A Bit-by-Bit Re-Writable Eflash in a Generic 65 nm Logic Process for Moderate-Density Nonvolatile Memory Applications

Seung-Hwan Song, Ki Chul Chun, and Chris H. Kim, *Senior Member, IEEE*

***Abstract*—Embedded nonvolatile memory (eNVM) is considered to be a critical building block in future system-on-chip and microprocessor systems. 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 eflash 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 work, we present details of a bit-by-bit re-writable eflash memory implemented in a generic 65 nm logic process which addresses this key challenge. The proposed 6 T eflash memory cell can improve the overall cell endurance by eliminating redundant program/erase cycles while preventing disturbance issues in the unselected wordlines. We also provide details of special high voltage circuits such as a voltage-doubler based charge pump circuit and a multistory high-voltage switch, for generating a reliable high-voltage output without causing damage to the standard logic transistors.**

***Index Terms*—Charge pump, embedded nonvolatile memory (eNVM), negative high-voltage switch, nonvolatile memory (NVM), single-poly embedded flash memory.**

I. Introduction

EMBEDDED nonvolatile memory (eNVM) technologies have been deployed in a growing number of high density (∼Mb) and moderate density (∼kb) applications as shown in Fig. 1(a). High-density eNVM such as dual-poly embedded flash (eflash), FeRAM, STT-MRAM, and RRAM are targeted for on-chip code and data storage [1]–[4]; however, they typically incur a significant process overhead compared with a standard logic technology. One-time programmable (OTP) memories such as e-fuse (electronic-fuse) and anti-fuse can be built in a generic logic process and have been widely used for memory redundancy, circuit trimming, and digital calibration [5]–[9]. However, the main shortcoming of OTP memories is that the program operation is irreversible. Single-poly eflash, on the other hand, can be built in a generic logic process and can be programmed multiple times, making it a viable alternative to OTP [10]–[17] for secure moderate-density NVM applications. Unlike dual-poly eflash cell where the floating gate (FG) and control gate (CG) are integrated into a single device, as shown in Fig. 1(b), single-poly eflash cells typically consist of three discrete devices: namely, the coupling, read, and write devices. The gates of these devices are back-to-back connected to form an explicit FG. The n-well of the coupling transistor acts like the CG in a dual-poly eflash cell. Consequently, single-poly

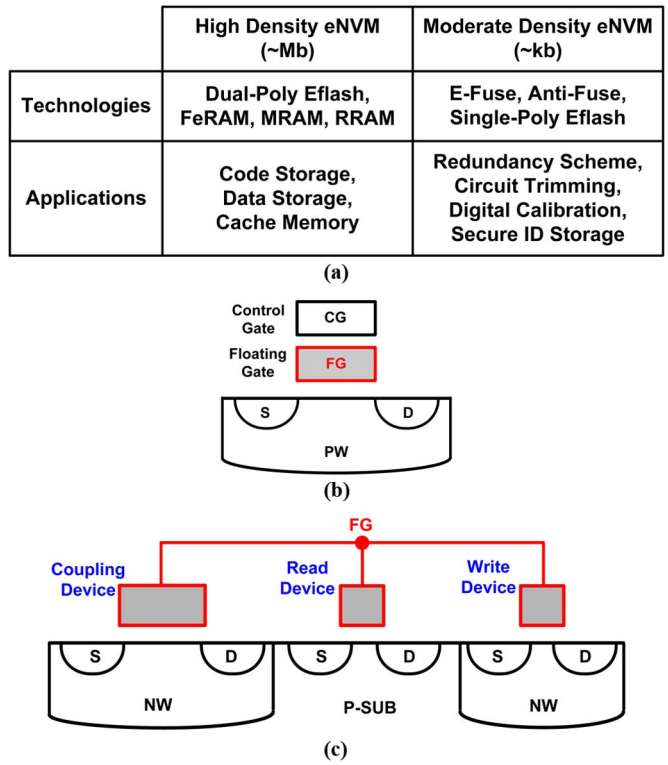


Fig. 1. (a) Embedded NVM technologies and their applications. (b) Dual-poly eflash cell comprises a single transistor with an FG and a CG. (c) Single-poly eflash cell consists of coupling, read, and write transistors. The FG is formed by the back-to-back gate connection of discrete transistors. N-well of the coupling transistor acts like the control gate of the dual-poly eflash cell and, thus, no additional process overhead is incurred compared to a generic logic technology.

Manuscript received December 08, 2013; revised February 04, 2014; accepted March 11, 2014. Date of publication April 16, 2014; date of current version July 21, 2014. This paper was approved by Guest Editor Andrea Mazzanti.

S.-H. Song was with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. He is now with the Storage Architecture Laboratory, HGST, San Jose, CA 95135 USA (e-mail: songx278@umn.edu).

K. C. Chun was with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. He is now with Memory Division, Samsung Electronics, Gyeonggi-do, Korea.

C. H. Kim is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2014.2314445

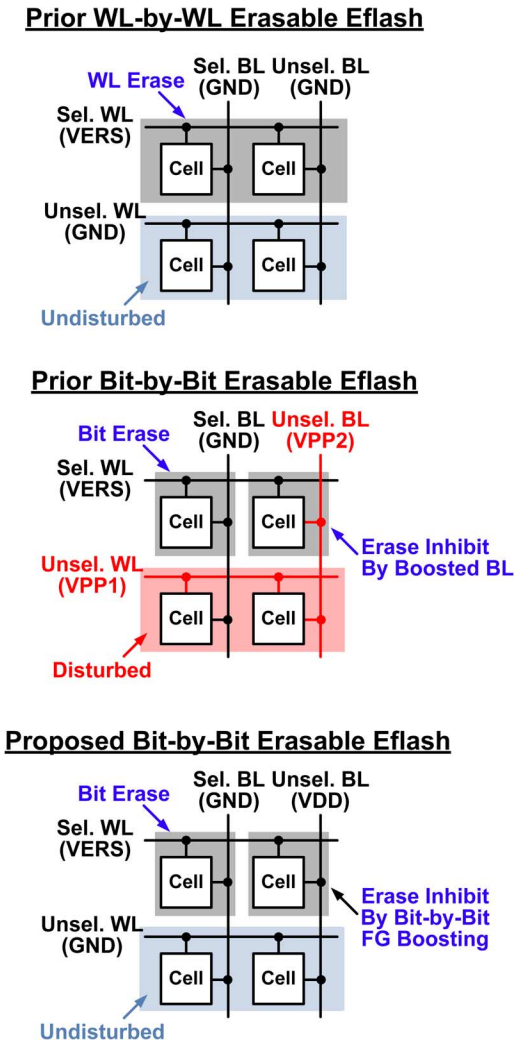


Fig. 2. Disturbance issue comparison between WL-by-WL and bit-by-bit erasable single-poly eflash cells. The prior WL-by-WL erasable eflash requires all cells in the selected WL to be erased simultaneously, which results in unnecessary erase cycles for cells whose data remain unchanged. The prior bit-by-bit 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. 3) minimizing disturbance issues.

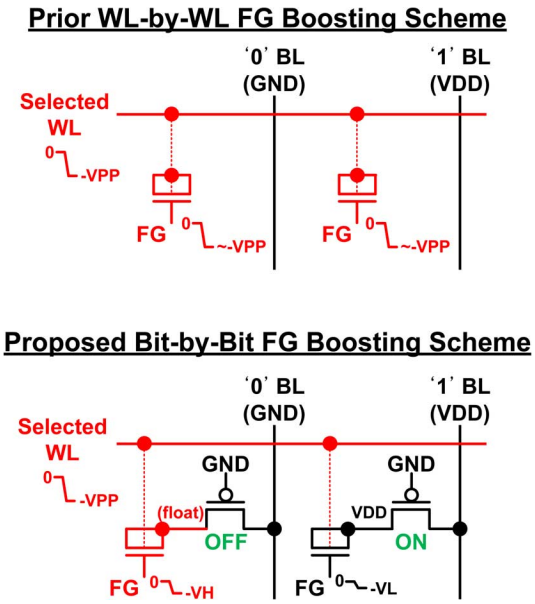


Fig. 3. Bias conditions of the coupling TR compared between the prior WL-by-WL and the proposed bit-by-bit FG boosting schemes. The former boosts all the FG's in the selected WL irrespective of the BL levels while the latter selectively boosts the FG depending on the write data (i.e. $-\mathrm{VPP} < -\mathrm{VH} < -\mathrm{VL}$).

eflash can be implemented using standard I/O devices that are readily available in a generic logic technology.

One of the key endurance-limiting factors for single-poly eflash is the large number of unnecessary erase cycles undergone by a cell when random data are 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. 2. In all three cases, a boosted voltage VERS that is three to four times the I/O VDD is applied to the selected WL to induce Fowler–Nordheim (FN) tunneling in the cells to be erased. Prior WL-by-WL erasable eflash memories [13]–[16] require the entire WL to be erased simultaneously prior to the program operation, which results in unnecessary erase cycles (Fig. 2, top). Prior bit-by-bit erasable eflash cells [12], on the other hand, can erase the selected cells without incurring unnecessary erase cycles (Fig. 2, 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 WLs. The problem here is that, for typical VPP2 and VPP1 voltage levels (e.g., 10 and 5 V), the difference between the two is still greater than the nominal I/O VDD (=2.5 V or 3.3 V) so high voltage disturbance issues cannot be completely avoided in the unselected WL cells. The proposed eflash illustrated in Fig. 2 (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 [18], [19] 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.

In this paper, we propose a bit-by-bit re-writable eflash in a generic logic process based on the bit-by-bit erasable cell structure in [20]. The remainder of this paper is organized as follows. Section II describes the proposed 6 T eflash memory and cell operation principles. Section III describes the negative high-voltage switch and charge pump circuit design. Section IV presents the measurement results from a test chip fabricated in a standard 65 nm logic process. Comparison with other logic compatible eNVM designs is given in Section V, followed by a conclusion in Section VI.

II. Proposed 6 T Eflash Memory

A. Bit-by-Bit FG Boosting Scheme

To accomplish a bit-by-bit write operation without disturbing the cells in the unselected WLs, we propose the bit-by-bit FG boosting scheme described in Fig. 3. Here, the bias conditions

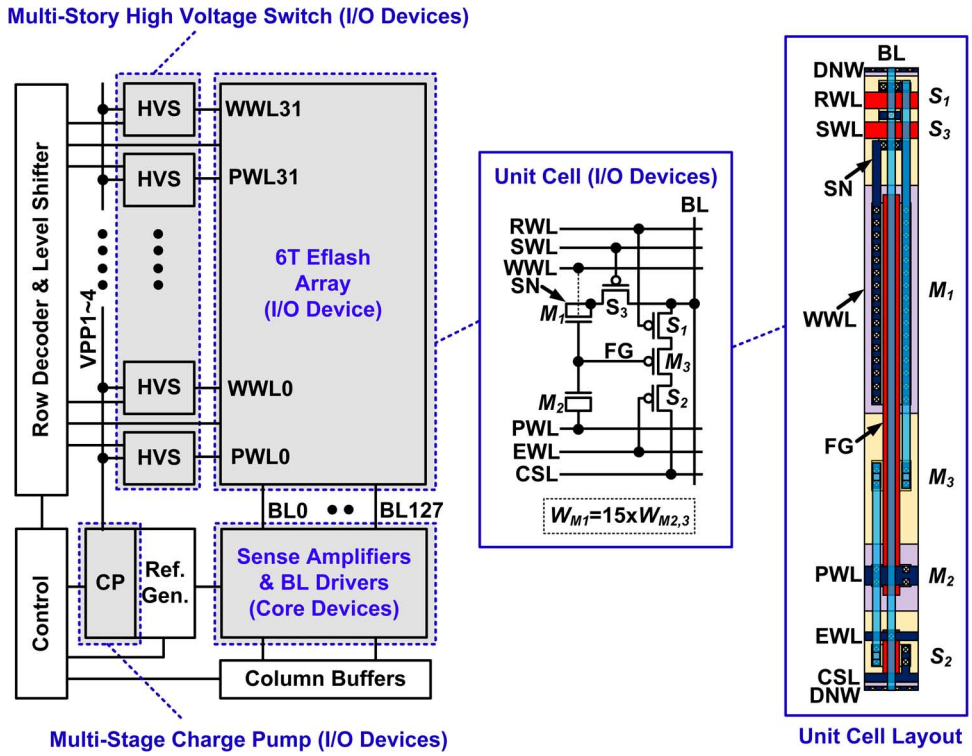


Fig. 4. Test chip diagram of the proposed bit-by-bit re-writable eflash memory. The 6 T eflash cell array, multistory high-voltage switch, and multistage charge pump are implemented using standard 2.5 V I/O devices with a 5 nm gate oxide. The sense amplifiers and BL drivers are implemented using 1.2 V core devices.

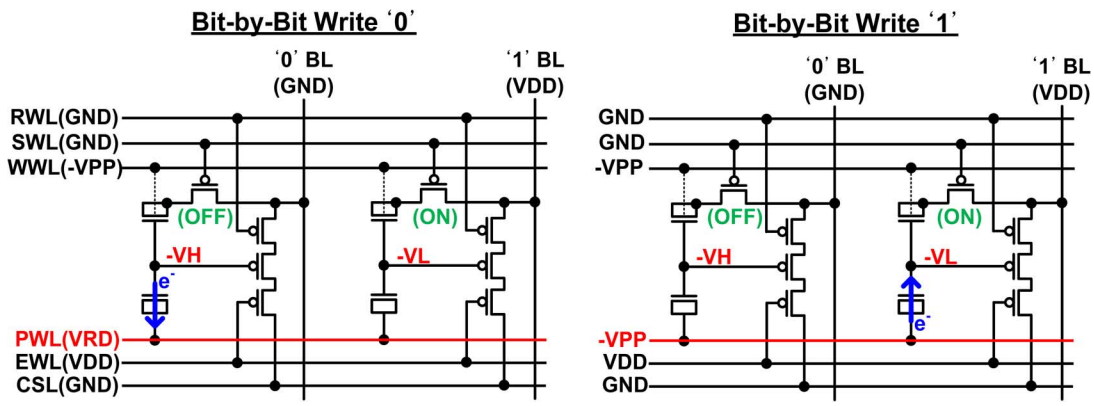


Fig. 5. Bit-by-bit write "0" and write "1" phases of the proposed 6 T eflash cell. The "0" BL cell loses electrons from FG during a write "0" phase, whereas the "1" BL cell adds electrons in FG during a write "1" phase via electron FN tunneling.

of the coupling transistor are compared with those of the prior WL-by-WL FG boosting scheme. When the selected WL is switched from 0 V to a boosted write voltage -VPP, 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. $-\mathrm{VPP} < -\mathrm{VH} < -\mathrm{VL}$), as the source and drain of the coupling transistors (i.e., SN node in Fig. 4) are boosted together for a "0" BL but tied to VDD for a "1" BL through the additional pass transistor connected to the BL. An NMOS coupling transistor was chosen over a PMOS one as the latter would have resulted in a floating channel during read operation when a positive read reference (i.e., VRD in Fig. 6) is applied to the body. This could potentially worsen the cell threshold voltage variation as the voltage level of the floating channel is unknown.

B. 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 along with the schematic and layout of the new 6 T cell. The 6 T eflash cell is composed of three main transistors $(M_1 - M_3)$ and three pass transistors $(S_1 - S_3)$. 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 6 T eflash cell array, multistory high-voltage switch (HVS), and

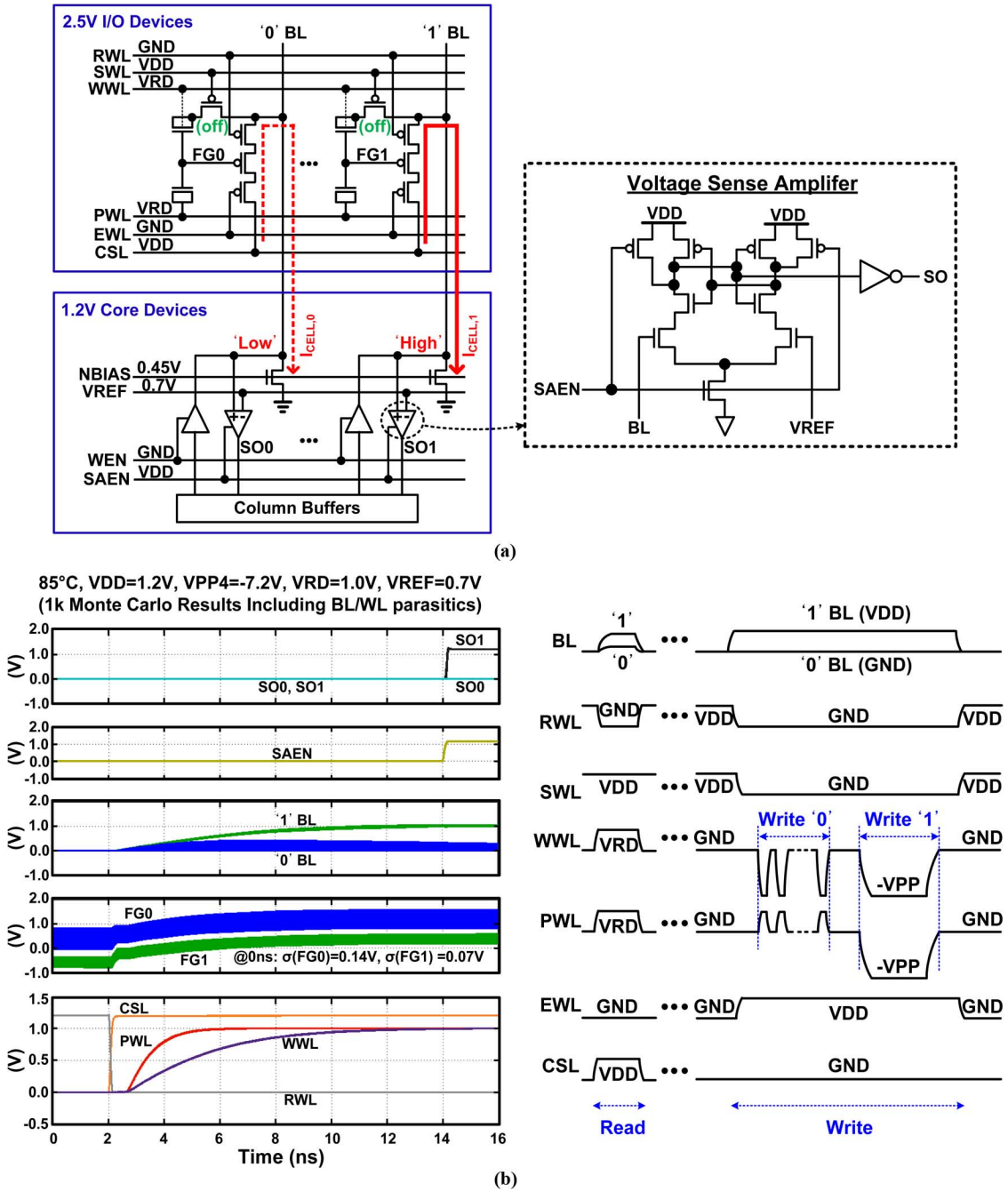


Fig. 6. (a) Read bias condition and sense amplifier schematic. (b) Simulated read waveforms and timing diagram for a full bit-by-bit update sequence.

multistage charge pump (CP) are implemented using 2.5 V I/O devices having a 5 nm gate oxide, while the sense amplifiers and BL drivers are implemented using 1.2 V core devices.

C. Cell Operation of the Proposed 6 T Eflash Memory

The complete write operation of the proposed 6 T eflash consists of the two phases illustrated in Fig. 5. First, SWL is driven to GND while BL is driven to either GND or VDD depending on the write data. Subsequently, WWL is switched to a negative boosted voltage, -VPP (e.g. −7.2 V). Then, the pass transistor in a '0' BL cell (i.e. S_3 in Fig. 4) is turned off, while this transistor is turned on in a "1" BL cell. Consequently, the FG node voltage of the '0' BL cell is boosted to a large negative voltage (-VH) while the "1" BL cell sees only a small negative voltage (-VL) according to the aforementioned bit-by-bit FG boosting. During write "0" phase, PWL is driven to a small positive voltage, VRD (e.g., 1.6 V) and thus electron FN tunneling occurs in the "0" BL cell. During write "1" phase, PWL is switched to the negative boosted voltage -VPP 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 -VH and -VL are simulated as −4.2 V and −0.6 V during write "0" phase, and −5.2 V and −1.9 V during write "1" phase, respectively, for a -VPP of −7.2 V and VRD of 1.6 V.

Fig. 6(a) shows the read bias condition of the proposed 6 T eflash cell along with the sense amplifier schematic. Simulated

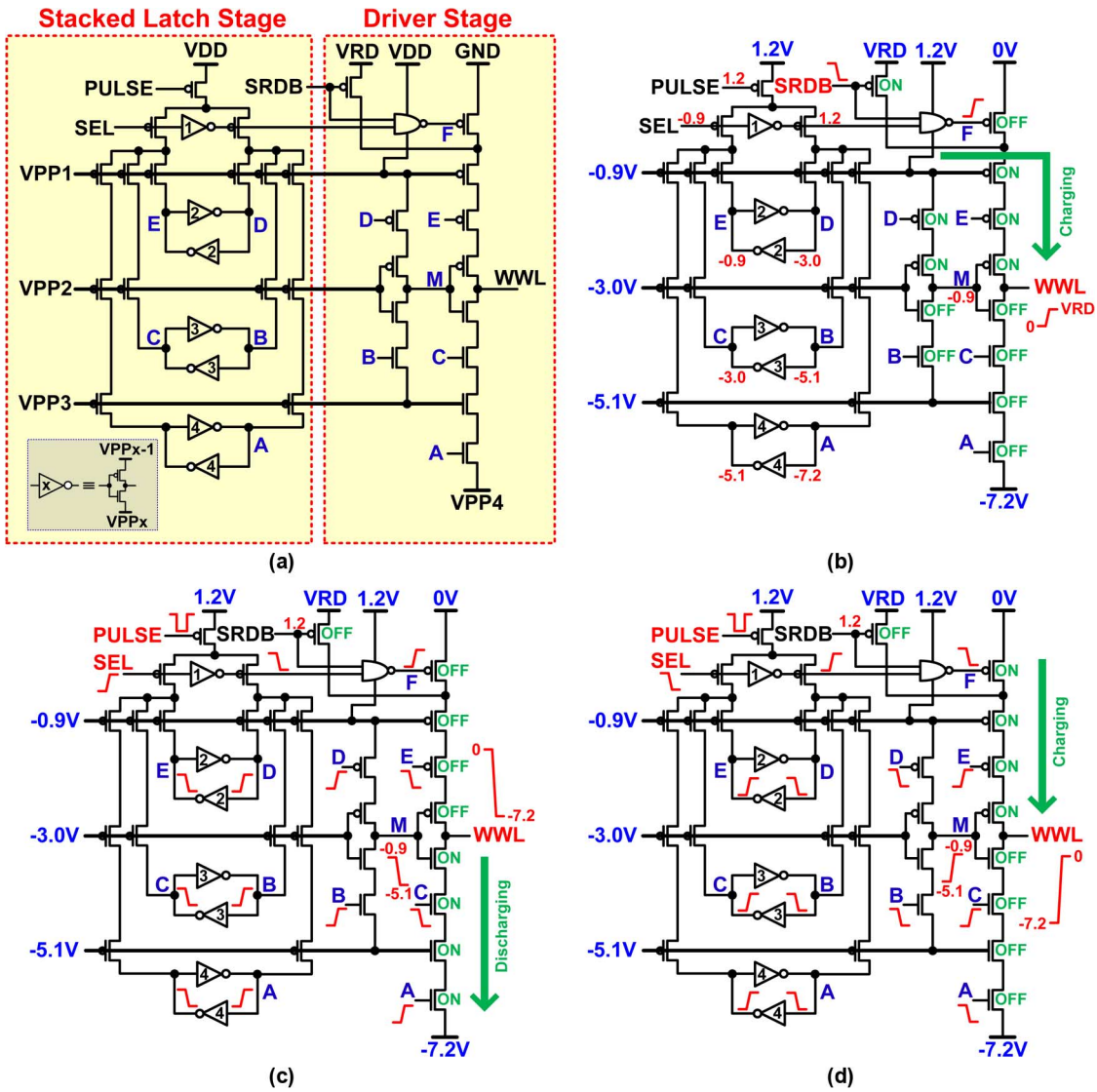


Fig. 7. (a) Multistory negative HVS consists of a stacked latch stage and a driver stage which prevents gate overstress during read and write operations. (b) During read, WWL is driven to VRD through the PMOS string without changing the latch states. (c), (d) During write, WWL is switched between VPP4 and GND by the SEL signal.

waveforms during read operation and the timing diagram for a full bit-by-bit update sequence are given in Fig. 6(b). During read operation, the pass devices (i.e., S_3 in Fig. 4) in the selected WL cells are 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 RWL and CSL are driven to GND and VDD, respectively. Then, BL starts to charge at a different rate according to the data stored in the cell. The FG node of the "1" cell contains more electrons than the "0" cell, thereby generating a higher cell current and a higher BL voltage. When the SA enable signal (i.e., SAEN) is activated, the sense amplifier compares the BL voltage levels with the reference level VREF to produce the digital output (i.e., SO in Fig. 6). The write enable signal (i.e., WEN) controlling the BL driver circuit is switched to GND during read operation and VDD during write operation. The read access time based on 1 k Monte Carlo simulations is 14 ns for BL and WL parasitic capacitances of C_{BL} = 70 fF, C_{WWL} = 800 fF, and C_{PWL} = 220 fF.

The full bit-by-bit update sequence of the proposed 6 T eflash cell consists of a read and a write operation. First, 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. 13 (top left). This is due to the strong coupling between the SN (shown in Fig. 4) nodes of adjacent cells that reduces the FG boosting effect for "0" BL cells explained in Fig. 3.

III. Negative High-Voltage and Switch Charge Pump

A. Negative High-Voltage Switch

The proposed multistory negative HVS is illustrated in Fig. 7. This is a modified version of the original multistory positive

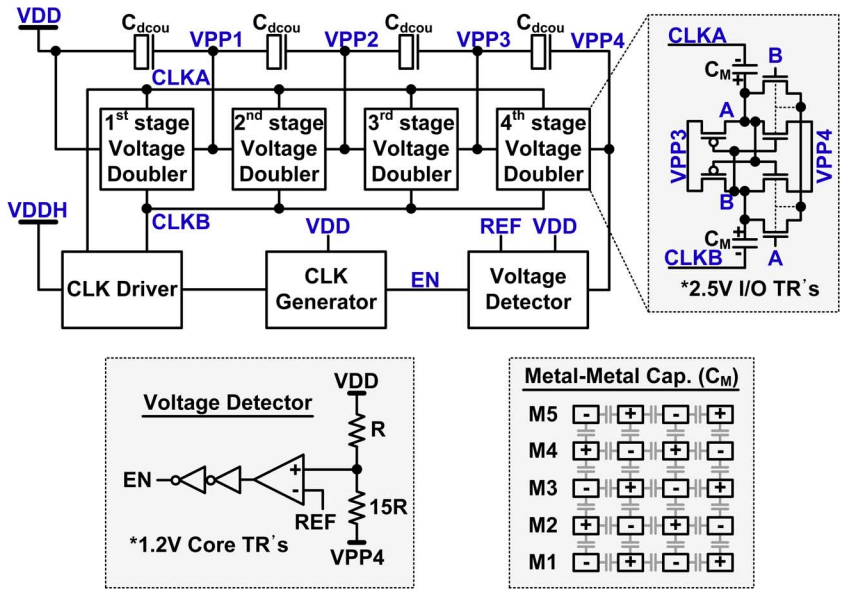


Fig. 8. A negative charge pump generating multiple boosted negative voltage levels (VPP1-VPP4) is implemented in a 65 nm standard logic process by cascading four voltage doubler stages (C_{dcou} = 300 fF, C_M = 900 fF).

HVS published in [16]. The HVS consists of stacked latch and driver stages, and are implemented using 2.5 V standard I/O devices. The stacked configuration effectively prevents gate overstress during the read and write operations as the transistor node voltages in the HVS are controlled by the precise voltage levels set by the charge pumps. The HVS is used as the WWL and PWL drivers. The boosted negative voltages $\mathrm{VPP1} \sim 4$ are supplied from the negative CP shown in Fig. 8. The nominal boosted voltage levels for $\mathrm{VPP1} \sim 4$ are -0.9, -3.0, -5.1, and -7.2 V, 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. 7(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.

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, 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, 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 [16], the PULSE signal width and the transistor sizes are optimized such that the latch states switch reliability while static power consumption is kept small so as to minimize the current loading of the negative CP.

B. Negative Charge Pump

The proposed negative HVS requires multiple boosted negative voltages, and these multiple boosted negative voltages are generated from the voltage doubler-based [21] on-chip negative CP shown in Fig. 8. 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 [22], [23], 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. The clock driver is supplied by the I/O voltage (VDDH) to generate the two-phase pump clock signals (CLKA and CLKB). The pump capacitors (C_M) are implemented using parasitic metal-metal capacitors. The capacitance of each pump capacitor is approximately 900 fF. Extra decoupling capacitors (C_{dcout}, ~300 fF) are added between the VPP levels to minimize the VPP1–VPP4 voltage overshoot while the proposed negative HVS is switching.

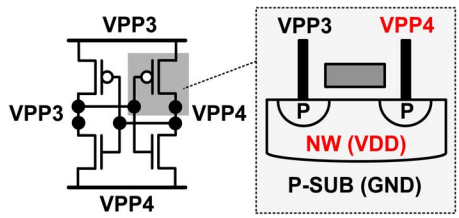


Fig. 9. Diagram of the junction breakdown issue in the proposed negative HVS and CP. Junction breakdown of the PMOS devices in the HVS bottom latch and the CP final stage limits the maximum negative VPP4 level.

C. Junction Breakdown Limit of the Designed Negative HVS and CP

Fig. 9 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 in [16] do not suffer from junction breakdown issues,

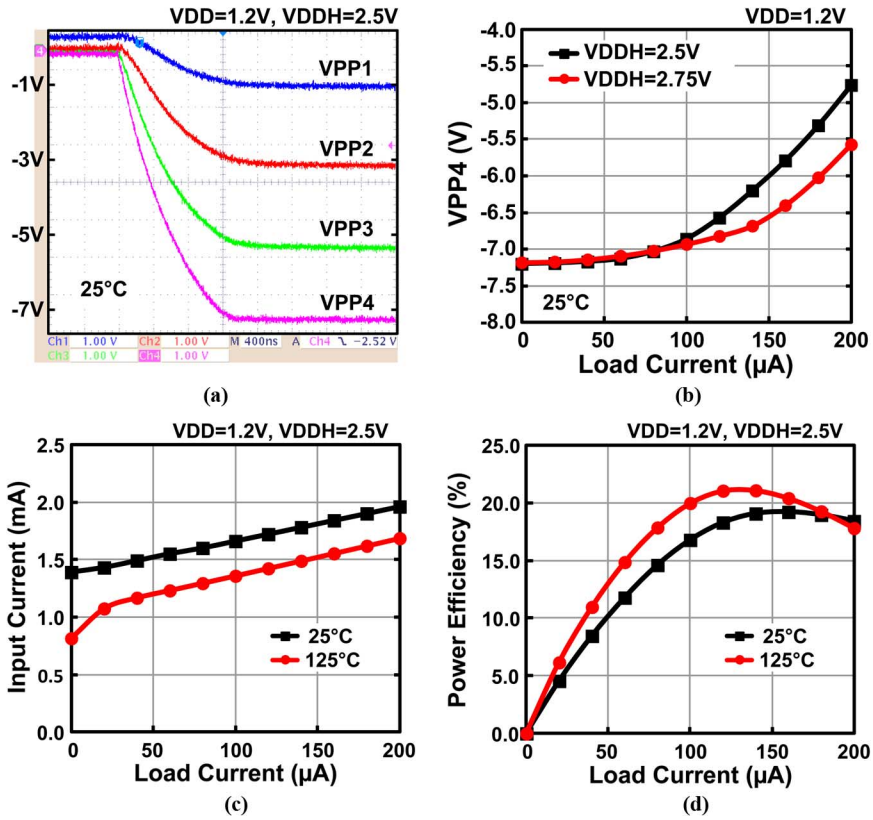


Fig. 10. (a), (b) Measured boosted negative voltages (VPP1–VPP4) and output characteristic of the fabricated negative charge pump (f_{CLK} = 90 MHz). (c), (d) Input current and power efficiency of charge pump versus load current.

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., 5 V). Using this configuration the junction reverse bias is limited to roughly half the breakdown voltage (e.g., 10 V).

IV. Test Chip Measurement Results

A 4 kb eflash test macro was implemented in a 65 nm low-power standard CMOS logic process to demonstrate the proposed circuit ideas. Fig. 10(a) and (b) shows the boosted negative voltages (VPP1–VPP4) and the output characteristic measured from the fabricated negative CP. A stable output voltage was measured for load currents exceeding the typical single WL write current ($\sim$ 0.1 μA) as can be seen in the figure. The simulated pump clock frequency (f_{CLK}) is 90 MHz and an I/O pad capacitance of $\sim$300 fF was added to the charge pump output nodes (VPP1–VPP4) for probing purposes. The equivalent output resistance of the charge pump according to the measured data is 23 kΩ, which is close to the theoretical prediction of 25 kΩ for a four-stage voltage doubler-based CP [21]. The current drawn by the CP and the power efficiency were measured for different load currents and are shown in Fig. 10(c) and (d). Here, power efficiency is defined as the output load power (=load current $\times$ (VDD1–VPP4)) divided by the input power. The measured average read current for a single WL access was around 140 μA which is much lower than the CP input current of around 1.4 mA. Fig. 11 shows the measured waveforms of the CP output VPP4 and the HVS outputs WWL and PWL for a bit-by-bit write "0" triggered by the WLS pulse.

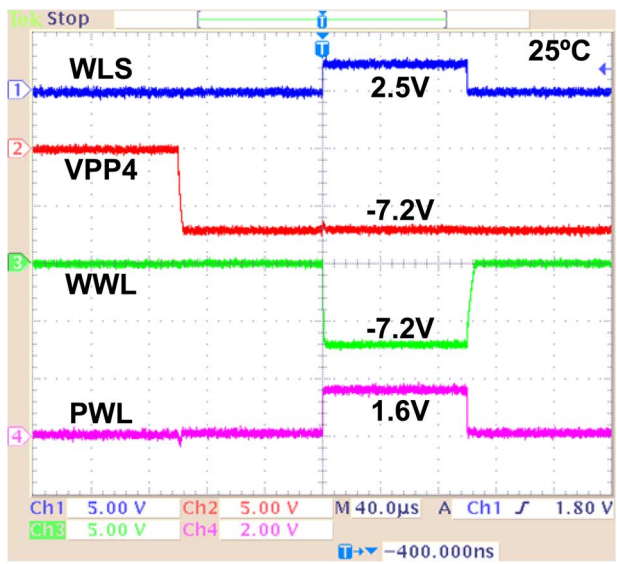


Fig. 11. Measured waveforms of the CP and HVS.

Fig. 12 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 $4n + 3$ $(n = 0, 1, 2, \ldots)$ are updated from "1" to "0" after a write "0" phase. Next, cells connects to BL 4n $(n = 0, 1, 2, \ldots)$ 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.

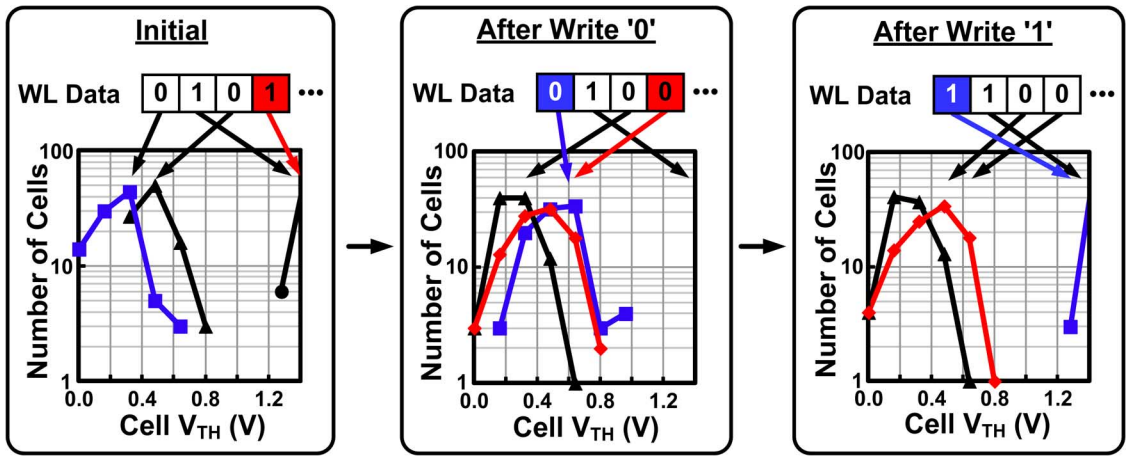
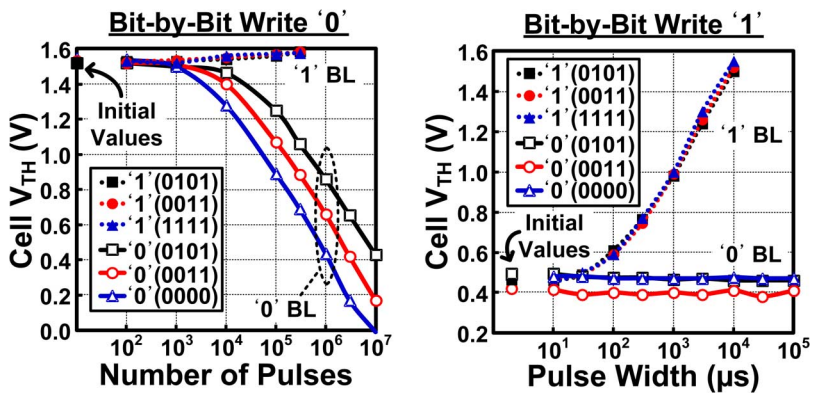


Fig. 12. Measured bit-by-bit update result from pattern (0101) to pattern (1100).



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

Fig. 13. Top: measured cell V_{TH} shift from the 6 T eflash test chip. Note that multiple write pulses with a fixed pulse width of 10 μs 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.

The bit-by-bit update from (0101) pattern to (1100) pattern is therefore achieved without any cells being unnecessarily erased.

Fig. 13 (top) shows the measured bit-by-bit write and disturbance results of the proposed 6 T 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. 13 (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). The measured result shows that the (0101) pattern has the slowest write "0" speed suggesting a large inter SN node coupling capacitance (C_{ISN}) for this pattern. A higher C_{ISN} reduces the FG boosting effect for "0" BL cells which in turn slows 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.

Fig. 14 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 1 k pre-cycles meet a one-year retention time at 85 °C maintaining a cell V_{TH} margin of ~0.7 V. To estimate the overall endurance improvement of the proposed 6 T eflash compared to the prior 5 T eflash [16], the average number of stress cycles for each state transition was compared in Fig. 15 based on the measured results. The cell V_{TH} transition plot in Fig. 15(a) shows an example for

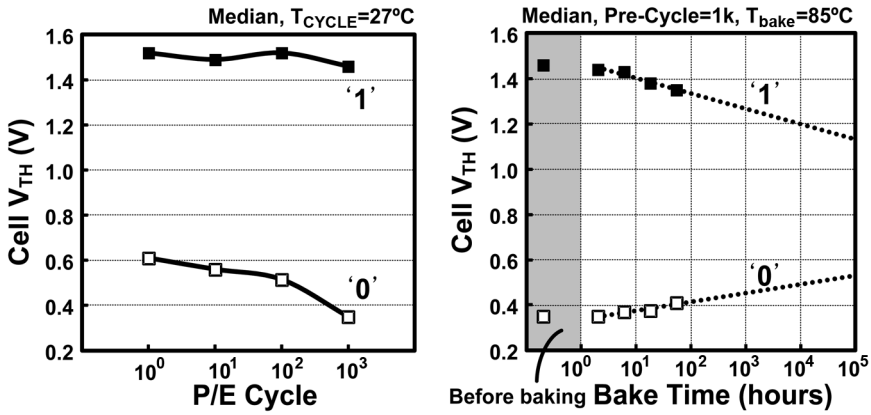
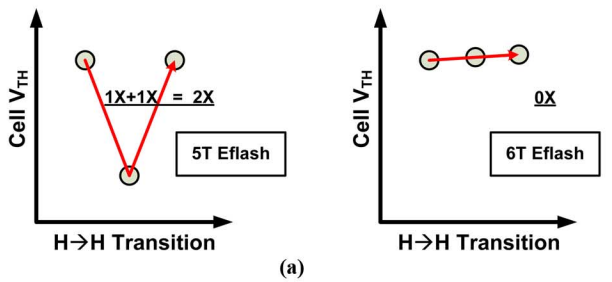
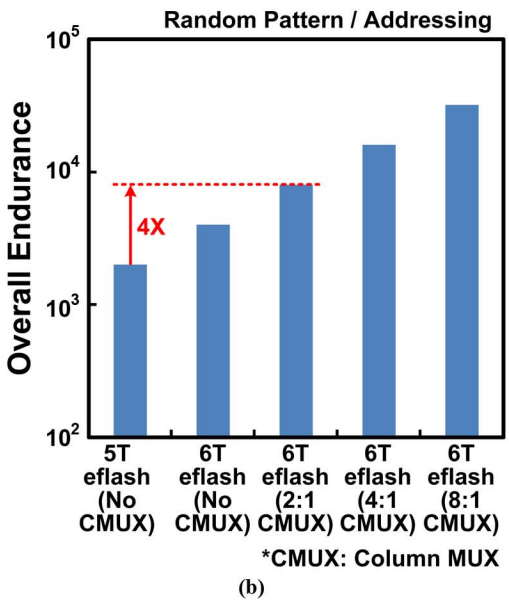


Fig. 14. Measured cell endurance and retention characteristics.

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



(a)



(b)

Fig. 15. (a) Average number of stress cycles for different data transitions and cell V_{TH} transition plot for a high-to-high transition in the prior 5 T eflash [16] and the proposed 6 T eflash cells. (b) Overall endurance estimated based on the average stress cycle count in (a). A smaller word size (i.e., larger column multiplexing ratio) improves the overall endurance for the proposed 6 T eflash.

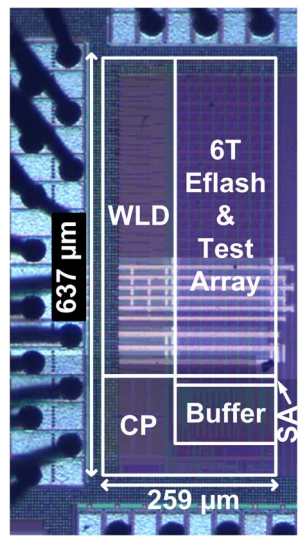


Fig. 16. Die photograph of 4 kb eflash test chip implemented in a 65 nm generic logic process.

a high-to-high transition. The prior WL-by-WL erasable 5 T eflash undergoes two stress cycles while the new 6 T eflash experiences relatively insignificant cell V_{TH} shift. Based on the information listed in Fig. 15(a) for the various transition cases, we can conclude that the overall cell V_{TH} shifts of the proposed 6 T eflash is roughly half compared to that of the prior 5 T eflash (no column multiplexing case). Half the number of stress cycles in the proposed 6 T eflash implies roughly half the number of traps generated, enhancing the overall endurance limit by twice compared to the prior 5 T 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. 15(b) for different eflash configurations assuming a random data pattern and random addressing. When the proposed 6 T eflash is used with a 2:1 column MUX, the overall endurance can be improved by around 4 times compared to the previous 5 T eflash. Finally, Fig. 16 shows the die photograph of the fabricated 4 kb eflash test chip.

TABLE I
LOGIC COMPATIBLE EMBEDDED NVM COMPARISON

CMOS Logic eNVM	1T1R E-Fuse [5]	2T Anti-Fuse [6]	5T Anti-Fuse [7]	2T Anti-Fuse [8]	10T Eflash [12]	3T Eflash [13]	C-Flash [15]	5T Eflash [16]	6T Eflash (This Work)
Process	32nm	32nm	65nm	0.18μm	0.18μm	65nm	0.18μm	65nm	65nm
Core Supply	1.0V	1.0V	1.2V	1.8V	1.2V	1.2V	1.8V	1.2V	1.2V
Acc. / Cell Dev.	1.8V I/O TR	1.8V I/O TR	3.3V I/O TR	3.3V I/O TR	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	None	None	None	None	7nm	5nm	7nm	5nm	5nm
Writing Method	Electro-Migration	Gate Oxide Breakdown	Gate Oxide Breakdown	Gate Oxide Breakdown	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Writing Voltage	1.9V	4.5V	6.5V	6.6V	10V	8V	5, -5V	10V	-7.2V
Bit-by-Bit Rewrite	No	No	No	No	Yes	No	No	No	Yes
Unsel. WL Disturb	No	No	No	No	Yes	Yes	Yes	No	No
Unit Cell Area	$1.37\mu m^2$	$1.01\mu m^2$	$15.3\mu m^2$	$4.88\mu m^2$	$220\mu m^2$	N. A.	$72\mu m^2$	$8.62\mu m^2$	$15.3\mu m^2$
CP Area	N. A.	N. A.	$0.0512mm^2$ (est.)	N. A.	N. A.	N. A.	N. A.	N. A.	$0.0214mm^2$
Macro Area	N. A.	N. A.	$0.244mm^2$	$0.133mm^2$	N. A.	N. A.	$0.0336mm^2$	$0.0859mm^2$	$0.165mm^2$
Capacity	4kb	1kb	8kb	2kb	192b	N. A.	256b	2kb	4kb

V. COMPARISON WITH PRIOR LOGIC-COMPATIBLE eNVM

Table I compares various eNVMs implemented in a standard logic process for moderate-density eNVM applications. Kulkarni *et al.* presented a 4 kb 1T1R e-fuse and a 1 kb 2 T anti-fuse OTP in a 32 nm logic process using 1.8 V I/O transistor [5], [6]. Permanent metal electro-migration and gate-oxide breakdown were used as the program method. Matsufuji *et al.* presented an 8 kb 5 T anti-fuse OTP memory in 65 nm logic technology using 3.3 V I/O transistors. The unique feature of this design is that the broken path can be tested using a 5 T cell structure [7]. 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 [10]–[17]. Feng *et al.*. proposed a bit-by-bit re-writable 192 b 10 T eflash in a 0.18 μm logic process using 3.3 V I/O transistors [12], but this cell structure suffers from the disturbance issue of the unselected WL cells. Chen *et al.* proposed a 3 T eflash which can be built in an advanced logic process [13], and Roizin *et al.* proposed a 256 b C-Flash in a 0.18 μm logic process using 3.3 V I/O transistor using a bipolar writing voltage (i.e. 5, -5 V) [14], [15]; however, these designs do not support a bit-by-bit re-write and the unselected WL cells suffer from disturbance issues. In [16], a 2 kb 5 T eflash was implemented in a 65 nm logic process using 2.5 V I/O transistor. Here, the unselected WL's are undisturbed [16]; 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 6 T eflash is the only bit-by-bit re-writable eNVM that eliminates disturbance in the unselected WL cells.

VI. CONCLUSION

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 [16] suffers from unnecessary erase and program cycles resulting in poor endurance characteristics. Previous bit-by-bit erasable eflash on the other hand [12] suffered from high voltage disturbance issues in the unselected WL's. In this work, we proposed a bit-by-bit re-writable 6 T 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 4 kb eflash test chip was demonstrated in a generic 65 nm logic process, confirming the functionality of the proposed techniques. The overall endurance was improved by $\sim$4 times compared to the prior WL-by-WL erasable 5 T eflash for a 2:1 column MUX configuration.

REFERENCES

[1] T. Kono *et al.*, "40 nm embedded SG-MONOS flash macros for automotive with 160 MHz random access for code and endurance over 10 M cycles for data," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2013, pp. 212–213.

[2] S. Bartling *et al.*, "An 8 MHz 75 μA/MHz zero-leakage non-volatile logic-based cortex-M0 MCU SoC exhibiting 100% digital state retention at VDD $=$ 0 V with $<$400 ns wakeup and sleep transitions," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2013, pp. 432–433.

[3] H. Yu *et al.*, "Cycling endurance optimization scheme for 1 Mb STT-MRAM in 40 nm technology," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2013, pp. 224–225.

[4] 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. Dig. Tech. Papers*, 2013, pp. 220–221.

[5] S. Kulkarni *et al.*, "A 4 kb metal-fuse OTP-ROM macro featuring a 2 V programmable 1.37 μm^2 1T1R bit cell in 32 nm high-k metal-gate CMOS," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 863–868, Apr. 2010.

[6] S. Kulkarni *et al.*, "A 32 nm high-k and metal-gate anti-fuse array featuring a 1.01 μm^2 1T1C bit cell," in *IEEE Symp. VLSI Technol. Dig.*, 2012, pp. 79–80.
[7] K. Matsufuji *et al.*, "A 65 nm pure CMOS one-time programmable memory using a two-port antifuse cell implemented in matrix structure," in *Proc. IEEE Asian Solid-State Circuits Conf.*, 2007, pp. 212–215.
[8] N. Phan, I. Chang, and J. Lee, "A 2 kb one-time programmable memory for UHF passive RFID tag IC in a standard 0.18 μm CMOS process," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 7, pp. 1810–1822, Jul. 2013.
[9] W. Chen *et al.*, "A 22 nm 2.5 MB slice on-die L3 cache for the next generation Xeon processor," in *IEEE Symp. VLSI Circuits Dig.*, 2013, pp. 132–133.
[10] J. Raszka *et al.*, "Embedded flash memory for security applications in a 0.13 μm CMOS logic process," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2004, pp. 46–47.
[11] L. Pan *et al.*, "Pure logic CMOS based embedded non-volatile random access memory for low power RFID application," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2008, pp. 197–200.
[12] 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 Integr. Circuits Conf.*, 2009, pp. 713–716.
[13] H. Chen, "Single Polysilicon Layer Non-Volatile Memory and Operating Method Thereof," U.S. Patent 8 199 578, Jun. 12, 2012, et al..
[14] Y. Roizin *et al.*, "C-flash: An ultra-low power single poly logic NVM," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop*, 2008, pp. 90–92.
[15] 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.
[16] S. Song, K. Chun, and C. H. Kim, "A logic-compatible embedded flash memory for zero-standby power system-on-chips featuring a multistory high voltage switch and a selective refresh scheme," *IEEE J. Solid-State Circuits*, vol. 48, no. 5, pp. 1302–1314, May 2013.
[17] S. Song, J. Kim, and C. H. Kim, "Program/erase speed, endurance, retention, disturbance characteristics of single-poly embedded flash cells," in *Proc. IEEE Int. Reliability Physics Symp.*, 2013, pp. MY.4.1–MY4.6.
[18] 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. VLSI Circuits Dig.*, 2012, pp. 134–135.
[19] 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. VLSI Technol. Dig.*, 2013, pp. 152–153.
[20] 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 Integr. Circuits Conf.*, 2013, pp. 1–4.
[21] 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.
[22] 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.
[23] 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.

Seung-Hwan Song received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, in 2004 and 2006, respectively, and the Ph.D. degree from University of Minnesota, Minneapolis, USA, in 2013, all in electrical engineering.

During his Ph.D. work, he interned with Broadcom and Qualcomm, designing the embedded memories and, at Seagate, evaluated the NAND flash memories for the SSD applications. From 2006 to 2009, he was with Samsung, performing research and development on the MLC/TLC/QLC NAND flash memories and their controllers. Currently, he is a Research Staff Member with the Storage Architecture Laboratory, HGST, San Jose, CA, USA. His recent research interests include embedded memory, nonvolatile memory, and storage systems.

Dr. Song was the recipient of ISLPED Low Power Design Contest Award in 2012 and graduate fellowship from the University of Minnesota in 2009.

Ki Chul Chun received the B.S. degree in electronics engineering from Yonsei University, Seoul, Korea, in 1998, the M.S. degree in electrical engineering from KAIST, Daejeon, Korea, in 2000, and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2012.

In 2000, he joined the Memory Division, Samsung Electronics, Gyeonggi-Do, Korea, where he has been involved in DRAM circuit design. After his Ph.D. work at the University of Minnesota, he rejoined Samsung Electronics in 2012, where he has worked for Low-Power DRAM development. His research interests include digital, mixed-signal and memory circuit designs with special focus on DRAM, PRAM, and STT-MRAM in scaled technologies.

Dr. Chun was the recipient of ISLPED Low Power Design Contest Awards (2009 and 2012) and a Samsung Ph.D. Scholarship.

Chris H. Kim (M'04–SM'10) received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA.

He spent a year with Intel Corporation, where he performed research on variation-tolerant circuits, on-die leakage sensor design, and crosstalk noise analysis. He joined the Electrical and Computer Engineering Faculty, University of Minnesota, Minneapolis, MN, USA, in 2004 where he is currently an Associate Professor. He is an author/coauthor of over 100 journal and conference papers His research interests include digital, mixed-signal, and memory circuit design in silicon and nonsilicon (such as organic TFT and spin) technologies.

Prof. Kim was the recipient of a National Science Foundation CAREER Award, a Mcknight Foundation Land-Grant Professorship, a 3M Non-Tenured Faculty Award, DAC/ISSCC Student Design Contest Awards, IBM Faculty Partnership Awards, an IEEE Circuits and Systems Society Outstanding Young Author Award, ISLPED Low Power Design Contest Awards, and an Intel Ph.D. Fellowship. He has served as a technical program committee chair for the 2010 International Symposium on Low Power Electronics and Design (ISLPED).