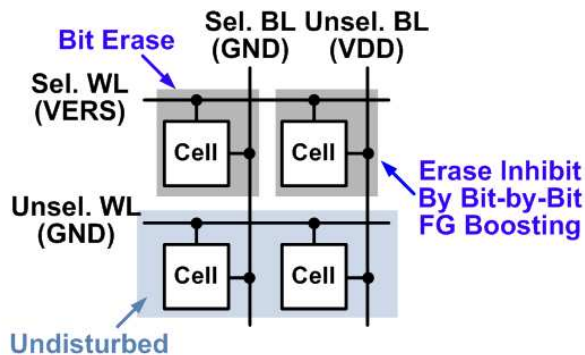


Fig. 4.1 Disturbance issue comparison between WL-by-WL and bit-by-bit erasable single-poly eflash cells. The prior WL-by-WL erasable eflash requires all cells in the selected WL to be erased simultaneously, which results in unnecessary erase cycles for cells whose data remain unchanged. The prior bit-by-bit erasable eflash requires a boosted BL voltage (VPP2) for erase inhibition of the unselected BL cell. This will cause high voltage disturbance issues in the unselected WL. In contrast, the proposed eflash enables bit-by-bit erase via a novel FG boosting scheme (See Fig. 4.2) minimizing disturbance issues.

In this chapter, we propose a bit-by-bit re-writable eflash in a generic logic process based on the bit-by-bit erasable cell structure in [57]. The remainder of this paper is organized as follows. Section 4.2 describes the proposed 6T eflash memory and cell operation principles. Section 4.3 describes the negative high voltage switch and charge pump circuit design. Section 4.4 presents the measurement results from a test chip fabricated in a standard 65nm logic process. Comparison with other logic compatible eNVM designs is given in Section 4.5, followed by a conclusion in Section 4.6.

4.2 Proposed Bit-by-Bit Re-Writable 6T Eflash Memory

4.2.1 Bit-by-Bit Floating Gate Boosting Scheme

To accomplish a bit-by-bit write operation without disturbing the cells in the unselected WL's, we propose the bit-by-bit FG boosting scheme described in Fig. 4.2. Here, the bias conditions of the coupling transistor are compared to those of the prior WL-by-WL FG boosting scheme. When the selected WL is switched from 0V to a boosted write voltage $-V_{PP}$, the prior WL-by-WL scheme boosts all the FG in the selected WL irrespective of their BL levels. This is because the source and drain of the coupling transistors are tied to a shared body (i.e. Selected WL). In contrast, the proposed scheme selectively boosts the FG depending on the BL data. For example, FG boosting is stronger for a '0' BL cell compared to that of a '1' BL cell (i.e. $-V_{PP} < -V_H < -V_L$), as the source and drain of the coupling transistors are boosted together for a '0' BL but tied to VDD for a '1' BL through the additional pass transistor. An NMOS coupling transistor

was chosen over a PMOS one, as the later suffers from the floating channel issue when a positive read reference (i.e. VRD in Fig. 4.5) is used during read operation.

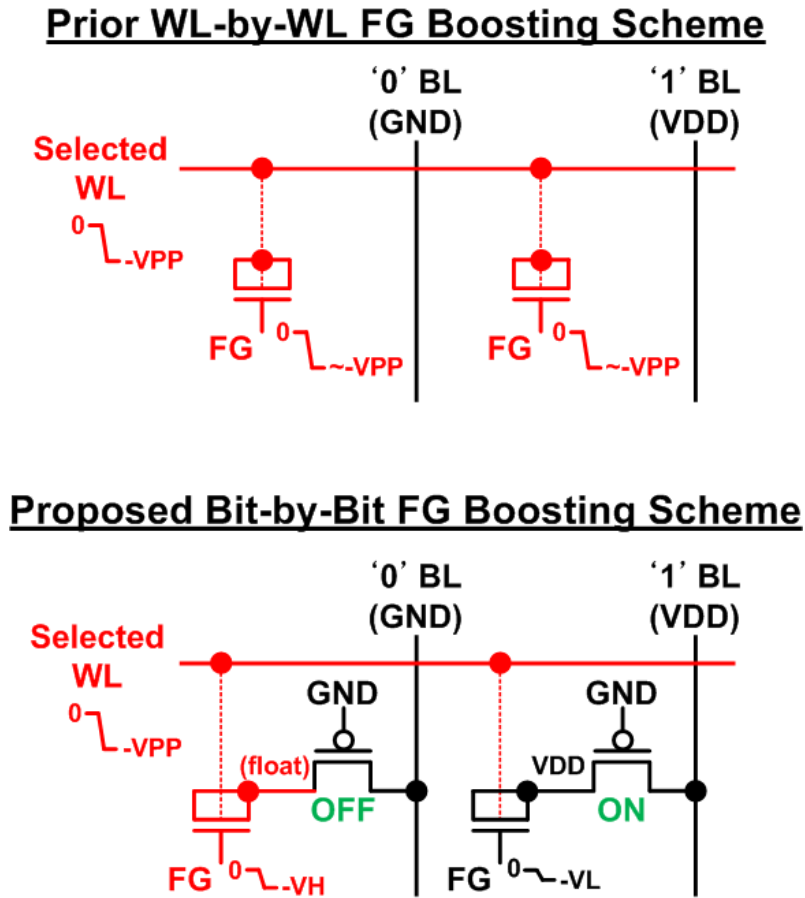


Fig. 4.2 Bias conditions of the coupling TR compared between the prior WL-by-WL and the proposed bit-by-bit FG boosting schemes. The former boosts all the FG's in the selected WL irrespective of the BL levels while the later selectively boosts the FG depending on the write data (i.e. $-VPP < -VH < -VL$).

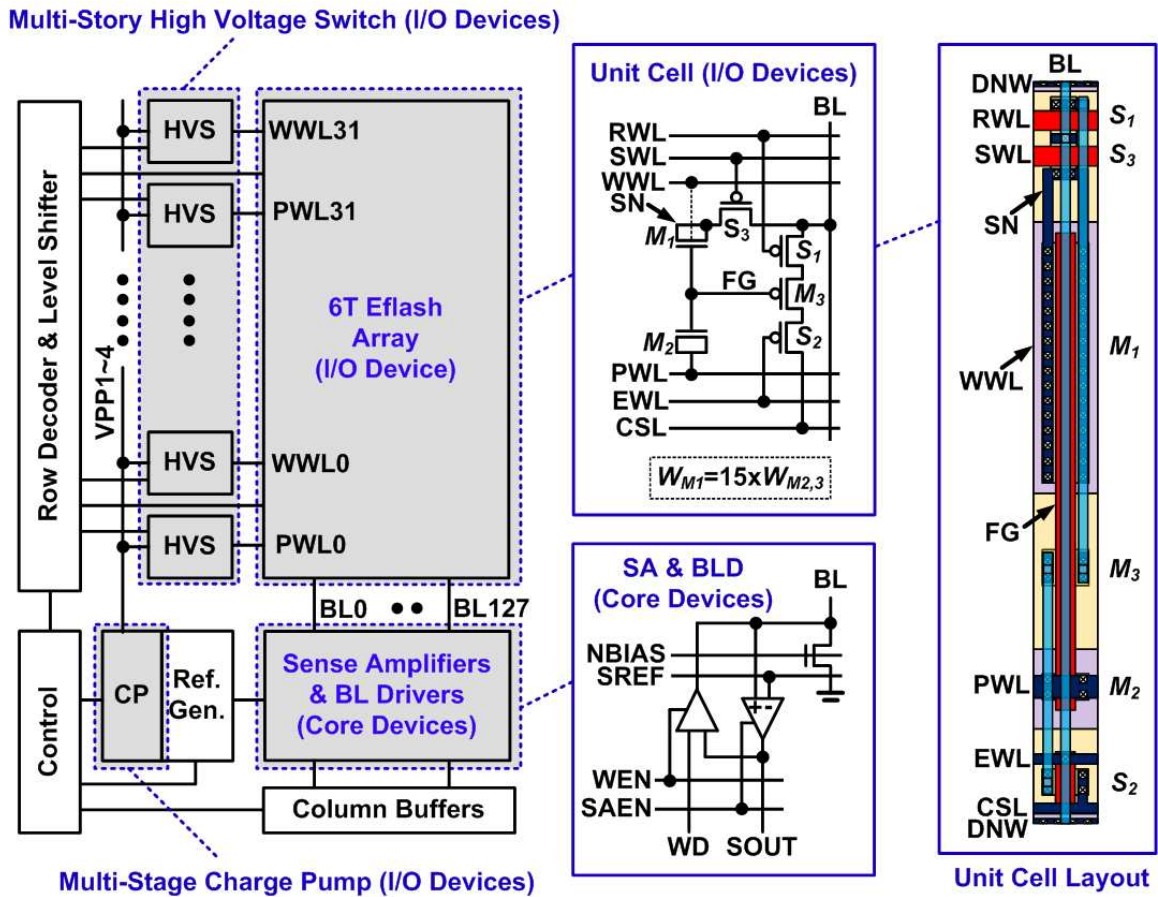


Fig. 4.3 Test chip diagram of the proposed bit-by-bit re-writable eflash memory. The 6T eflash cell array, multi-story high voltage switch, and multi-stage charge pump are implemented using standard 2.5V I/O devices with a 5nm gate oxide. The sense amplifiers and BL drivers are implemented using 1.2V core devices.

4.2.2 Bit-by-Bit Re-Writable Eflash Memory Overview

The array architecture of the proposed bit-by-bit re-writable eflash is shown in Fig. 4.3 along with the schematic and layout of the new 6T cell. The 6T eflash cell is composed of three main transistors (M_1 - M_3) and three pass transistors (S_1 - S_3) such that the SN node is selectively connected to BL through pass transistor S_3 depending on the BL signal. This enables the aforementioned bit-by-bit FG boosting scheme. The width of the coupling

transistor (M_1) is designed to be 15 times wider than that of the write (M_2) and read (M_3) transistors. By doing so, we can significantly lower the program and erase voltages compared to dual-poly eflash. The p-wells and Deep N-Well (DNW) are shared in the WL direction for a compact layout. The 6T eflash cell array, multi-story High Voltage Switch (HVS), and multi-stage Charge Pump (CP) are implemented using 2.5V I/O devices having a 5nm gate oxide, while the sense amplifiers and BL drivers are implemented using 1.2V core devices.

4.2.3 Cell Operation of the Proposed 6T Eflash Memory

The complete write operation of the proposed 6T eflash consists of the two phases illustrated in Fig. 4.4. Firstly, SWL is driven to GND while BL is driven to either GND or VDD depending on the write data. Subsequently, during the bit-by-bit write '0' phase, WWL is switched to a negative boosted voltage, $-V_{PP}$ (e.g. $-7.2V$). Then, the pass transistor in a '0' BL cell (i.e. S_3 in Fig. 4.3) is turned off, while this transistor is turned on in a '1' BL cell. Under this signal bias condition, the '0' BL cell experiences a stronger FG boosting compared to the '1' BL cell. Consequently, the FG node voltage of the '0' BL cell is boosted to a large negative voltage ($-V_H$) while the '1' BL cell sees only a small negative voltage ($-V_L$). PWL is driven to a small positive voltage, VRD (e.g. $1.6V$) during the write '0' phase and thus electron FN tunneling occurs in the '0' BL cell. During write '1' phase, PWL is switched to the negative boosted voltage $-V_{PP}$ which generates a sufficiently high electric field for electron FN tunneling in the '1' BL cells. Assuming no initial charge in FG, the typical boosted FG node voltages $-V_H$ and -

VL are simulated as -4.2V and -0.6V during write '0' phase, and -5.2V and -1.9V during write '1' phase, respectively, for a -VPP of -7.2V and VRD of 1.6V.

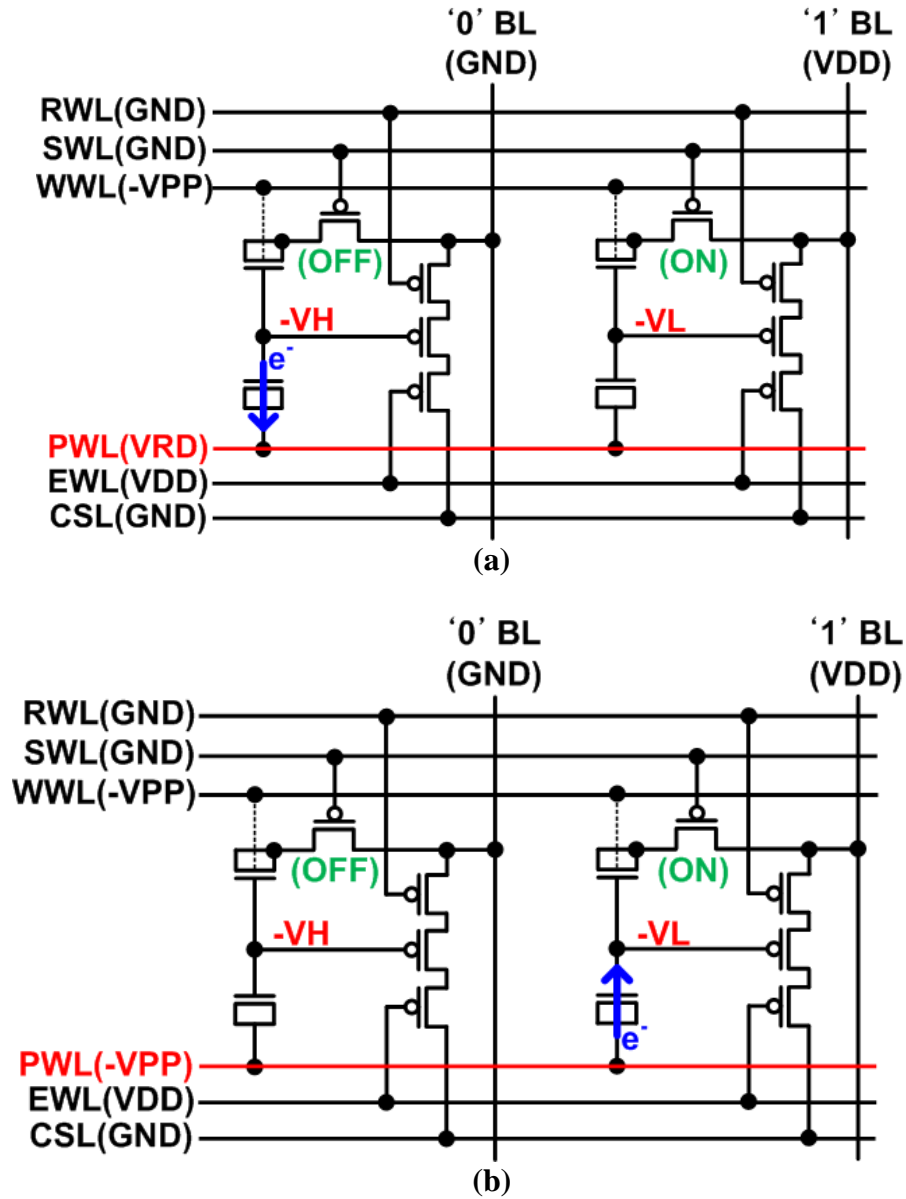
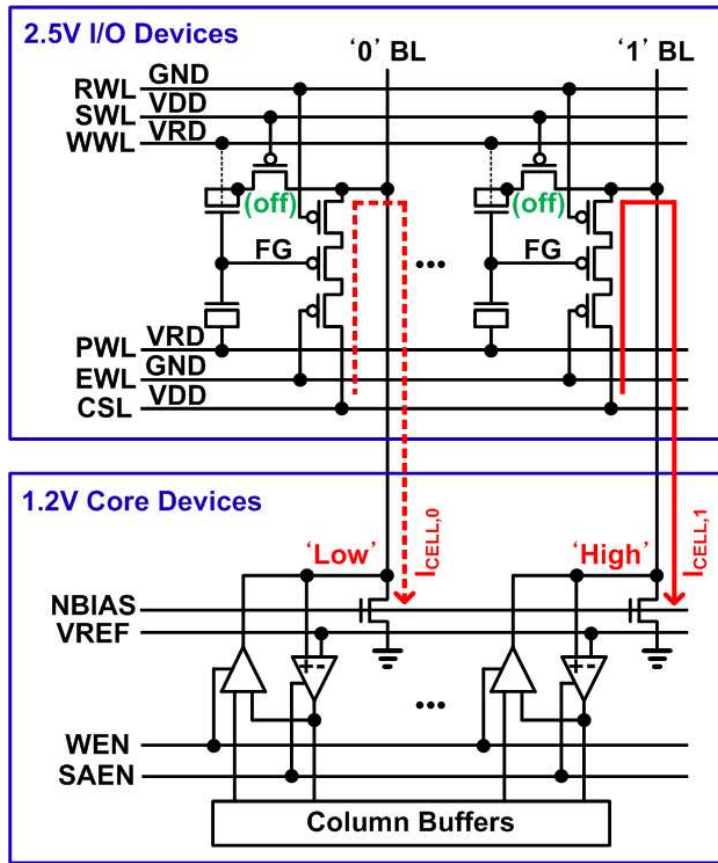


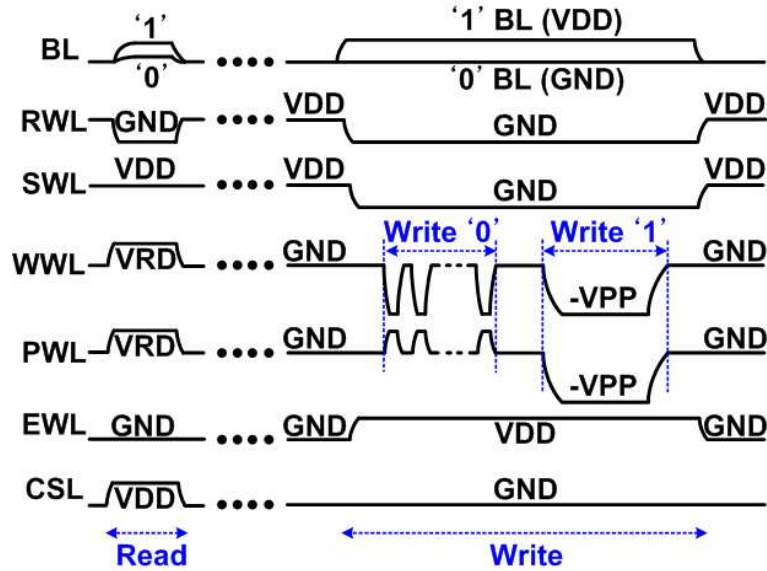
Fig. 4.4 Bit-by-bit (a) write '0' and (b) write '1' phases of the proposed 6T eflash cell. The '0' BL cell loses electrons from FG during a write '0' phase, whereas the '1' BL cell adds electrons in FG during a write '1' phase via electron FN tunneling.

Fig. 4.5 shows the read bias condition of the proposed 6T eflash cell along with the timing diagram for the full bit-by-bit update sequence. During read operation, the pass transistor S_3 (i.e. S_3 in Fig. 4.3) in all selected WL cells is turned off as SWL is driven to VDD which is greater than the nominal read reference level (VRD). Subsequently, VRD is applied to WWL and PWL while CSL is driven to VDD. FG node of the '1' cell contains more electrons than that of the '0' cell, thereby generating a higher BL current. The BL voltage levels are compared to the reference level VREF, using conventional voltage sense amplifiers.

The full bit-by-bit update sequence of the proposed 6T eflash cell consists of a read and a write operation. Firstly, the read operation is conducted on the selected WL and the sensed data is stored in the column buffers. After replacing the old data stored in the column buffers with the new values, write operation is carried out to the same selected WL. As noted earlier, the write operation comprises a write '0' phase and a write '1' phase. Multiple short pulses need to be applied during write '0' phase for sufficient cell V_{TH} margin according to the measured write '0' speed shown in Fig. 4.12 (top left). This is due to the strong coupling between the SN (shown in Fig. 4.3) nodes of adjacent cells that reduces the FG boosting effect for '0' BL cells explained in Fig. 4.2.



(a)



(b)

Fig. 4.5 (a) Read bias condition of the proposed 6T eflash cell and (b) Timing diagram for the full bit-by-bit update sequence.

4.3 Negative High Voltage Switch and Charge Pump

4.3.1 Negative High Voltage Switch

The proposed multi-story negative HVS is illustrated in Fig. 4.6. This is a modified version of the original multi-story positive HVS published in [54, 55]. The HVS consists of stacked latch and driver stages, and are implemented using 2.5V standard I/O devices. The stacked configuration effectively prevents gate overstress during the read and write operations. The HVS is used as the WWL and PWL drivers. The boosted negative voltages VPP1~4 are supplied from the negative CP shown in Fig. 4.7. The nominal boosted voltage levels for VPP1~4 are -0.9, -3.0, -5.1, and -7.2V, respectively.

For read operation, the SRDB signal switches from VDD to VPP1, so that VRD is connected to WWL through the PMOS stack as illustrated in Fig. 4.6 (b). This PMOS signal path enables high speed WWL activation, as the stacked latches do not change their states during read operation. When the SRDB signal returns from VPP1 to VDD, WWL is discharged to GND.

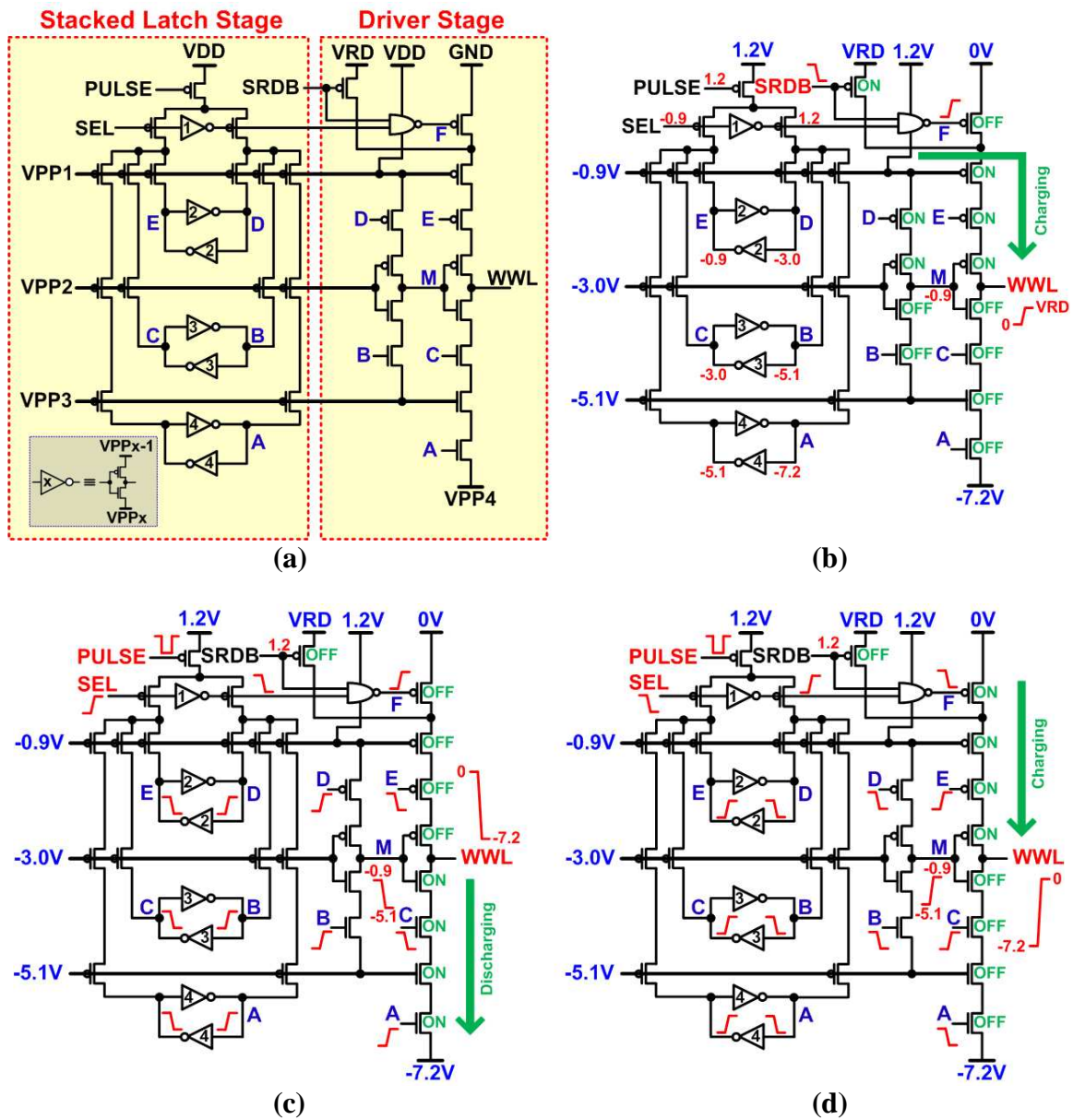


Fig. 4.6 (a) Multi-story negative high voltage switch consists of a stacked latch stage and a driver stage which prevents gate overstress during read and write operation. (b) During read, WWL is driven to VRD through the PMOS string without changing the latch states. (c, d) During write, WWL is switched between VPP4 and GND by the SEL signal.

During write operation, WWL switches from GND to VPP4, which is triggered by the SEL signal changing the three stacked latch states. When SEL switches from VPP1 to VDD (Fig. 4.6 (c)), nodes C and E are pulled-down to VPP3 and VPP2, and nodes A, B, D, and F are pulled-up to VPP3, VPP2, VPP1, and VDD, respectively, making the intermediate node 'M' and output node WWL connected to VPP3 and VPP4 levels, respectively. When SEL switches from VDD to VPP1 (Fig. 4.6 (d)), nodes C and E are pulled up to VPP2 and VPP1, and nodes A, B, D, and F are pulled down to VPP4, VPP3, VPP2, and VPP1, respectively. As a result, node 'M' and output nodes WWL are driven to VPP1 and GND. Similar to the previous positive HVS design described in chapter 2 [54, 55], the PULSE signal width and the transistor sizes are optimized such that the latch states switch reliability while static power consumption kept small so as to minimize the current loading of the negative CP.

4.3.2 Negative Charge Pump

The proposed negative HVS requires multiple boosted negative voltages, and these multiple boosted negative voltages are generated from the voltage doubler [61] based on-chip negative CP shown in Fig. 4.7. Each voltage doubler stage is cascaded to provide multiple boosted supplies (VPP1-VPP4) without experiencing gate oxide reliability issues. Similar cascading techniques have been widely adopted for high efficiency charge pump designs in a standard CMOS logic process [62, 63], as this configuration can prevent threshold voltage drop without a complicated clocking scheme. A deep n-well

surrounds the VPP1-VPP4 p-wells for the isolation purpose. The write voltage level (VPP4) is regulated by comparing the resistively divided voltage level against a reference voltage (REF) and gating on or off the pumping clock. Parasitic metal-to-metal capacitances are utilized for the pumping capacitors (C_M).

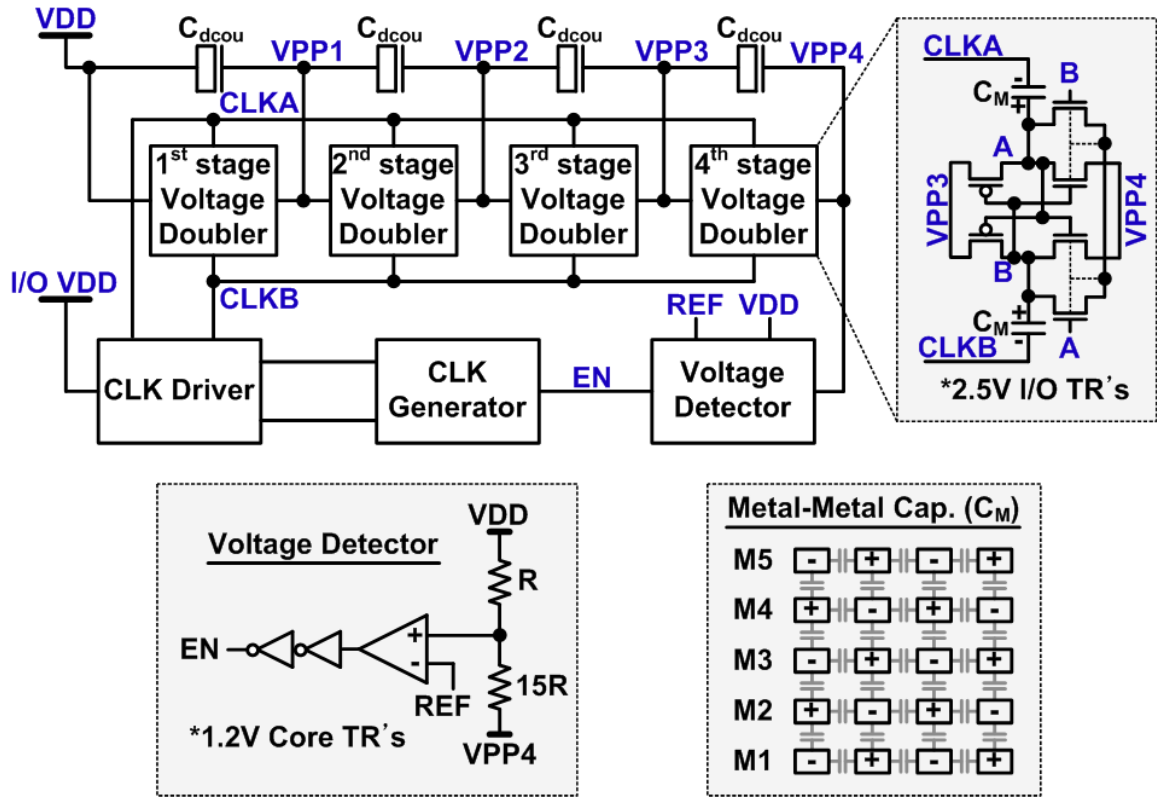


Fig. 4.7 A negative charge pump generating multiple boosted negative voltage levels (VPP1-VPP4) is implemented in a 65nm standard logic process by cascading four voltage doubler stages.

4.3.3 Junction Breakdown Issue of the Designed Negative HVS and CP

Fig. 4.8 illustrates the junction breakdown issue in the proposed negative HVS and CP. As the body of the PMOS devices in the HVS bottom latch and the CP final stage are

connected to VDD, junction breakdown in these devices limits the maximum negative boosted VPP4 level. Note that the previous HVS and CP designs illustrated in chapter 2 [55] do not suffer from junction breakdown issues, as the body of the NMOS devices in the HVS top latch and the CP final stage is connected to a high voltage (e.g. 5V). Using this configuration the junction reverse bias is limited to roughly half the breakdown voltage (e.g. 10V).

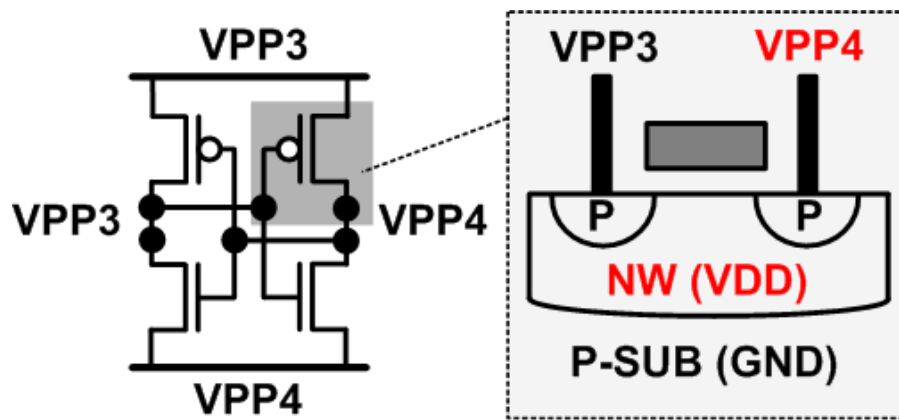


Fig. 4.8 Illustration of the junction breakdown issue in the proposed negative HVS and CP. Junction breakdown of the PMOS devices in the HVS bottom latch and the CP final stage limits the maximum negative VPP4 level.

4.4 Test Chip Measurement Results

A 4kb eflash test macro was implemented in a 65nm low power standard CMOS logic process to demonstrate the proposed circuit ideas. Fig. 4.9 shows the boosted negative voltages (VPP1-4) and the output characteristic measured from the fabricated negative CP. A reliable output voltage was demonstrated for load currents level above the typical operating range ($<1\mu\text{A}$). Fig. 4.10 shows the measured waveforms of the CP output

(VPP4) and HVS output (WWL and PWL) signals for a bit-by-bit write '0' triggered by the WLS pulse.

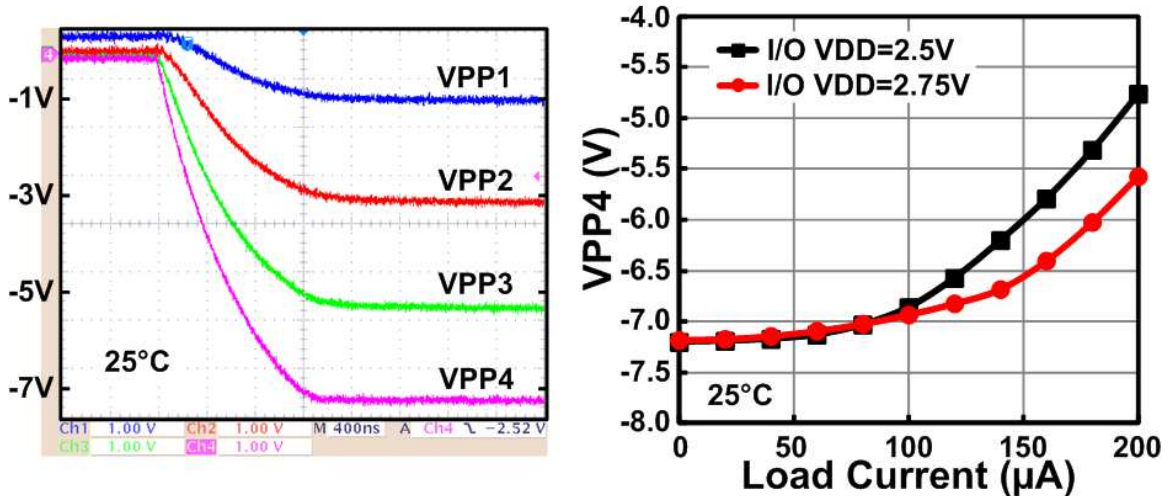


Fig. 4.9 Measured boosted negative voltages (VPP1-VPP4) and output characteristics of the negative charge pump.

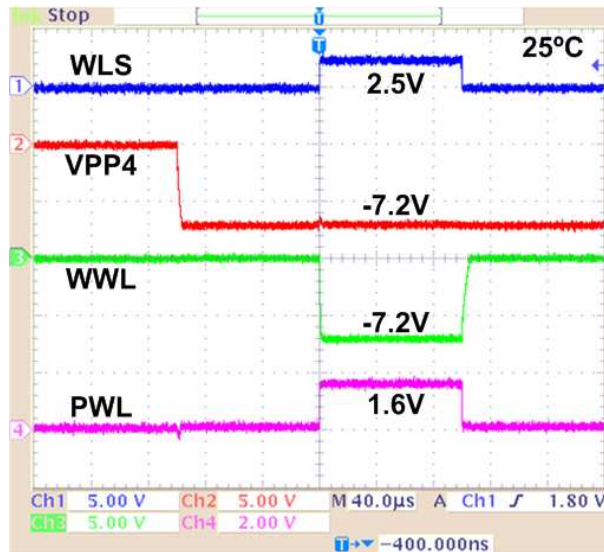


Fig. 4.10 Measured waveforms of the charge pump and high voltage switch.

Fig. 4.11 shows the measured bit-by-bit update result from pattern (0101) to (1100). In this test, the 4 bit data pattern is repeated for the entire WL. Initially, pattern (0101) is

stored in the WL. Then, the cells connected to BL $4n+3$ ($n=0,1,2,\dots$) are updated from '1' to '0' after a write '0' phase. Next, cells connects to BL $4n$ ($n=0,1,2,\dots$) are updated from '0' to '1' upon a write '1' phase. The corresponding cell threshold voltage distributions of each BL group are shown in the figure. The bit-by-bit update from (0101) pattern to (1100) pattern is therefore achieved without any cells being unnecessarily erased.

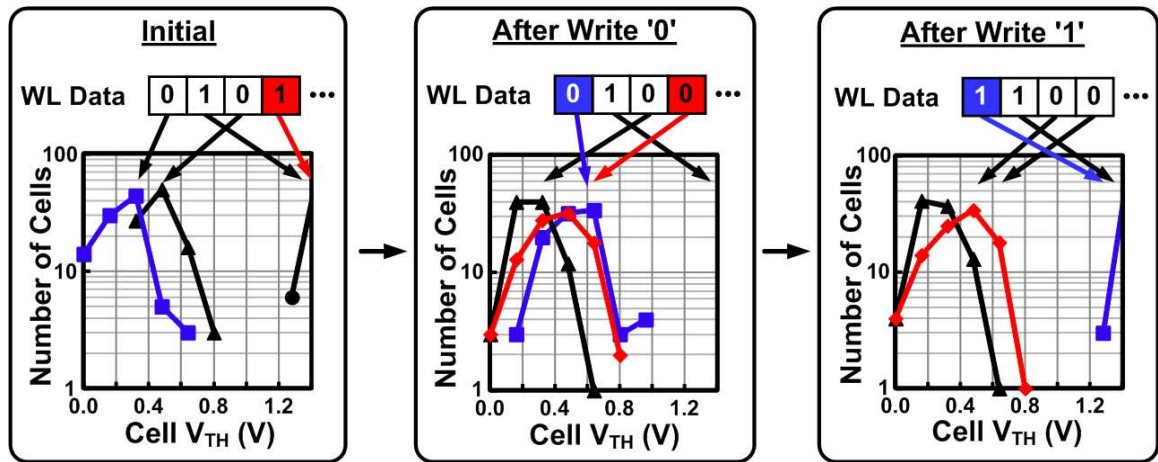
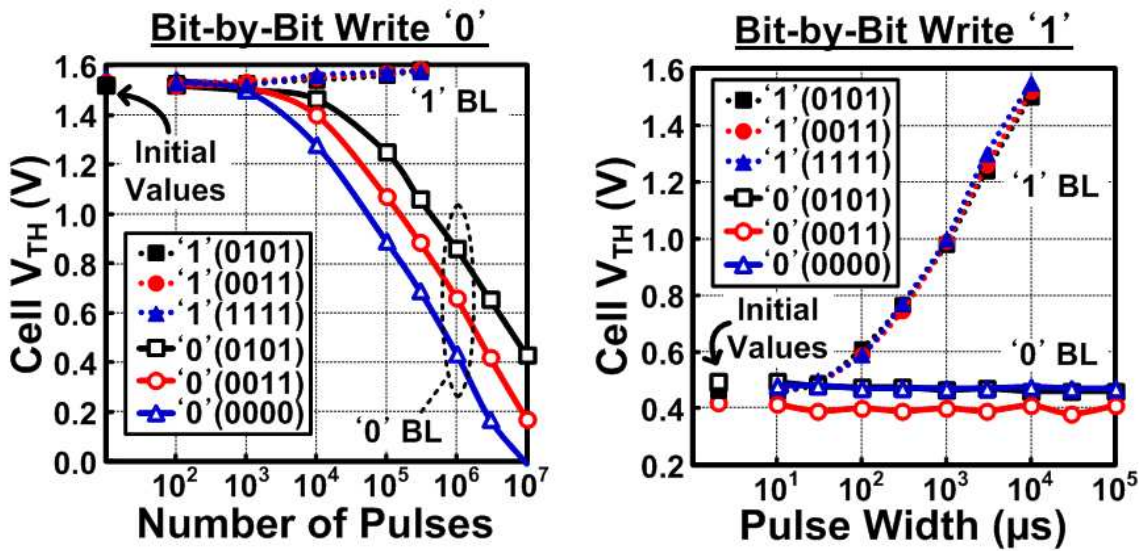


Fig. 4.11 Measured bit-by-bit update result from pattern (0101) to pattern (1100).

Fig. 4.12 (top) shows the measured bit-by-bit write and disturbance results of the proposed 6T eflash. We measure the cell V_{TH} indirectly by simultaneously sweeping the WWL and PWL voltage levels while checking whether the sensed data has flipped. The '0' BL cells show a larger shift in threshold voltage after consecutive write '0' pulses. The disturbance of '1' BL cells increase the signal margin between the '0' and '1' BL cells. The tested write patterns in this measurement are shown in Fig. 4.12 (bottom). Each test pattern induced a different amount of the coupling between the SN nodes of cells located on adjacent BLs (refer to Fig. 4.3). The measured result shows that the (0101)

pattern has the slowest write ‘0’ speed, as the inter SN node coupling capacitance (C_{ISN}) is largest for this pattern. A higher C_{ISN} reduces the FG boosting effect for ‘0’ BL cells, slowing down the write ‘0’ speed. Similar to the write ‘0’ case, only the ‘1’ BL cells show increased threshold voltages after write ‘1’ pulses. However, disturbance of ‘0’ BL cells was not clearly observed. Dependence of ‘1’ BL cell write speed on the data pattern was not apparent either.



Pattern Name	BL #					C_{ISN}
	0	1	2	3	...	
(1111)	1	1	1	1	...	0
(0101)	0	1	0	1	...	++
(0011)	0	0	1	1	...	+
(0000)	0	0	0	0	...	0

Fig. 4.12 (top) Measured cell V_{TH} shift from the 6T eflash test chip. Note that multiple write pulses with a fixed pulse width of $10\mu s$ were applied for the bit-by-bit write ‘0’, whereas a single write pulse was applied for the bit-by-bit write ‘1’. (bottom) Different test patterns give different coupling between adjacent cells.

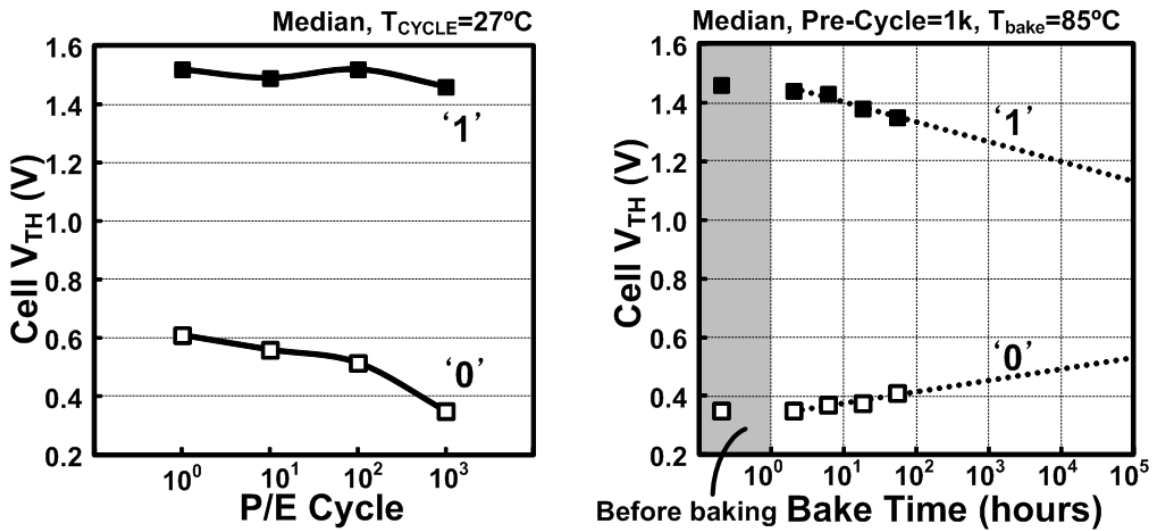


Fig. 4.13 Measured cell endurance and retention characteristics.

Fig. 4.13 shows the measured cell endurance and retention characteristic. Cycling was performed at room temperature (27°C) and all cells in the selected WL experienced the same write '0' and '1' pulses during cycling. The measured data confirms that the median cells with 1k pre-cycles meet a 1 year retention time at 85°C maintaining a cell V_{TH} margin of ~0.7V. To estimate the overall endurance improvement of the proposed 6T eflash compared to the prior 5T eflash discussed in chapter 2 [54, 55], the average number of stress cycles for each state transition was compared in Fig. 4.14 (top) based on the measured results. The cell V_{TH} transition plot in Fig. 4.14 (bottom) shows an example for a high-to-high transition. The prior WL-by-WL erasable 5T eflash undergoes two stress cycles while the new 6T eflash experiences relatively insignificant cell V_{TH} shift. Based on the information listed in Fig. 4.14 for the various transition cases, we can conclude that the overall cell V_{TH} shifts of the proposed 6T eflash is roughly half compared to that of the prior 5T eflash (no column multiplexing case). Half the number

of stress cycles in the proposed 6T eflash implies roughly half the number of traps generated, enhancing the overall endurance limit by twice compared to the prior 5T eflash. The total number of stress cycles can be reduced further with a higher column multiplexing ratio as data stored in the unselected BL's remain unchanged. Based on these observations, the overall endurance is shown in Fig. 4.15 for different eflash configurations assuming a random data pattern and random addressing. When the proposed 6T eflash is used with a 2:1 column MUX, the overall endurance can be improved by around 4 times compared to the previous 5T eflash. Finally, Fig. 4.16 shows the die photograph of the fabricated 4kb eflash test chip.

State Transition	BL 0				BL 1				Average
	L→L	L→H	H→L	H→H	L→L	L→H	H→L	H→H	
5T Eflash (No CMUX)	0X	1X	1X	2X	0X	1X	1X	2X	$8X/8 = 1X$
6T Eflash (No CMUX)	0X	1X	1X	0X	0X	1X	1X	0X	$4X/8 = 0.5X$
6T Eflash (2:1 CMUX)	0X	1X	1X	0X	0X	0X	0X	0X	$2X/8 = 0.25X$

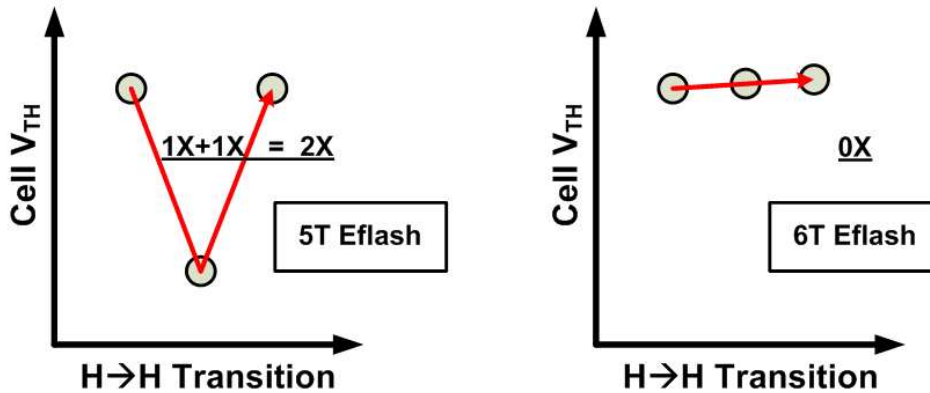


Fig. 4.14 (top) Average number of stress cycles for different data transitions and (bottom) cell V_{TH} transition plot for a high-to-high transition in the 5T eflash discussed in chapter 2 [54, 55] and the proposed 6T eflash cells.

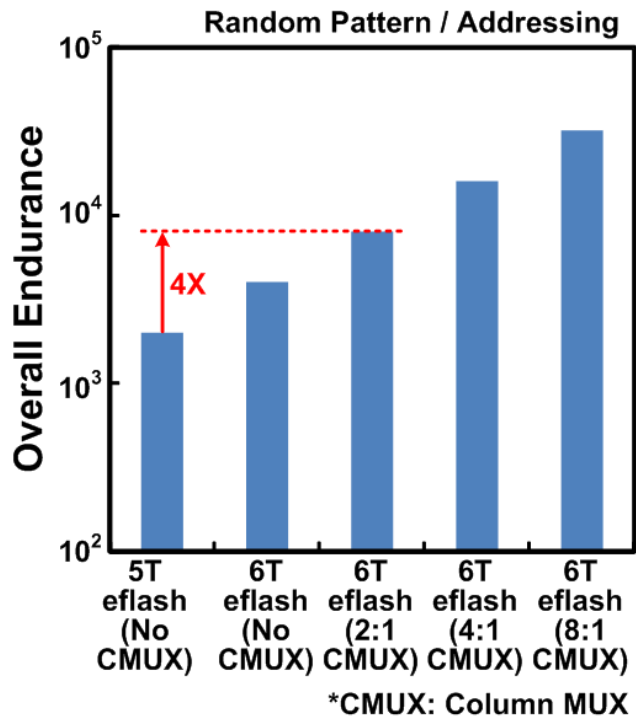


Fig. 4.15 Overall endurance estimated based on the average stress cycle count in Fig. 4.14. A smaller word size (i.e. larger column multiplexing ratio) improves the overall endurance for the proposed 6T eflash.

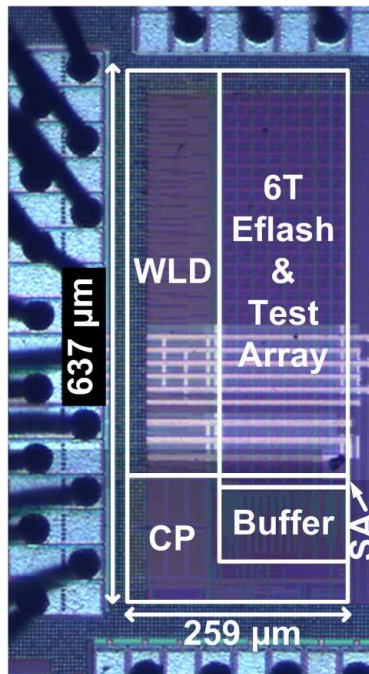


Fig. 4.16 Die photograph of 4kb eflash test chip implemented in a 65nm generic logic process.

4.5 Comparison with Other CMOS Logic Embedded NVM Options

Table 4.1 and 4.2 compare various eNVMS implemented in a standard logic process for moderate density eNVM applications. Kulkarni *et al.* presented a 4kb 1T1R e-fuse and a 1kb 2T anti-fuse OTP in a 32nm logic process using 1.8V I/O transistor [17, 18]. Permanent metal electro-migration and gate-oxide breakdown were used as the program method. Matsufuji *et al.* presented an 8kb 5T anti-fuse OTP memory in 65nm logic technology using 3.3V I/O transistors. The unique feature of this design is that the broken path can be tested using a 5T cell structure [19]. Such e-fuse and anti-fuse designs, however, cannot be written more than once. On the other hand, various single-poly eflash memories capable of multiple write operations were presented in [29-40]. Feng *et al.* proposed a bit-by-bit re-writable 192b 10T eflash in a 0.18 μm logic process using 3.3V I/O transistors [36], but this cell structure suffers from the disturbance issue of the unselected WL cells. Chen *et al.* proposed a 3T eflash which can be built in an advanced logic process [38], and Roizin *et al.* proposed a 256b C-Flash in a 0.18 μm logic process using 3.3V I/O transistor using a bipolar writing voltage (i.e. 5, -5V) [39, 40]; however, these designs do not support a bit-by-bit re-write and the unselected WL cells suffer from disturbance issues. In chapter 2 [54, 55], a 2kb 5T eflash was implemented in a 65nm logic process using 2.5V I/O transistor. Here, the unselected WL's are undisturbed; however, it is not capable of a bit-by-bit write, which increases the number of unnecessary stress cycles. Compared to all the prior work, the proposed 6T eflash is the only bit-by-bit re-writable eNVM that eliminates disturbance in the unselected WL cells.

Table 4.1 Logic Compatible Embedded NVM Comparison (OTP)

CMOS Logic eNVM	1T1R E-Fuse [17]	2T Anti-Fuse [18]	5T Anti-Fuse [19]	2T Anti-Fuse [20]
Process	32nm	32nm	65nm	0.18 μ m
Supply Voltage	1.0V	1.0V	1.2V	1.8V
Acc. / Cell Dev.	1.8V I/O TR	1.8V I/O TR	3.3V I/O TR	3.3V I/O TR
Tunnel Oxide	None	None	None	None
Writing Method	Electro-Migration	Gate Oxide Breakdown	Gate Oxide Breakdown	Gate Oxide Breakdown
Writing Voltage	1.9V	4.5V	6.5V	6.6V
Bit-by-Bit Rewrite	No	No	No	No
Unsel. WL Disturb	No	No	No	No
Unit Cell Area	1.37 μ m ²	1.01 μ m ²	15.3 μ m ²	4.88 μ m ²
CP Area	N. A.	N. A.	0.0512mm ² (est.)	N. A.
Macro Area	N. A.	N. A.	0.244mm ²	0.133mm ²
Capacity	4kb	1kb	8kb	2kb

Table 4.2 Logic Compatible Embedded NVM Comparison (Eflash)

CMOS Logic eNVM	10T Eflash [36]	3T Eflash [38]	C-Flash [40]	5T Eflash [54, 55]	This Work [57]
Process	0.18 μ m	65nm	0.18 μ m	65nm	65nm
Supply Voltage	1.2V	1.2V	1.8V	1.2V	1.2V
Acc. / Cell Dev.	3.3V I/O TR	2.5V I/O TR	3.3V I/O TR	2.5V I/O TR	2.5V I/O TR
Tunnel Oxide	7nm	5nm	7nm	5nm	5nm
Writing Method	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Writing Voltage	10V	8V	5, -5V	10V	-7.2V
Bit-by-Bit Rewrite	Yes	No	No	No	Yes
Unsel. WL Disturb	Yes	Yes	Yes	No	No
Unit Cell Area	220 μ m ²	N. A.	72 μ m ²	8.62 μ m ²	15.3 μ m ²
CP Area	N. A.	N. A.	N. A.	N. A.	0.0214mm ²
Macro Area	N. A.	N. A.	0.0336mm ²	0.0859mm ²	0.165mm ²
Capacity	192b	N. A.	256b	2kb	4kb

4.6 Chapter Summary

Single-poly eflash memory is ideally suitable for moderate density eNVM applications, as it can be built using standard I/O devices readily available in a generic logic process. The previous WL-by-WL erasable eflash designed by our group suffers from unnecessary erase and program cycles resulting in poor endurance characteristics. Previous bit-by-bit erasable eflash on the other hand suffered from high voltage disturbance issues in the unselected WL's. In this work, we proposed a bit-by-bit rewritable 6T eflash which can prevent disturbance issues in the unselected WL's. This was accomplished by a novel bit-by-bit FG boosting scheme. A negative HVS and an on-chip voltage doubler based CP were designed to provide the appropriate WL voltage levels. A 4kb eflash test chip was demonstrated in a generic 65nm logic process, confirming the functionality of the proposed techniques. The overall endurance was improved by ~4 times compared to the prior WL-by-WL erasable 5T eflash for a 2:1 column MUX configuration.

Chapter 5 10T Differential Eflash Featuring Multi-Configurable High Voltage Switch with No Boosted Read Supplies

Moderate density eNVMS have been adopted widely in many digital and analog building blocks such as processor, SRAM, display driver IC, RF-ID tag, mixed signal circuit, and wireless sensor to deal with the circuit variability issues for yield improvement [17-25]. Moreover, they are expected to perform a critical role in the future VLSI technologies to solve the circuit aging issues by storing critical data on them for the run-time calibration. Logic compatible single-poly eflash memories supporting multiple-time program operation, therefore, have been investigated with great interest [26-40, 54-57]. Among them, the logic compatible 5T and 6T eflash memories discussed in prior chapters uniquely provided the overstress-free multi-story HVS circuits [54-57], while completely removing the high voltage disturbance issues of the unselected WL cells.

These prior eflash designs, however, have the issues such as non-reconfigurable HVS requiring boosted read supplies and limited retention time.

In this chapter, we present 10T differential eflash featuring the multi-configurable HVS with no boosted read supplies for lower read mode power, faster wake-up and enhanced retention time, making the proposed eflash being a competitive moderate density eNVM candidate in a standard CMOS logic process.

5.1 Issues in 5T and 6T Eflash Memory Designs

Fig. 5.1 illustrates the issues in 5T and 6T eflash memory designs discussed in the prior chapters. They adopt the non-reconfigurable HVS requiring boosted supplies (i.e. VPP1 to VPP4) such that the highest boosted level (VPP4) becomes around 3 to 4 times the nominal I/O supply voltage during read mode as well as during write mode [54-57], whereas the e-fuse, anti-fuse, and many other single poly eflash designs [29, 40] typically do not require boosted supplies during read mode. As a result, 5T and 6T eflash memories discussed in chapter 2-4 consume higher read power than other moderate density eNVM not requiring the boosted read supplies. Moreover, they have slower wake-up speed, since the boosted supplies need to be stabilized prior to the read operations.

Another issue in 5T and 6T eflash designs discussed in the prior chapters is the limited retention time, since the optimal read reference voltage (VRD) level is not well defined for the wide range of eflash cell usage conditions such as P/E pre-cycle count and

retention temperature with a single cell architecture having the sensing margin within the minimum difference between the VRD level and P/E states as illustrated in Fig. 5.1.

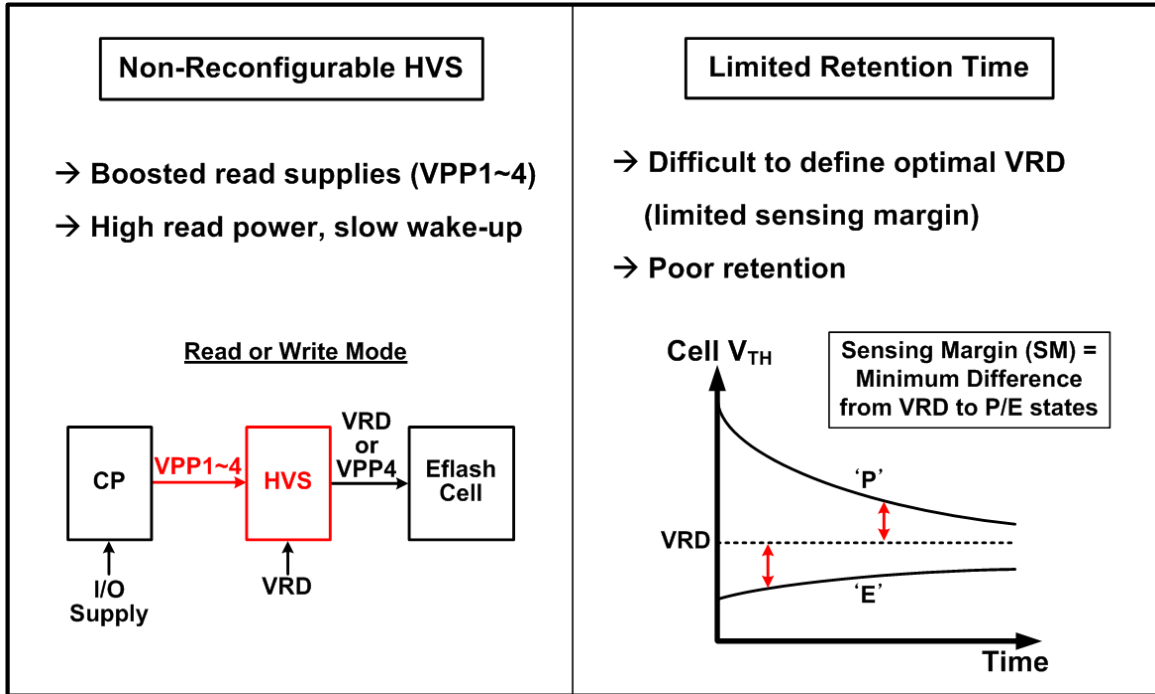


Fig. 5.1 Issues in 5T and 6T eflash memory designs: non-reconfigurable HVS and limited single cell sensing margin.

5.2 Proposed Solutions

5.2.1 Multi-Configurable HVS

Fig. 5.2 compares the non-reconfigurable HVS in prior 5T and 6T eflash designs to the multi-configurable HVS in a proposed eflash design. The non-reconfigurable HVS required boosted supplies (VPP1-VPP4) in order to provide VRD to the eflash cell during read mode as well as during write mode, whereas the multi-configurable HVS with an additional VPP switch does not require the boosted supplies during read mode, but still

can provide the suitable write pulses during write mode. The new eflash design with this multi-configurable HVS therefore operates with no boosted supplies during read and hold mode operations, achieving lower read mode power and faster wake-up compared to 5T and 6T eflash designs discussed in prior chapters. The details of the proposed multi-configurable multi-story HVS, VPP switch, and CP circuits are described in Section 5.3.

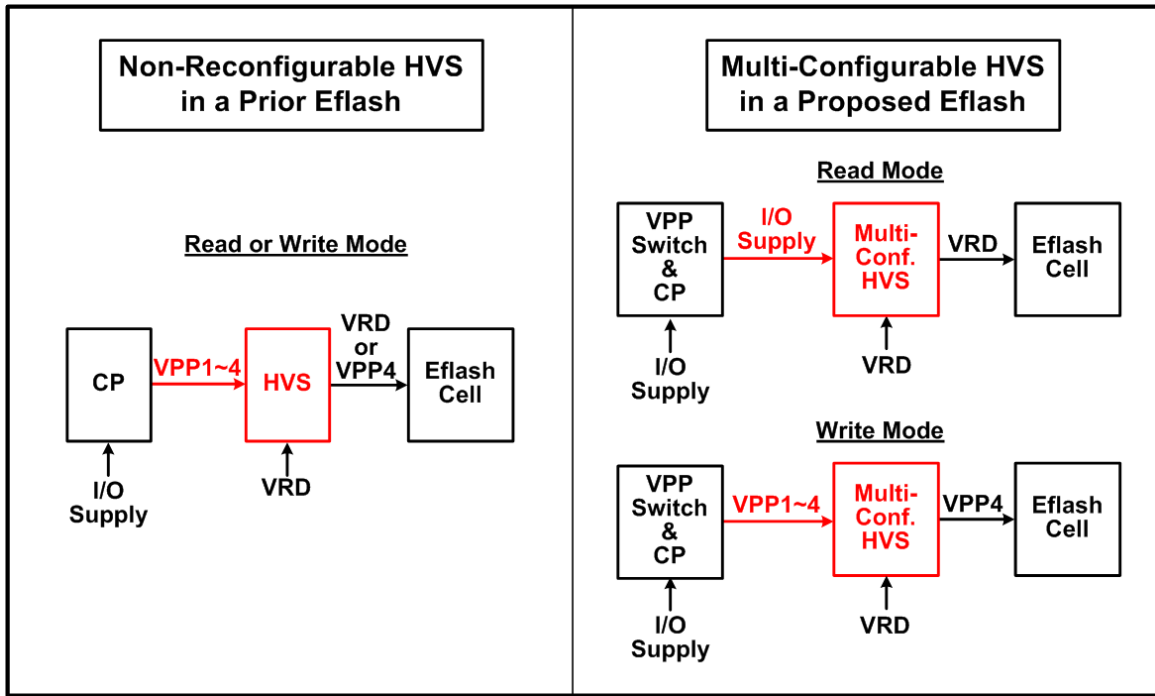


Fig. 5.2 Non-reconfigurable HVS in 5T and 6T eflash designs is compared to the multi-configurable HVS in a proposed eflash.

5.2.2 Differential Cell Architecture

Fig. 5.3 compares the sensing margins of single and differential eflash cell architectures. The single cell architecture has limited sensing margin within the minimum difference between the VRD level and P/E states, since the VRD level is difficult to be optimized for wide range of eflash cell conditions such as P/E pre-cycle count and

retention temperature, whereas the differential cell architecture has larger sensing margin which is defined to be the cell V_{TH} difference between worst 'P' and 'E' states, since the eflash cell usage conditions such as P/E pre-cycle count and retention temperature are common mode signals like VRD and power supplies. Thus, with the differential cell architecture, the retention time is significantly improved compared to the single cell architecture. Indeed, the differential cell architecture was previously reported to have the lower failure rate during retention mode compared to the single cell architecture in [33]. The proposed 10T differential eflash cell and its differential sensing scheme are described in Section 5.3, and the estimated retention improvement by the proposed differential cell architecture is discussed in Section 5.4.

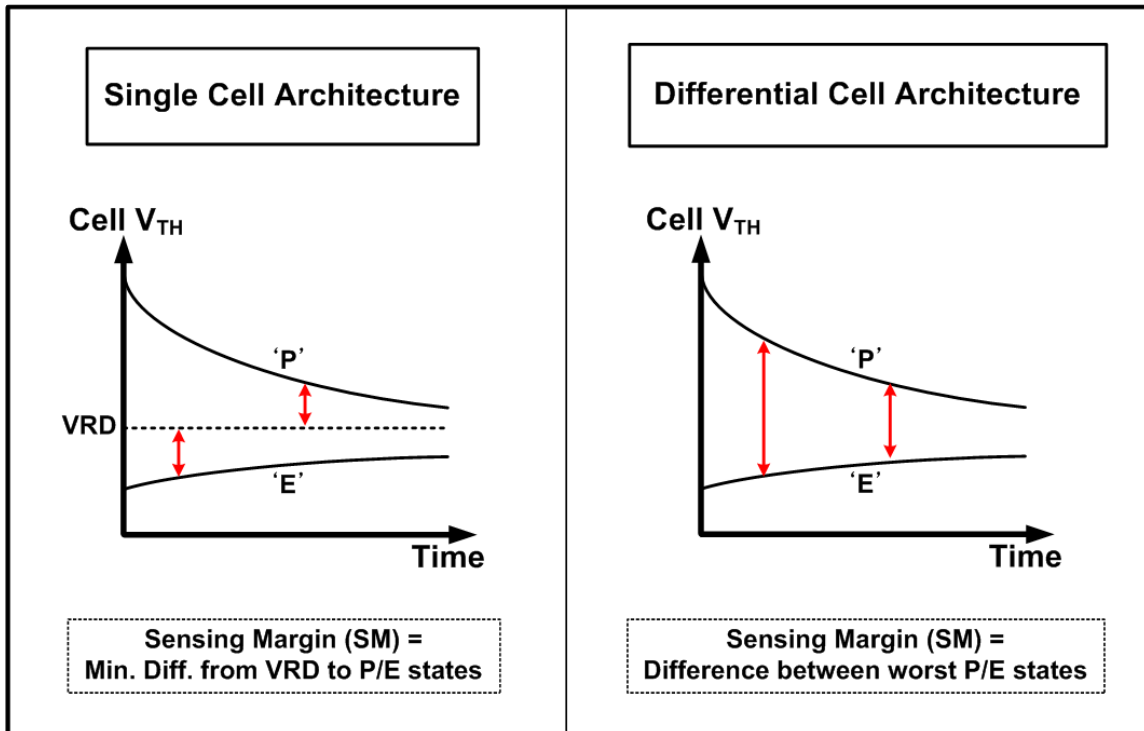


Fig. 5.3 Sensing margin of the single cell architecture is compared to that of the differential cell architecture.

5.3 Proposed 10T Differential Eflash memory

5.3.1 Overall Architecture

Fig. 5.4 shows the overall architecture of the proposed 10T differential eflash memory consisting of 10T differential cell array, multi-configurable high voltage switches, VPP switch and charge pump, differential current sense amplifiers, and other logic blocks. All the building blocks are implemented using standard I/O and core transistors with no overstress voltage causing reliability issues. The supply voltages of the proposed multi-configurable HVS (VPS1~4, VPO1~3) are provided via VPP switch and charge pump. Each multi-configurable HVS provides suitable read and write pulses to Program WL (PWL) and Write WL (WWL) of the eflash cell. The 10T differential cell array consists of 16 row and 112 columns and each row is simultaneously sensed by differential current sense amplifiers implemented using low voltage core transistors for higher read performance. Simple ECC encoder and decoder are included in the column circuitry for higher reliability. The detailed high voltage circuit and 10T differential cell operations are described in the later sections.

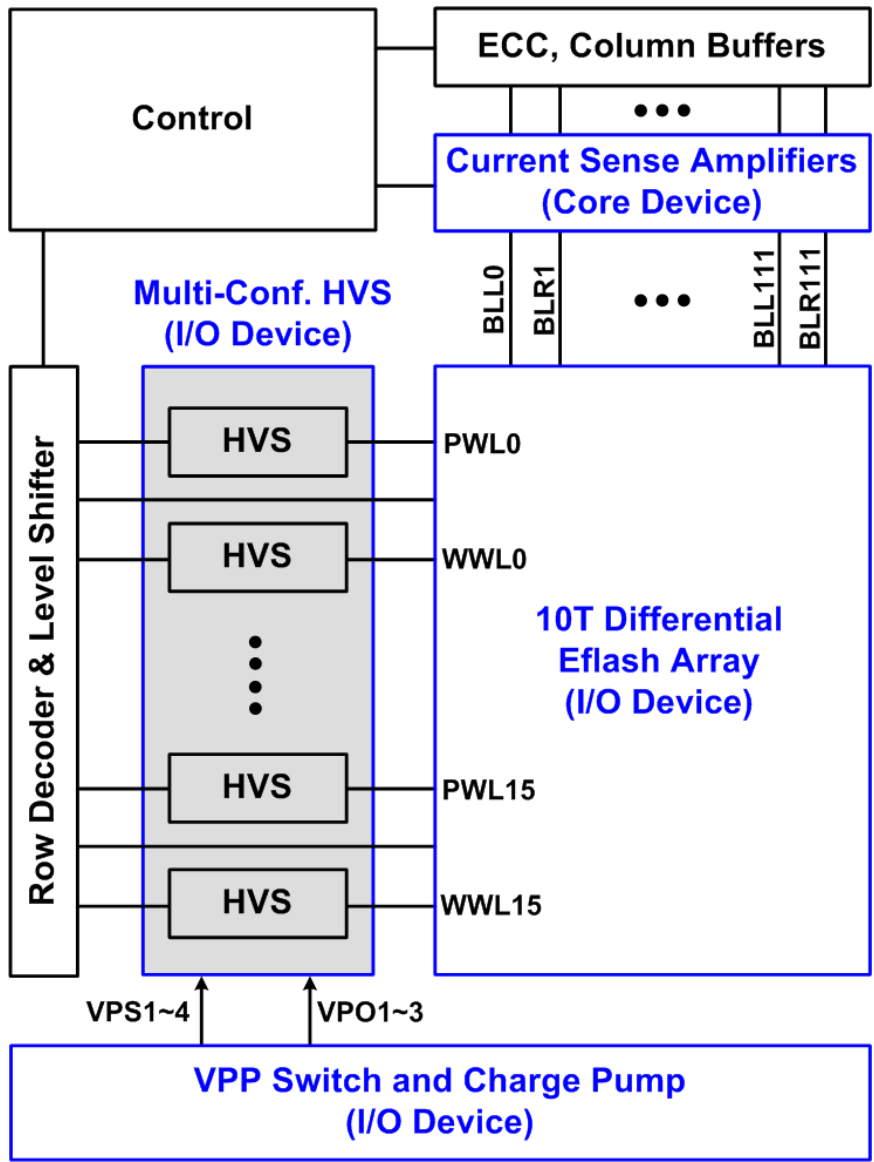


Fig. 5.4 The proposed eflash memory consists of 10T differential cell array, multi-configurable high voltage switches, VPP switch and charge pump, differential current sense amplifiers, and other logic blocks implemented using the standard core and I/O devices in a generic logic process.

5.3.2 Proposed Multi-Configurable High Voltage Switch

The proposed multi-configurable HVS shown in Fig. 5.5 consists of stacked latch stage and $4\times$ I/O supply driver like the non-reconfigurable HVS shown in Fig. 2.6 (bottom) and Fig. 4.6 (a). In the proposed multi-configurable HVS, the top latch is differently placed compared to the non-reconfigurable HVS shown in Fig. 2.6 (bottom). The supplies of the multi-configurable HVS (i.e. VPS1~4 and VPO1~3) are switched for read and write modes. That is, VPS1-VPS4 and VPO1-VPO3 are connected to the I/O supply (VDDE) and VSS, respectively, for read mode, whereas they are connected to the boosted voltages VPP1-VPP4 and VPP1-VPP3 with the highest level VPP4 being 3 to 4 times the nominal I/O voltage, respectively, for write mode. The boosted voltages VPP1, VPP2, and VPP3 levels are designed to be about 0.25, 0.5, 0.75 times the VPP4 level, respectively.

The operation details of the proposed multi-configurable HVS during read and write modes are illustrated in Figs. 5.6 and 5.7, respectively. During read mode, the proposed multi-configurable HVS switches between VSS and VRD depending on SRD signal without changing the latch states for a high speed WL activation. When SRD goes to high, WWL is charged to VRD level through the NMOS string in the final stage. When SRD goes to low, WWL is discharged to VSS level through another NMOS transistor path in the driver stage. During write mode, the proposed multi-configurable HVS operates similarly to the multi-story HVS described in chapter 2 [54, 55] to provide the boosted write voltage (VPP4) pulse to WWL. When SWR1 and SWR2 switch from low to high, nodes A, B, D, and F are discharged to VPP3, VPP2, VPP1, and VSS,

respectively, while nodes C and E are pulled-up to VPP3 and VPP2. Then, node M and WWL are pulled up to VPP3 and VPP4. When SWR1 and SWR2 switch from high to low, the opposite transitions occur, making node M and WWL pulled down to VPP1 and VSS. During read and write operations, all the transistors operate without no over-stress voltage. The simulated waveforms of the proposed multi-configurable high voltage switch operations during read and write modes are shown in Figs. 5.8 and 5.9, respectively. During read operation, no boosted supply is applied to the multi-configurable HVS, while it successfully drives WWL to VRD level. During write operation, on the other hand, the boosted supplies are applied to the multi-configurable HVS via VPP switch and charge pump described in Section 5.3.3, while it drives WWL to VPP4 level without an overstress voltage in HVS.

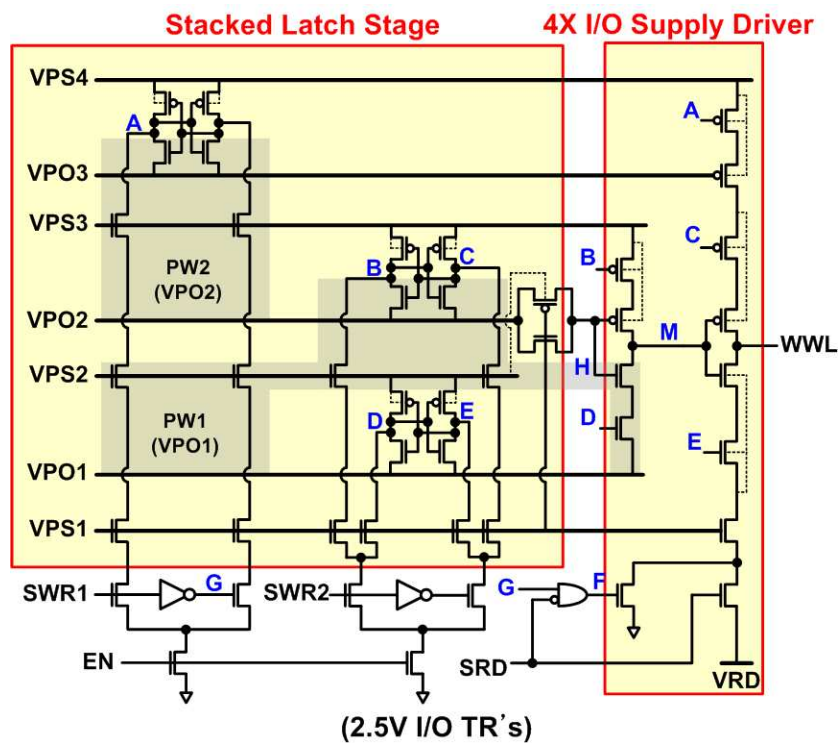


Fig. 5.5 The proposed multi-configurable high voltage switch.

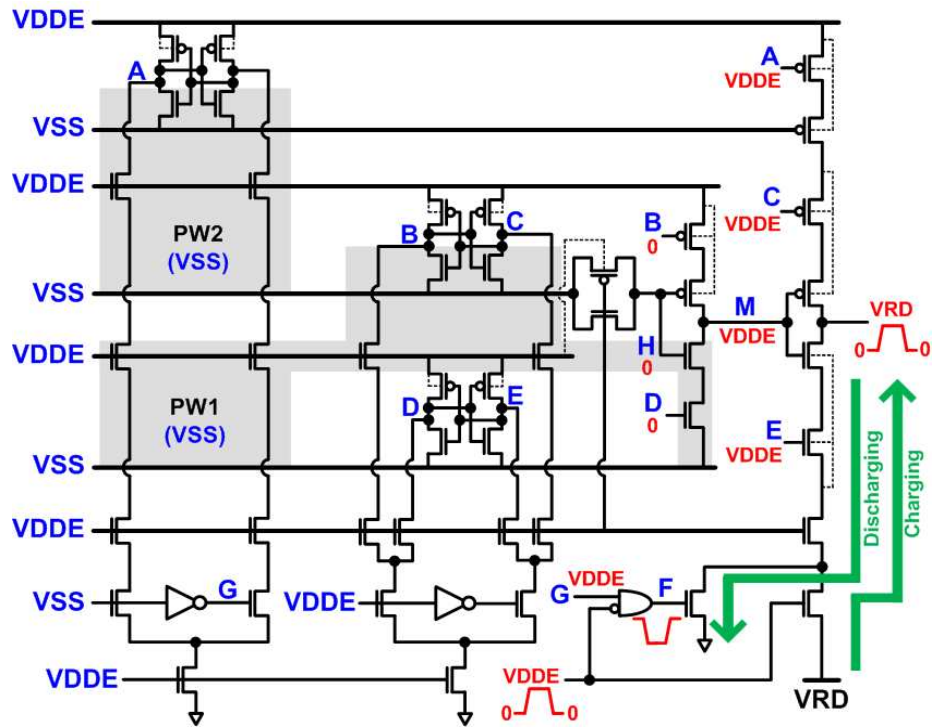


Fig. 5.6 Operation of the proposed multi-configurable HVS during read mode.

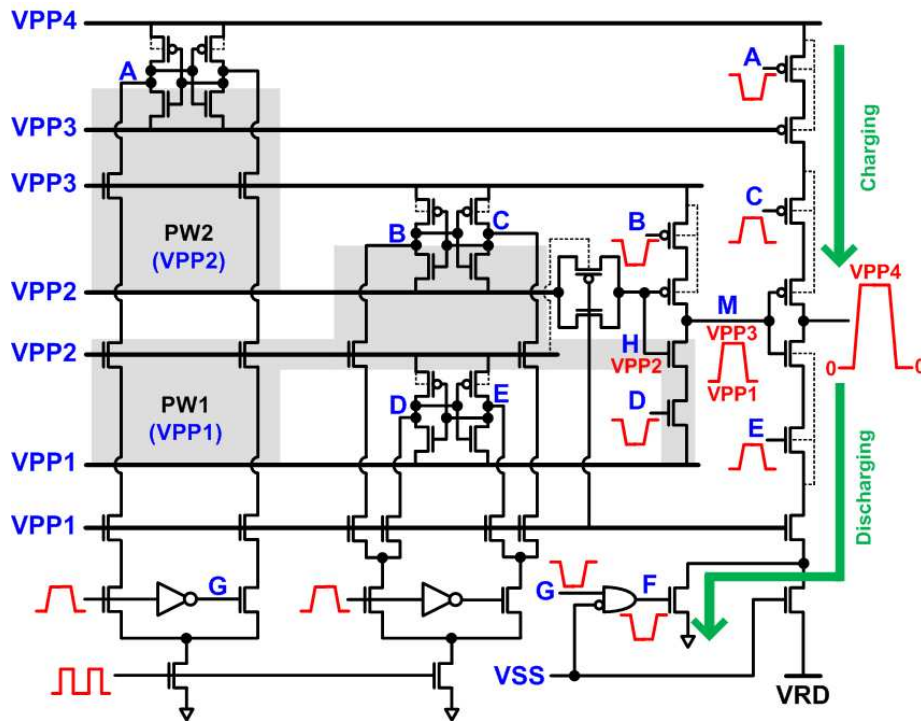
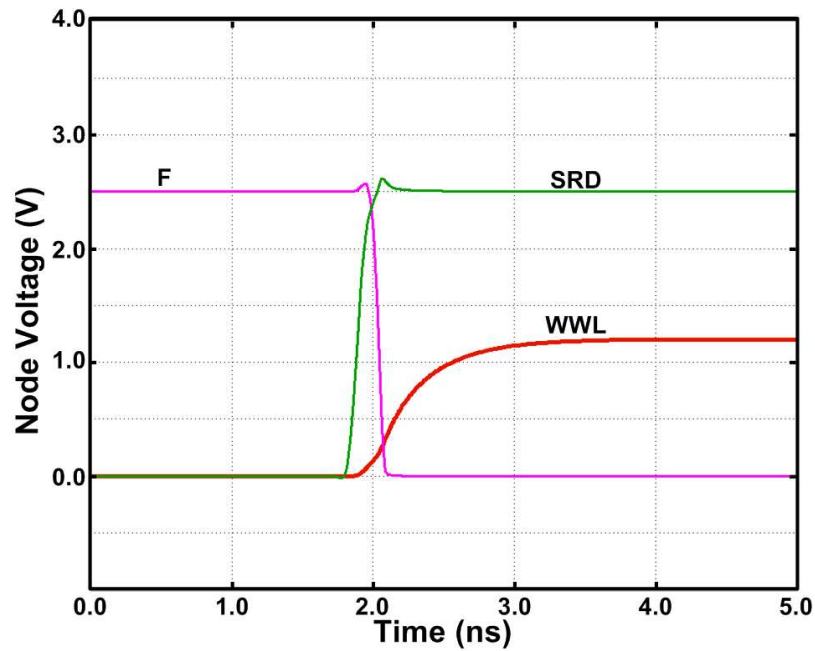
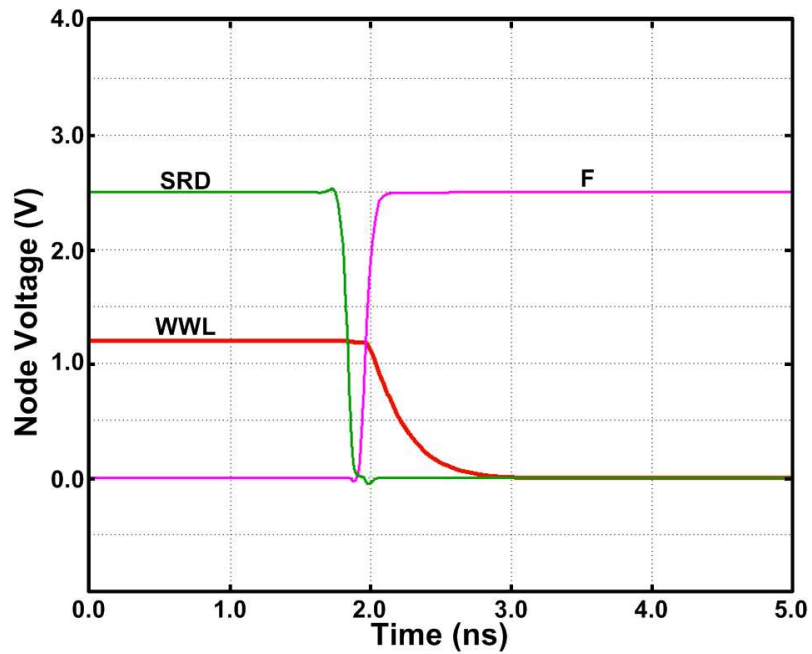


Fig. 5.7 Operation of the proposed multi-configurable HVS during write mode.

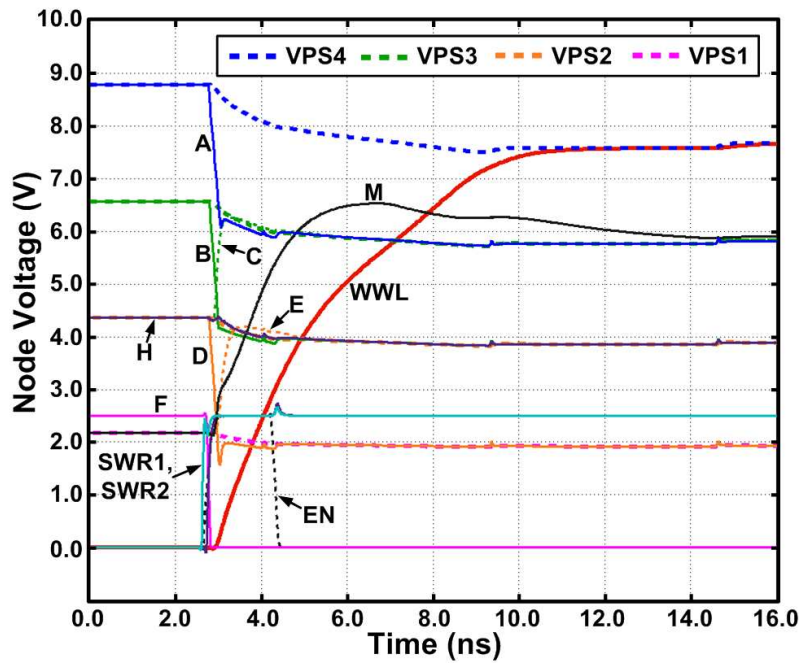


(a)

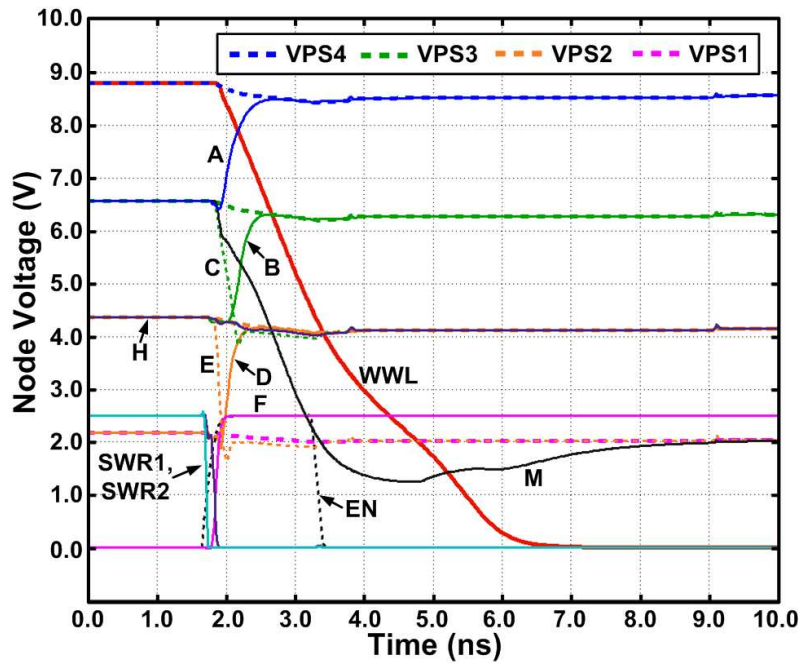


(b)

Fig. 5.8 The simulated waveforms of the proposed multi-configurable high voltage switch operations during read mode ($V_{DDE}=2.5V$, $V_{RD}=1.2V$, No Boosted Supply, $Temp.=25^{\circ}C$): (a) low to high transition of WWL, (b) high to low transition of WWL.



(a)



(b)

Fig. 5.9 The simulated waveforms of the proposed multi-configurable high voltage switch operations during write mode ($V_{DDE}=2.5V$, $V_{RD}=1.2V$, $V_{PP1}=2.2V$, $V_{PP2}=4.4V$, $V_{PP3}=6.6V$, $V_{PP4}=8.8V$, $Temp.=25^{\circ}C$): (a) low to high transition of WWL, (b) high to low transition of WWL.

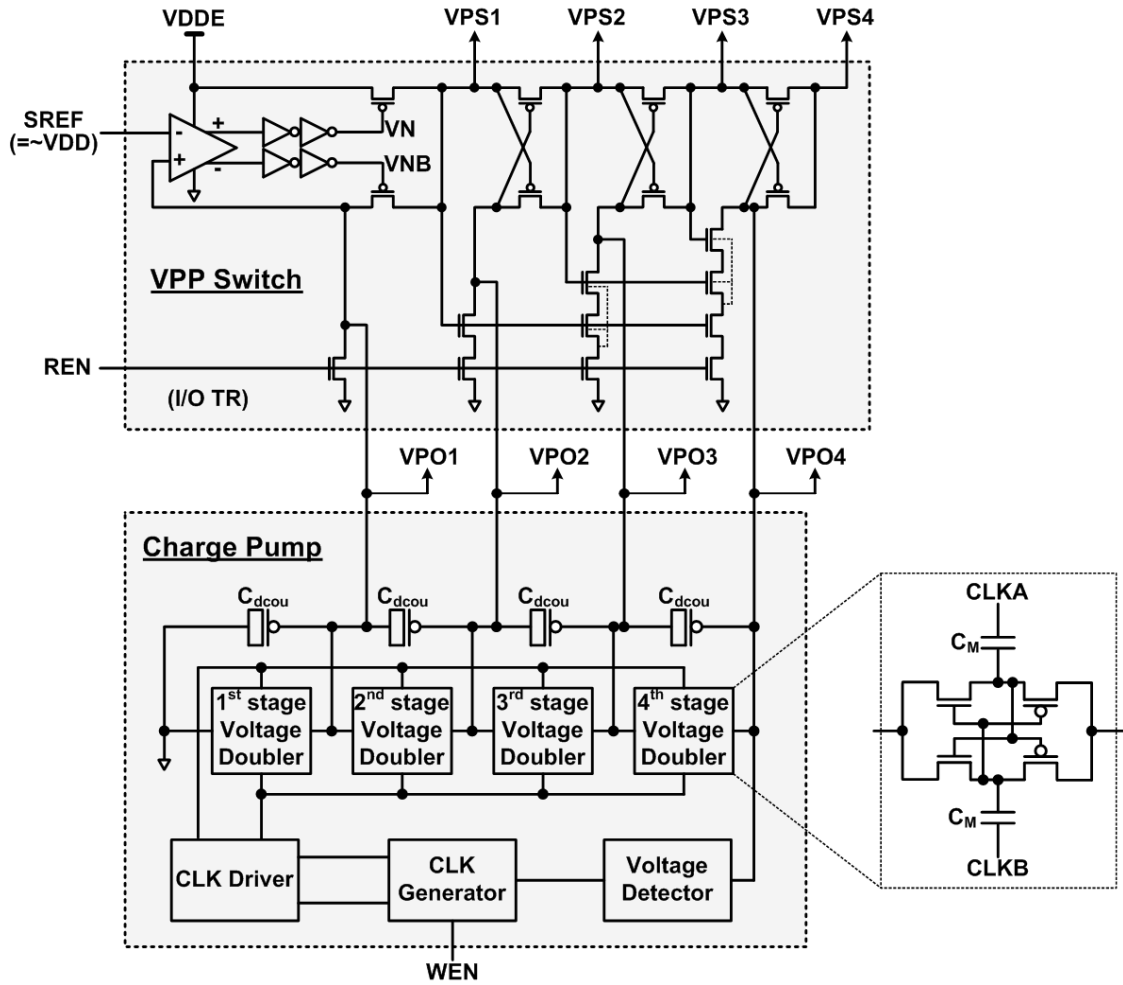


Fig. 5.10 The proposed VPP switch and the voltage doubler based charge pump.

5.3.3 Proposed VPP Switch and Charge Pump

The proposed VPP switch and charge pump circuits are shown in Fig. 5.10. They drive VPS1-VPS4 and VPO1-VPO4 nodes to appropriate voltage levels depending on the operation modes required for the aforementioned multi-configurable HVS. Similarly to the negative charge pump shown in Fig. 4.7, parasitic metal-to-metal capacitances are utilized for the pumping capacitors (C_M). During read mode, the charge pump outputs VPO1-VPO4 are discharged to VSS, whereas during write mode, they are regulated to

the boosted levels VPP1-VPP4 by the voltage detector and clock generator circuits. Then, the proposed VPP switch selectively connects VDDE or VPO1-VPO4 to its outputs VPS1-VPS4 by comparing the VPO1 level against SREF ($=\sim VDD$).

Fig. 5.11 shows the simulated node voltage waveforms of the proposed VPP switch and charge pump which change the operation mode from read to write, and from write to read. The node voltage and transistor operation conditions corresponding to time 0 or 1100ns (i.e. read mode), and time 500 or 1000ns (i.e. write mode) are also illustrated. When WEN and REN signals switch to VDD and VSS respectively, the charge pump starts boosting the VPO1-VPO4 levels up to the appropriate VPP1-VPP4, as illustrated in Fig. 5.11 (a, left). At the moment that VPO1 level exceeds SREF, the comparator in VPP switch toggles the outputs, which in turn connect VN and VNB to VDDE and VSS, respectively, as shown in Fig. 5.11 (b, right). This triggers the boosted charge pump outputs VPO1, VPO2, VPO3, and VPO4 to be sequentially driven to VPS1, VPS2, VPS3, and VPS4. Then, the VPP switch and charge pump turn into write mode configuration. Since the VPO1-VPO4 are boosted-up up to VPP1-VPP4 levels, VPS1-VPS4 are also driven to these VPP1-VPP4 levels during write mode. When WEN and REN signals switch to VSS and VDDE respectively, on the other hand, the charge pump is disabled, and VPO1-VPO4 levels start to be pulled down to VSS as illustrated in Fig. 5.11 (a, right). At the moment that VPO1 level goes below SREF, the comparator in VPP switch toggles the outputs, which in turn connect VN and VNB to VSS and VDDE, respectively, as shown in Fig. 5.11 (b, left). This triggers VDDE to be sequentially driven to VPS1, VPS2, VPS3, and VPS4, discharging VPO1-VPO4 to VSS. Then, the VPP switch and

charge pump turn into read mode configuration. Note that the proposed VPP switch and charge pump do not require any overstress voltage during read and write modes, and the transition periods between the two operation modes.

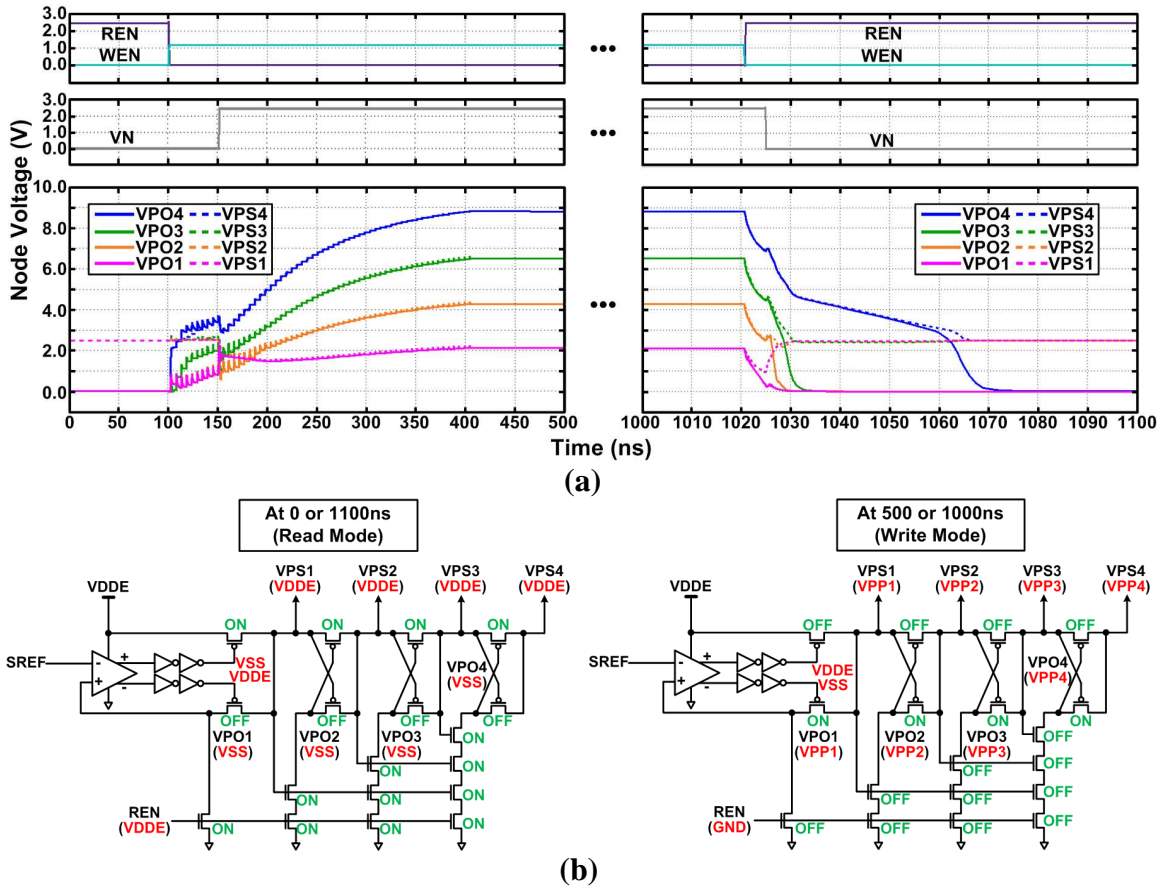


Fig. 5.11 (a) The simulated node voltage waveforms of the proposed VPP switch and charge pump which change the operation mode (left) from read to write, and (right) from write to read ($VDD=1.2V$, $VDDE=2.5V$, $VRD=1.2V$, $VPP1=2.2V$, $VPP2=4.4V$, $VPP3=6.6V$, $VPP4=8.8V$, $Temp.=25^{\circ}C$). (b) The node voltage and transistor operation conditions corresponding to time 0 or 1100ns (i.e. read mode), and time 500 or 1000ns (i.e. write mode).

5.3.4 Operation of the Proposed 10T Differential Eflash Memory Cell

The proposed 10T differential eflash memory cell and its erase/program/read operation bias conditions are shown in Fig. 5.12. The 10T differential eflash cell consists

of a pair of 5T eflash cell where the PMOS coupling transistors (M_1 and M_4), and NMOS erase transistors (M_2 and M_5) are preferred for higher performance and reliability, whereas NMOS program/read transistors (M_3 and M_6) are adopted for self-boostered program method as discussed in chapters 2 and 3 [54-56]. The coupling transistors (M_1 and M_4) are upsized 8 times larger than the erase and program/read transistors (M_2 , M_3 , M_5 , and M_6) for optimized erase and program performance considering area overhead as discussed in chapter 3.

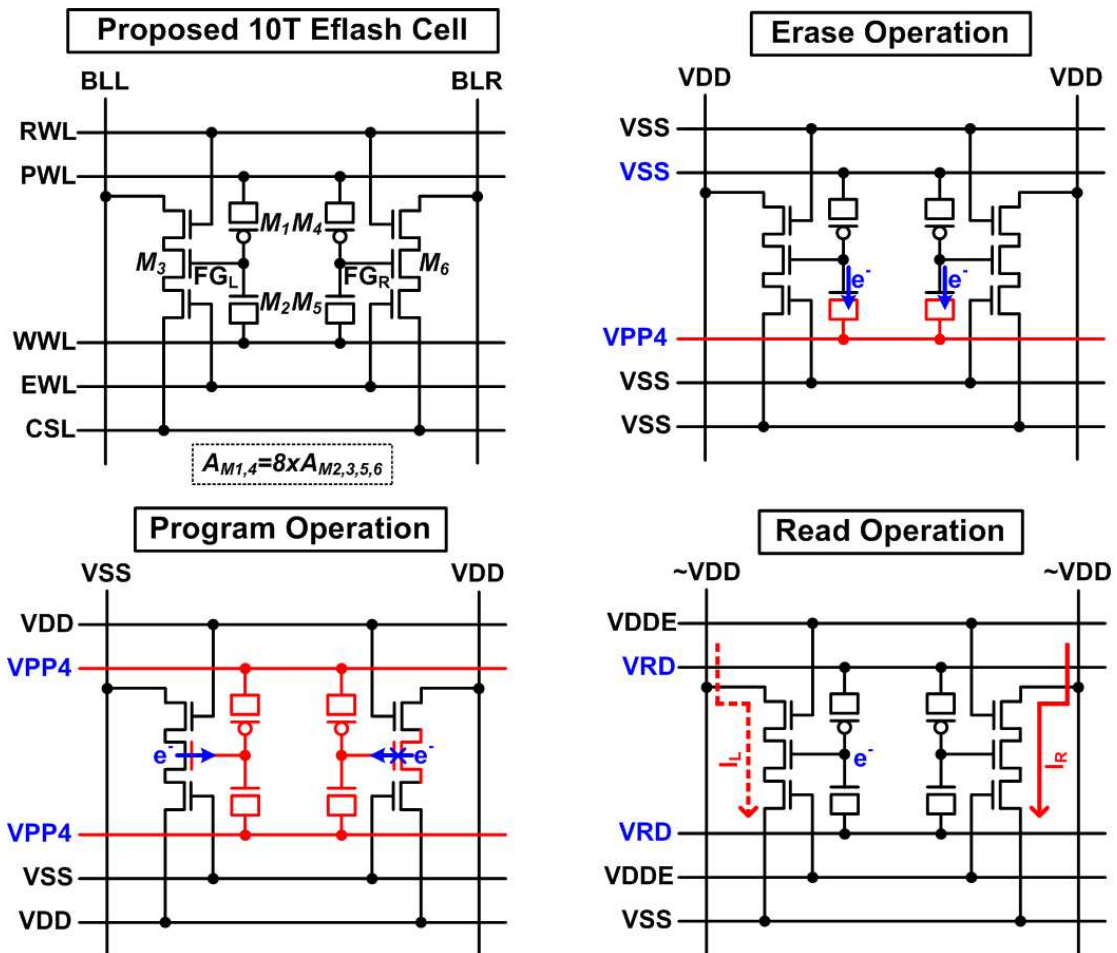


Fig. 5.12 The proposed 10T differential eflash cell, and its erase/program/read operation bias conditions.

During erase operation, a high voltage (V_{PP4}) is applied to WWL, while PWL is biased at VSS. The large coupling from PWL to floating gates (FG_L and FG_R) via the coupling transistors (M_1 and M_4) maintains the FG nodes to be close to VSS level, while the large enough electric fields are generated in the gate oxide of the erase transistors (M_2 and M_5). As a result, FN electron tunneling is enabled through the gate oxide of the erase transistors (M_2 and M_5).

During program operation, the high voltage (V_{PP4}) is applied to both WWL and PWL, while the complementary voltage levels are applied to left BL (BLL) and right BL (BLR), respectively. The half 5T cell connected to VSS enables the FN electron tunneling, while the other half 5T cell connected to VDD inhibits the FN electron tunneling via self-boosting [58, 59] as explained in chapters 2 and 3 [54-56]. Differently from the prior differential cell architecture [29, 33], a high voltage is not applied to BL's during erase and program operations, removing the disturbance issues of the multiple unselected WL's.

During read operation, the read reference voltage (V_{RD}) is applied to both WWL and PWL, while BLL and BLR are maintained close to VDD so that the current difference between the half cells can be detected. The differential current sensing scheme [76] is employed in this work, since it does not discharge BL parasitic capacitance for sensing which is attractive for high speed read access unlike the prior voltage sensing scheme of the 5T eflash discussed in chapter 2 [54, 55]. Fig. 5.13 shows the designed differential current sense amplifier and the read timing diagram of the proposed 10T differential eflash cell. More negatively charged floating gate (i.e. FG_L) results in the lower half cell

current during read operation, which increases the corresponding sense node voltage (i.e. SL) after WL signals are activated. Later, these sense output levels (i.e. SL and SR) turn to digital values when the VSEN signal is activated. Fig. 5.14 shows 1k Monte Carlo run waveforms during read operation of the proposed 10T differential eflash cell with the device mismatch and post-layout parasitic. The shorter than 4ns read latency is achieved from these simulation results.

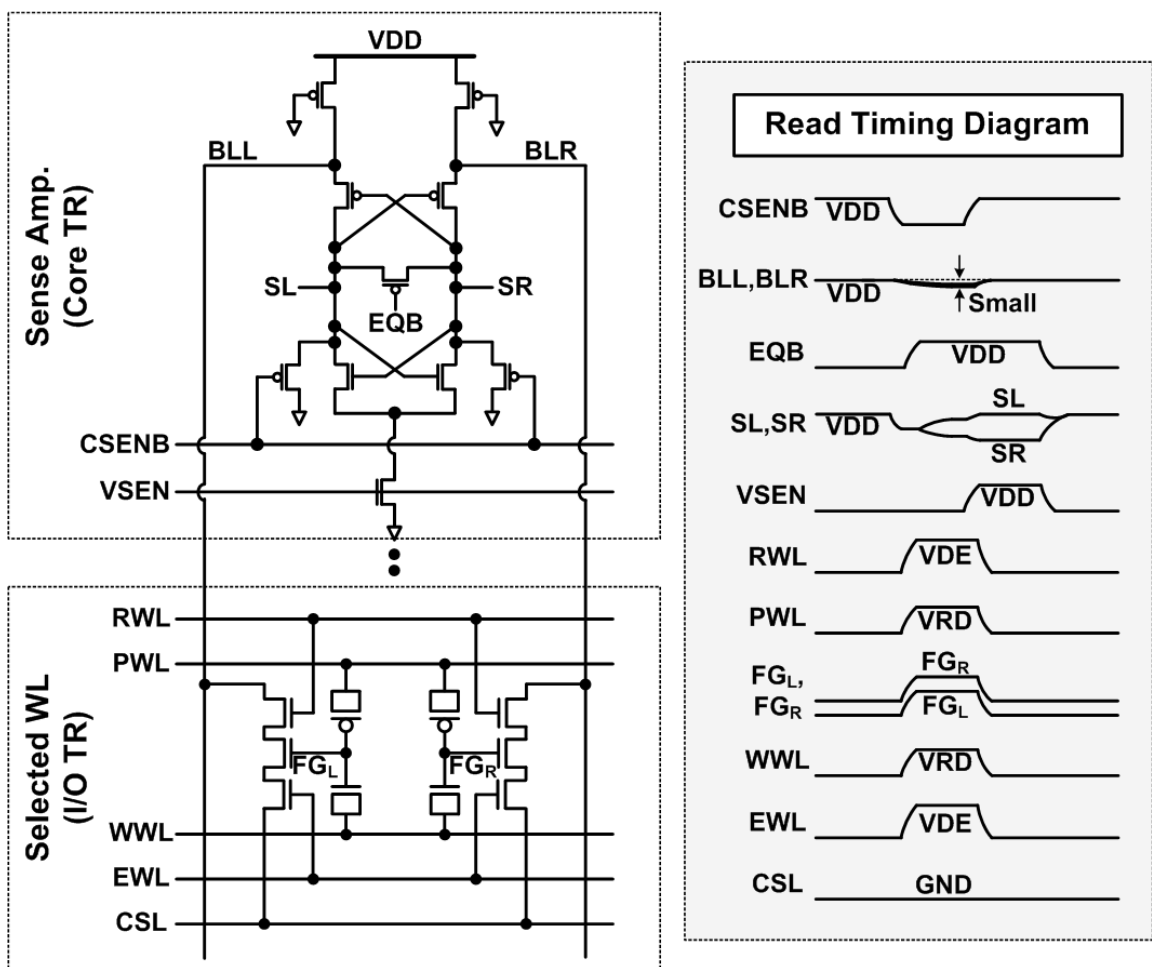


Fig. 5.13 The differential current sense amplifier and the read timing diagram of the proposed 10T differential eflash cell.

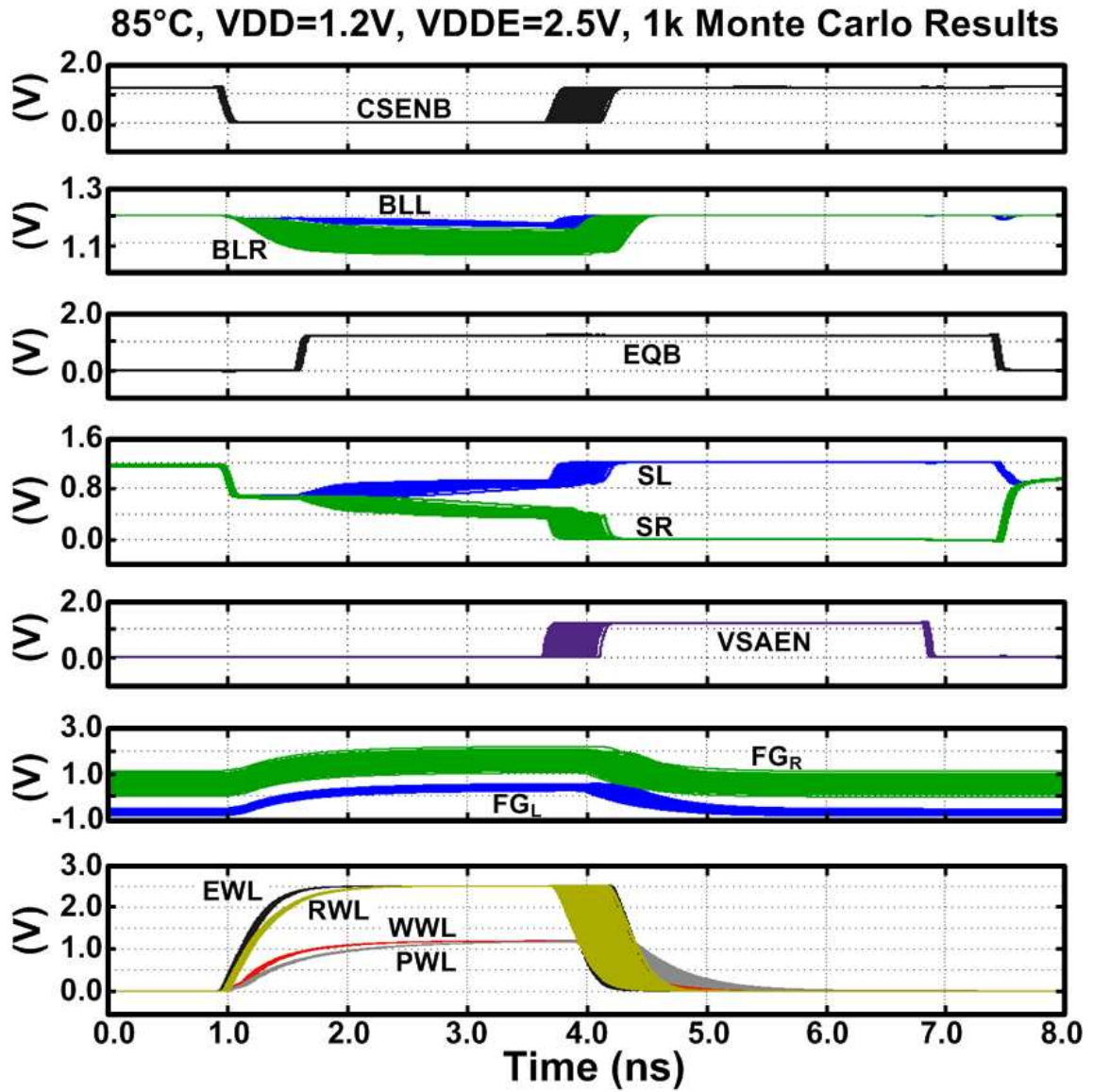


Fig. 5.14 1k Monte Carlo run waveforms during read operation of the proposed 10T differential eflash cell.

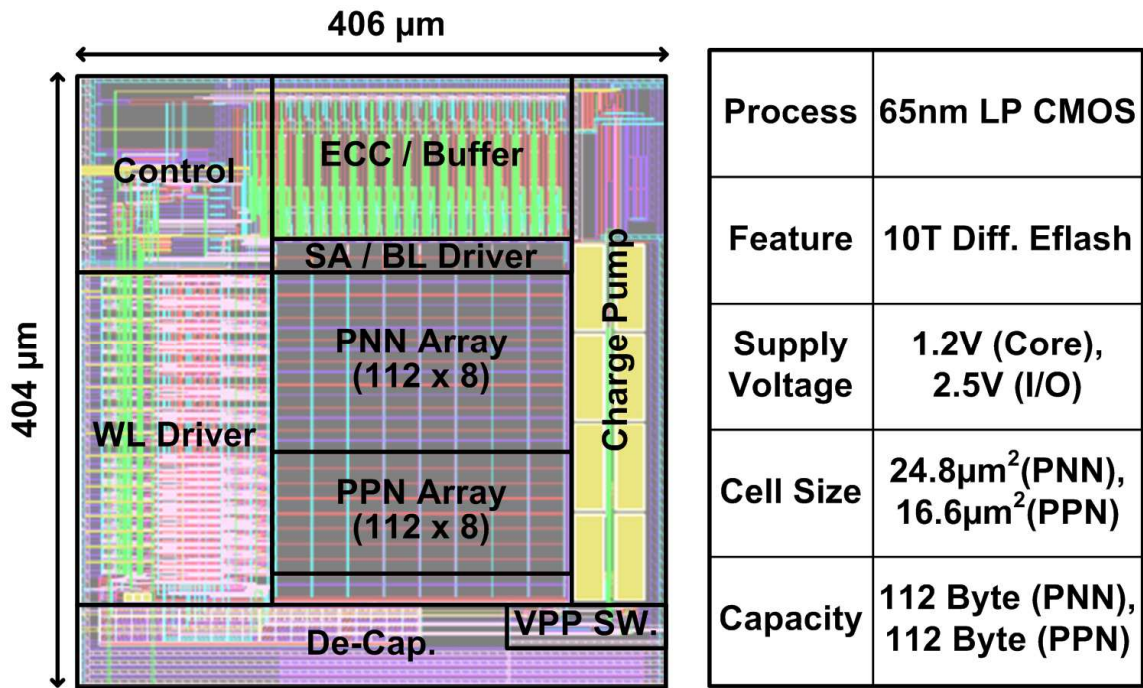


Fig. 5.15 Test chip layout and feature summary.

5.4 Test Chip Design and Discussion

Fig. 5.15 shows the 1792b test chip layout and the feature summary of the proposed 10T differential eflash memory implemented in a 65nm standard logic technology. The cell having NMOS erase TR (i.e. PNN cell) has ~50% cell area overhead compared to the cell having PMOS erase TR (i.e. PPN cell) because of the separation of the deep NW layer.

Fig. 5.16 (left) shows the measured retention characteristic of the 5T eflash cell [54, 55] for different P/E pre-cycle counts. Fig. 5.16 (right) shows the measured single cell sensing margin with VRD of 1.0V and the emulated differential cell sensing margin extracted from the worst case 100, 1k, 10k P/E pre-cycled 5T eflash cells. The

differential cell has more than twice sensing margin during retention time compared to the single cell where the sensing margin is limited by worst case erased and programmed cells. The differential cell is estimated to have greater than 0.6V sensing margin after 10 year retention time for the 10k P/E pre-cycled cells, which is not achievable with the single cell structure.

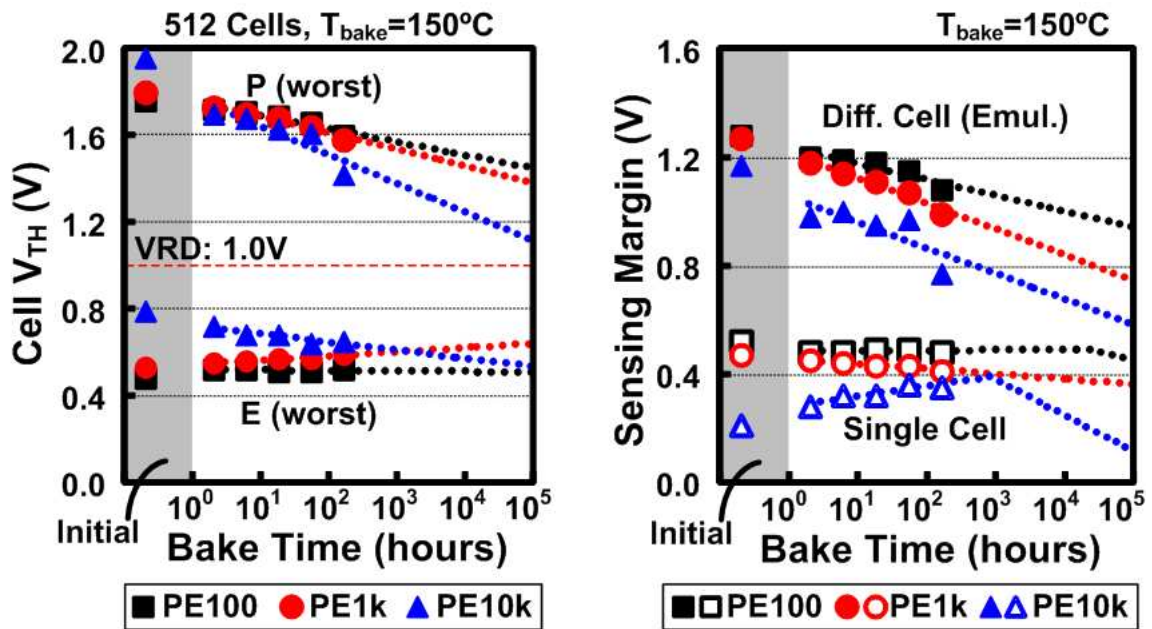


Fig. 5.16 (left) Measured retention characteristic of 5T eflash cell at 150°C. (right) Measured single cell and emulated differential cell sensing margins from the worst case 100, 1k, 10k P/E pre-cycled 5T eflash cells.

Fig. 5.17 illustrates the proposed eflash operations and power consumption during each operation mode. The read power can be minimized with the proposed multi-configurable HVS not requiring the boosted read supplies as discussed in this chapter, though a large amount of the write power is still consumed for generating the boosted write voltage levels. Thus, the proposed eflash is preferred for the applications requesting

the write operation infrequently, while it can achieve zero-standby power during idle periods, which is highly attractive characteristic for many low power system-on-a-chip applications.

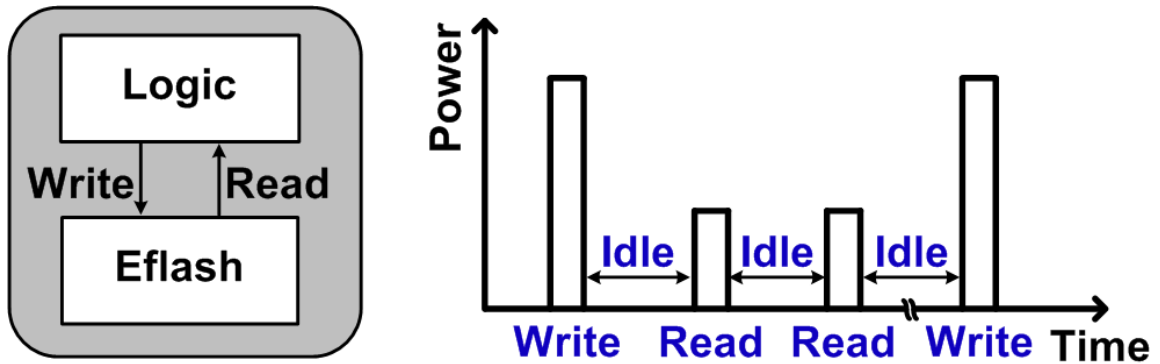


Fig. 5.17 Illustrations of the proposed eflash operations and power consumption during each operation mode.

Fig. 5.18 illustrates such reliability aware system-on-a-chip architecture where the embedded flash memory can be used to store the reliability information that is changed infrequently during the entire device life-time. Combined with the on-die circuit reliability monitor circuits illustrated in [46], the proposed logic compatible eflash memory may have a potential core building block as a cost-effective moderate density eNVM to improve the device reliability and life-time by storing the monitored reliability information and supporting the system reconfiguration based on it. For example, system parameters such as system clock and supply voltage levels can be adjusted dynamically, and memory redundancy schemes can be redefined, and run-time circuit calibration can be enabled through this system-on-a-chip architecture.

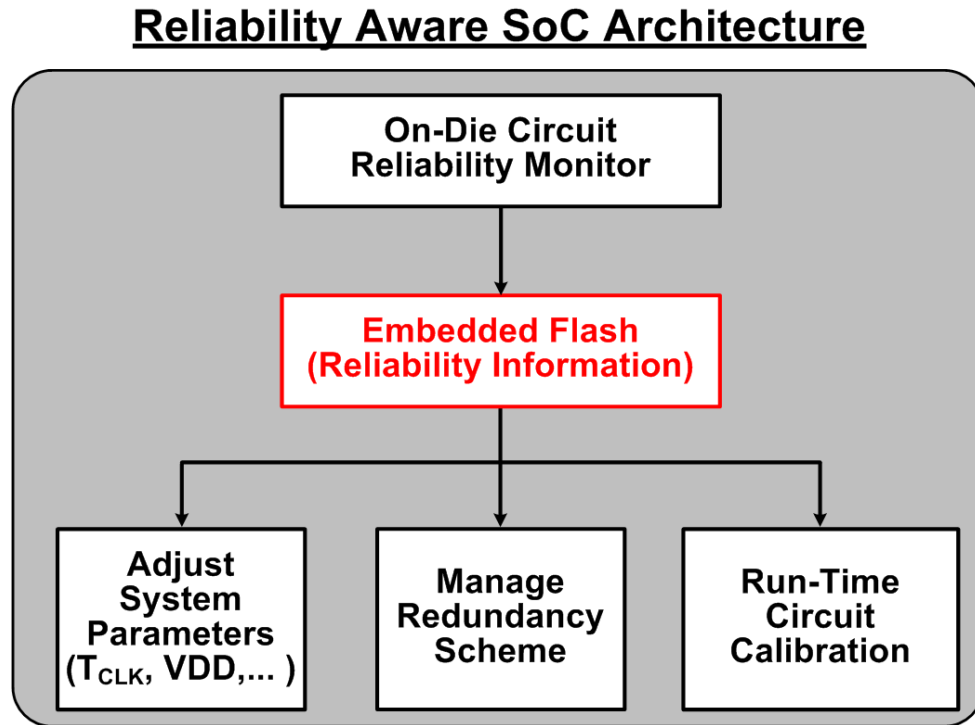


Fig. 5.18 Illustration of the reliability aware system-on-chip architecture where the embedded flash memory can be used to store the reliability information that is changed infrequently during the entire device life-time.

5.5 Chapter Summary

The 10T eflash featuring a multi-configurable HVS was proposed in this chapter. It does not require a boosted read supply, while improving the retention time significantly by the differential cell architecture, and was implemented in a 65nm generic logic process having 5nm tunnel oxide. Table 5.1 compares prior commercial logic compatible eflash memories (i.e. NOVEA [29], NOOEE [38], C-Flash [40]), 5T and 6T eflash discussed in chapters 2-4, and 10T eflash discussed in this chapter. The proposed 10T differential eflash does not require a boosted read supply like the prior commercial eflash memories but also functions without any unselected WL disturbance and HVS overstress issues like

the prior 5T and 6T eflash memories, making it a competitive candidate for the moderate density logic compatible and multiple-time programmable eNVM applications.

Table 5.1 Logic Compatible Embedded Flash Memory Comparison

Logic Compatible Embedded Flash	NOVEA [29]	NEOEE [38]	C-Flash [40]	5T Eflash [54, 55]	6T Eflash [57]	This Work
Process	0.13 μ m Logic	65nm Logic	0.13 μ m Logic	65nm Logic	65nm Logic	65nm Logic
Cell Transistor (Tunnel Oxide)	3.3V I/O TR (7nm)	2.5V I/O TR (5nm)	3.3V I/O TR (7nm)	2.5V I/O TR (5nm)	2.5V I/O TR (5nm)	2.5V I/O TR (5nm)
Unsel. WL Disturb	No	Yes	Yes	No	No	No
HVS Overstress	0.7V	N. A.	3.2V	0V	0V	0V
Erase/Write Unit	Block	WL	WL	WL	Bit	WL
Boosted Read Supply	None	None	None	VPP1~4	VPP1~4	None
Unit Cell Area	700 μ m ²	N. A.	72 μ m ²	*8.62 μ m ²	15.3 μ m ²	*16.6 μ m ² **24.8 μ m ²
Capacity	2kb	N. A.	256b	2kb	4kb	1792b

*PMOS Erase TR, **NMOS Erase TR

Chapter 6 Conclusions

An on-chip embedded NVM enables zero-standby power system-on-a-chip with a smaller form factor, faster access speed, lower access power, and higher security than the off-chip NVM. Differently from the high density eNVM technologies such as dual-poly eflash, FeRAM, STT-MRAM, and RRAM that typically require process overhead beyond logic technology, the moderate density eNVM technologies such as e-fuse, anti-fuse, and single-poly eflash can be built in a standard logic process without additional process steps; therefore, they have been adopted in many digital and analog integrated circuits for higher yield and performance. On the other hand, a single-poly eflash memory is multiple-times programmable, whereas e-fuse and anti-fuse cell are OTP; therefore, the single-poly eflash is expected to play a key role in mitigating run-time variability and reliability issues of the future VLSI technologies that the traditional OTP memories cannot easily have solved.

Typically, a single-poly eflash cell is implemented using standard I/O devices that are supposed to operate within the nominal I/O supply level; however, the unselected eflash cells in an array structure are typically prone to be driven to higher stress voltage than the nominal I/O supply level during the erase and program operations of the selected cell

requiring the around 3 to 4 times higher voltage than the nominal I/O supply. On the other hand, the design of an HVS circuit switching from GND to such high erase or program voltage level is another challenge, since the HVS circuit has to be implemented using standard logic devices, which can cause reliability issues. The limited retention time is also a big concern, since a single-poly eflash cell is implemented using a standard I/O device having as thin as 5nm tunnel oxide for 2.5V I/O device.

This thesis has focused on alleviating such challenges of the single-poly eflash memory with three single-poly eflash designs proposed in a generic logic process for moderate density eNVM applications.

In chapter 2, a logic-compatible 5T eflash memory was proposed for zero-standby power system-on-a-chip and 2kb test chip results in a 65nm standard logic process were discussed. This proposed 5T eflash has 60 to 80 times smaller cell size than the prior eflash memories [29, 30], and features WL-by-WL accessible architecture with negligible high voltage disturbance, the overstress-free multi-story HVS expanding the cell sensing margin, and a selective WL refresh scheme for the cell endurance to more than 10k P/E cycles [54, 55].

In chapter 3, the various types of single-poly eflash cell topologies were studied in a 65nm standard logic process having four kinds of 2.5V standard I/O devices to find the optimal eflash cell configuration [56]. A 5T eflash cell structure having PMOS coupling device, NCAP electron ejection device, and NMOS read device with two additional pass transistors for self-boosting is turned out to be the most favorable configuration when the performance, endurance, retention, and disturbance characteristics are all considered.

In chapter 4, a bit-by-bit re-writable 6T eflash was proposed for improving the overall endurance of the eflash array, and the 4kb test chip results in a 65nm standard logic process were discussed [57]. This proposed 6T eflash features the bit-by-bit re-write capability without disturbing the unselected WL cells. The on-chip negative HVS and CP provided the appropriate WL pulses for read and write operations of the proposed 6T eflash operations. With the proposed bit-by-bit re-writable 6T eflash, the overall endurance is estimated to be improved by around 4 times with 2:1 column MUX compared to the WL-by-WL erasable 5T eflash proposed in chapters 2 and 3.

In chapter 5, a 10T differential eflash was proposed for low power and high performance read operation with an improved cell sensing margin, and the 1792b test chip was designed in a 65nm standard logic process. This proposed 10T eflash features the multi-configurable HVS that does not require a boosted supply during read operation, and the improved retention time by the differential cell architecture which is less affected by the common mode signal such as read reference level (VRD), retention temperature, and power supply variation. This proposed 10T eflash is preferred for the applications requesting the write operation infrequently such as the reliability aware system-on-a-chip architecture.

All the proposed eflash memories in this thesis are implemented in a 65nm standard logic process, and the test chip measurement results confirm the functionality of the proposed designs without high voltage disturbance and overstress issues, making them competitive candidates for the moderate density on-chip multiple-time programmable NVM.

Bibliography

- [1] S. Hanson et al., “A low-voltage processor for sensing applications with picowatt standby mode,” *IEEE J. of Solid-State Circuits*, vol. 44, no. 4, pp. 1145-1155, Apr. 2009.
- [2] H. Hidaka, “Evolution of embedded flash memory technology for MCU,” in *Proc. IEEE Int. Conf. on IC Design and Technol. (ICICDT)*, 2011, pp. 1-4.
- [3] R. Strenz, “Embedded flash technologies and their applications: status & outlook,” in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2011, pp. 211-214.
- [4] H. Kojima et al., “Embedded flash on 90nm logic technology & beyond for FPGAs,” in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2007, pp. 677-680.
- [5] C. Deml, M. Jankowski, and C. Thalmaier, “A 0.13 μ m 2.125MB 23.5ns embedded flash with 2GB/s read throughput for automotive microcontrollers,” *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2007, pp. 478-479.

- [6] Y. Lee et al., "2T-FN eNVM with 90nm logic process for smart card," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW)*, 2008, pp. 26-27.
- [7] T. Ikehashi et al., "A 60ns access 32kByte 3-transistor flash for low power embedded applications," in *IEEE Symp. on VLSI Circuits Dig.*, 2000, pp. 162-165.
- [8] H. Lee et al., "NeoFlash - true logic single poly flash memory technology," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW)*, 2006, pp. 15-16.
- [9] M. Fliesler, D. Still, and J. Hwang, "A 15ns 4Mb NVSRAM in 0.13 μ m SONOS technology," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW)*, 2008, pp. 83-86.
- [10] J. Yater et al., "16Mb Split Gate Flash Memory with Improved Process Window," in *Proc. IEEE Int. Memory Workshop (IMW)*, 2009, pp. 1-2.
- [11] S. Kang et al., "High performance nanocrystal based embedded flash microcontrollers with exceptional endurance and nanocrystal scaling capability," in *Proc. IEEE Int. Memory Workshop (IMW)*, 2012, pp. 1-4.
- [12] Y. Yano, "Take the Expressway to go Greener," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 24-30.
- [13] T. Kono et al., "40nm embedded SG-MONOS flash macros for automotive with 160MHz random access for code and endurance over 10M cycles for data," in

- IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2013, pp. 212-213.
- [14] S. Bartling et al., "An 8MHz 75 μ A/MHz zero-leakage non-volatile logic-based cortex-M0 MCU SoC exhibiting 100% digital state retention at VDD=0V with <400ns wakeup and sleep transitions," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2013, pp. 432-433.
- [15] H. Yu et al., "Cycling endurance optimization scheme for 1Mb STT-MRAM in 40nm technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2013, pp. 224-225.
- [16] A. Kawahara et al., "Filament scaling forming technique and level-verify-write scheme with endurance over 10^7 cycles in ReRAM," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2013, pp. 220-221.
- [17] S. Kulkarni et al., "A 4kb metal-fuse OTP-ROM macro featuring a 2V programmable 1.37 μ m² 1T1R bit cell in 32nm high-k metal-gate CMOS," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 863-868, Apr. 2010.
- [18] S. Kulkarni et al., "A 32nm high-k and metal-gate anti-fuse array featuring a 1.01 μ m² 1T1C bit cell," in *IEEE Symp. on VLSI Technology Dig.*, 2012, pp. 79-80.
- [19] K. Matsufuji et al., "A 65nm pure CMOS one-time programmable memory using a two-port antifuse cell implemented in matrix structure," in *Proc. IEEE Asian Solid-State Circuits Conf. (ASSCC)*, 2007, pp. 212-215.

- [20] N. Phan, I. Chang, and J. Lee, "A 2kb one-time programmable memory for UHF passive RFID tag IC in a standard 0.18 μ m CMOS process," *IEEE Trans. Circuits and Systems*, vol. 60, no. 7, pp. 1810-1822, Jul. 2013.
- [21] T. Oh and R. Harjani, "A 12-Gb/s multichannel I/O using MIMO crosstalk cancellation and signal reutilization in 65-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 48, no. 6, pp. 1383-1397, Jun. 2013.
- [22] M. Seok et al., "A portable 2-transistor picowatt temperature-compensated voltage reference operating at 0.5V," *IEEE J. Solid-State Circuits*, vol. 47, no. 10, pp. 2534-2545, Oct. 2012.
- [23] Y. Lee et al., "A sub-nW multi-stage temperature compensated timer for ultra-low-power sensor nodes," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2511-2521, Oct. 2013.
- [24] W. Chen et al., "A 22nm 2.5MB slice on-die L3 cache for the next generation Xeon processor," in *IEEE Symp. on VLSI Circuits Dig.*, 2013, pp. 132-133.
- [25] J. Shin et al., "The next-generation 64b SPARC core in a T4 SoC processor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 60-61.
- [26] R. McPartland and R. Singh, "1.25 volt, low cost, embedded flash memory for low density applications," in *IEEE Symp. on VLSI Circuits Dig.*, 2000, pp. 158-161.
- [27] S. Shukuri, K. Yanagisawa, K. Ishibashi, "CMOS process compatible ie-flash (inverse gate electrode flash) technology for system-on a chip," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2001, pp. 179-182.

- [28] M. Yamaoka et al., "A system LSI memory redundancy technique using an inverse-gate-electrode flash (inverse-gate-electrode flash) programming circuit," *IEEE J. of Solid-State Circuits*, vol. 37, no. 5, pp. 599-604, May 2002.
- [29] J. Raszka et al., "Embedded flash memory for security applications in a 0.13 μ m CMOS logic process," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2004, pp. 46-47.
- [30] Y. Yamamoto 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.
- [31] Y. Yamamoto et al., "Nonvolatile semiconductor memory device," US Patent 7,755,941, Jul. 13, 2010.
- [32] B. Wang et al., "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, Sep. 2007.
- [33] Y. Ma et al., "Floating-gate nonvolatile memory with ultrathin 5nm tunnel oxide," *IEEE Trans. on Electron Devices*, vol. 55, no. 12, pp. 3476-3481, Dec. 2008.
- [34] A. Pesavento, F. Bernard, and J. Hyde, "PFET Nonvolatile Memory," US Patent 7,221,596, May 22, 2007.
- [35] L. Pan et al., "Pure logic CMOS based embedded non-volatile random access memory for low power RFID application," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2008, pp. 197-200.

- [36] P. Feng, Y. Li, and N. Wu, "An ultra low power non-volatile memory in standard CMOS process for passive RFID tags," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2009, pp. 713-716.
- [37] C. Shin and O. Kwon, "TFT-LCD driver IC with embedded non-volatile memory for portable applications," *J. of the Society for Information Display*, vol. 17, no. 5, pp. 481-487, May 2009.
- [38] H. Chen et al., "Single polysilicon layer non-volatile memory and operating method thereof," US Patent 8,199,578, Jun. 12, 2012.
- [39] Y. Roizin et al., "C-flash: an ultra-low power single poly logic NVM," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSMW)*, 2008, pp. 90-92.
- [40] H. Dagan et al., "A low-power DCVSL-like GIDL-free voltage driver for low-cost RFID nonvolatile memory," *IEEE J. Solid-State Circuits*, vol. 48, no. 6, pp. 1497-1510, Jun. 2013.
- [41] M. Porter et. al, "Reliability considerations for implantable medical ICs," in *Proc. IEEE Int. Reliability Physics Symp. (IRPS)*, 2008, pp. 516-523.
- [42] S. Oesterle, P. Gerrish, and P. Cong, "New interfaces to the body through implantable system integration," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2011, pp. 9-14.
- [43] H. Kim et. al, "A configurable and low-power mixed signal SoC for portable ECG monitoring applications," in *IEEE Symp. on VLSI Circuits Dig.*, 2011, pp. 142-143.

- [44] S. Lee et. al, "A programmable implantable micro-stimulator SoC with wireless telemetry: application in closed-loop endocardial stimulation for cardiac pacemaker," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2011, pp. 44-45.
- [45] G. Chen et. al, "A cubic-millimeter energy-autonomous wireless intraocular pressure monitor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2011, pp. 310-311.
- [46] J. Keane and C. H. Kim, "On-chip silicon odometers and their potential use in medical electronics," in *Proc. IEEE Int. Reliability Physics Symp. (IRPS)*, 2012, pp. 4C.1.1-4C.1.8.
- [47] M. Liang and C. Hu, "Electron trapping in very thin thermal silicon dioxides," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 1981, pp. 396-399.
- [48] A. Hdiy et al., "Relaxation of interface states and positive charge in thin gate oxide after Fowler-Nordheim stress," *AIP J. of Appl. Phys.*, vol. 73, no. 7, pp. 3569-3570, Apr. 1993.
- [49] Y. Park and D. Schroder, "Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM," *IEEE Trans. on Electron Devices*, vol. 45, no. 6, pp. 1361-1368, Jun. 1998.
- [50] Y. Manabe et al., "Detailed observation of small leak current in flash memories with thin tunnel oxides," *IEEE Trans. on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 170-174, May 1999.

- [51] R. Bez et al., "Introduction to flash memory," *Proc. of the IEEE*, vol. 91, no. 4, Apr. 2003.
- [52] J. Lee et al., "Effects of interface trap generation and annihilation on the data retention characteristics of flash memory cells," *IEEE Trans. on Device Materials and Reliability*, vol. 4, no. 1, pp. 110-117, Mar. 2004.
- [53] N. Mielke et al., "Flash EEPROM threshold instabilities due to charge trapping during program/erase cycling," *IEEE Trans. on Device Materials and Reliability*, vol. 4, no. 3, pp. 335-344, Sep. 2004.
- [54] S. Song, K. Chun, and C. H. Kim, "A Logic-Compatible Embedded Flash Memory Featuring a Multi-Story High Voltage Switch and a Selective Refresh Scheme," in *IEEE Symp. on VLSI Circuits Dig.*, 2012, pp. 130-131.
- [55] S. Song, K. Chun, and C. H. Kim, "A logic-compatible embedded flash memory for zero-standby power system-on-chips featuring a multi-story high voltage switch and a selective refresh scheme," *IEEE J. Solid-State Circuits*, vol. 48, no. 5, pp. 1302-1314, May 2013.
- [56] S. Song, J. Kim, and C. H. Kim, "Program/erase speed, endurance, retention, and disturbance characteristics of single-poly embedded flash cells," in *Proc. IEEE Int. Reliability Phys. Symp.(IRPS)*, 2013, pp. MY.4.1-MY.4.6.
- [57] S. Song, K. Chun, and C. H. Kim, "A bit-by-bit re-writable eflash in a generic logic process for moderate-density embedded non-volatile memory applications," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2013, pp. 1-4.

- [58] K. Suh et al., "A 3.3V 32Mb NAND flash memory with incremental step pulse programming scheme," *IEEE Jour. of Solid-State Circuits*, vol. 30, no. 11, pp. 1149-1156, Nov. 1995.
- [59] T. Jung et al., "A 117mm² 3.3V only 128Mb multilevel NAND flash memory for mass storage applications," *IEEE Jour. of Solid-State Circuits*, vol. 31, no. 11, pp. 1575-1583, Nov. 1996.
- [60] M. Jefremow et al., "Bitline-capacitance-cancelation sensing scheme with 11ns read latency and maximum read throughput of 2.9GB/s in 65nm embedded flash for automotive," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 428-429.
- [61] P. Favrat, P. Deval., and M. Declercq, "A high-efficiency CMOS voltage doubler," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 410-416, Mar. 1998.
- [62] R. Pelliconi et al., "Power efficient charge pump in deep submicron standard CMOS technology," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 1068-1071, Jun. 2003.
- [63] M. Ker, S. Chen, and C. Tsai, "Design of charge pump circuit with consideration of gate-oxide reliability in low-voltage CMOS processes," *IEEE J. Solid-State Circuits*, vol. 41, no. 5, pp. 1100-1107, May 2006.
- [64] Y. Pan et al., "Quasi-nonvolatile SSD: trading flash memory nonvolatility to improve storage system performance for enterprise applications," in *Proc. IEEE Int. Symp. on High Performance Computer Architecture (HPCA)*, 2012, pp. 1-10.
- [65] Q. Wu, G. Dong, and T. Zhang, "A first study on self-healing solid-state drives," in *Proc. IEEE Int. Memory Workshop (IMW)*, 2011, pp. 1-4.

- [66] C. Miccoli et al., "Assessment of distributed-cycling schemes on 45nm NOR flash memory arrays," in *Proc. IEEE Int. Reliability Physics Symp. (IRPS)*, 2012, pp. 2A.1.1-2A.1.7.
- [67] S. Tanakamaru, Y. Yanagihara, K. Takeuchi, "Over-10x-extended-lifetime 76%-reduced-error solid-state drives (SSDs) with error-prediction LDPC architecture and error-recovery scheme," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 424-425.
- [68] A. Umezawa et al., "A new self-data-refresh scheme for a sector erasable 16Mb flash EEPROM," in *IEEE Symp. on VLSI Circuits Dig.*, 1993, pp. 99-100.
- [69] J. Lee, S. Hur, and J. Choi, "Effect of floating-gate interference on NAND flash memory cell operation," *IEEE Electron Device Letters*, vol. 23, no. 5, pp. 264-266, May 2002.
- [70] Y. Shi et al., "Polarity dependent gate tunneling currents in dual-gate CMOSFET's," *IEEE Trans. on Electron Devices*, vol. 45, no. 11, pp. 2355-2360, Nov. 1998.
- [71] K. Schuegraf and C. Hu, "Hole injection SiO₂ breakdown model for very low voltage lifetime extrapolation," *IEEE Trans. on Electron Devices*, vol. 41, no. 5, pp. 761-767, May 1994.
- [72] Y. Yeo, Q. Lu, and C. Hu, "MOSFET gate oxide reliability: anode hole injection model and its applications," *Int. J. of High Speed Electron. and Sys.*, vol. 11, no. 3, pp. 849-886, Sep. 2001.
- [73] S. Satoh et al., "A novel isolation-scaling technology for NAND EEPROMs," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 1997, pp. 291-294.
- [74] H. Fujii et al., "x11 performance increase, x6.9 endurance enhancement, 93% energy reduction of 3D TSV-integrated hybrid ReRAM/MLC NAND SSDs by

- data fragmentation suppression,” in *IEEE Symp. on VLSI Circuits Dig.*, 2012, pp. 134-135.
- [75] H. Lue et al., “A novel bit alterable 3D NAND flash using junction-free p-channel device with band-to-band tunneling induced hot-electron programming,” in *IEEE Symp. on VLSI Technology Dig.*, 2013, pp. 152-153.
- [76] K. Chun et al., “A 667MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches,” *IEEE J. Solid-State Circuits*, vol. 47, no. 2, pp. 547-559, Feb. 2012.