

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

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

Abstract—Embedded flash memory implemented using standard I/O devices can open doors to new applications and system capabilities, as it can serve as a secure on-chip non-volatile storage for VLSI chips built in standard logic processes. For example, it is indispensable for adaptive self-healing techniques targeted for mitigating process variation and circuit aging related issues where system information must be retained during power down periods. Embedded non-volatile memory can also enable zero-standby power systems by allowing them to completely power down without losing critical data. There has been numerous device and circuit level research on high-density non-volatile memories such as flash, 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 paper, a logic-compatible embedded flash memory that uses no special devices other than standard core and I/O transistors is demonstrated in a generic logic process having a 5 nm 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.

Index Terms—Embedded flash memory, embedded nonvolatile memory, multi-story high voltage switch, selective WL refresh, zero-standby power system-on-chip.

I. INTRODUCTION

ULTRA-LOW Power (ULP) systems such as wireless sensors and embedded microcontrollers typically employ power down modes (e.g., power gating) to minimize leakage during long standby periods [1], [2]. One important requirement during standby mode is to retain critical system information in an on-board non-volatile storage. Two approaches exist for this purpose as illustrated in Fig. 1. First option 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 due to the data retention voltage requirement. Second option is to use a re-writable embedded nonvolatile memory (eNVM) which allows the system to completely shut down without losing data thereby achieving zero-standby power dissipation. Among various eNVM candidates, embedded flash (eflash) memories based on floating gate technology have been successfully deployed in many SoC applications such as automotive microcontroller units (MCU) and smartcard ICs [2]–[7]. This approach can provide secure on-chip data and code storage without an on-board battery; however is only possible when a dedicated eflash process is available. This paper explores an application space that is different from traditional high-density non-volatile memories (e.g., flash, PRAM, emerging memories such as STT-MRAM and RRAM) where a moderate-density (e.g., few kilo-bytes) non-volatile memory built in a generic logic process can be used to enhance system reliability or enable zero standby power.

Prior to discussing the proposed circuits, we give an overview of high-density and moderate-density eflash memory candidates illustrated in Fig. 2. Kojima *et al.* presented a 1T dual-poly eflash that incurs additional process steps for forming the floating gate and thick oxide layer [4]. It utilizes the Channel Hot Electron (CHE) injection method for program operation which has a high power dissipation. Lee *et al.* presented a 2T dual-poly eflash utilizing Fowler-Nordheim (FN) tunneling for program operation to minimize program power dissipation [6], but the program voltage has to be increased to enable an efficient FN tunneling and a relatively high program inhibit voltage had to be applied to the unselected BL's during the program operation, dissipating significant dynamic power. Ikehashi *et al.* presented a 3T dual-poly eflash utilizing FN tunneling program and self-boosting technique for a program inhibit operation [7], but the required program and erase voltage levels were as high as 22 V due to the low coupling ratio between control gate (CG) and floating gate (FG). On the other hand, a charge trap based eflash technology was demonstrated in various literatures, as it can reduce the additional mask count [8], [9]. Later, Yater *et al.* presented the Split-Gate (SG) eflash that reduces the write voltage level and enhances the electron injection efficiency [10], but it still needs additional process steps such as SG

Manuscript received October 23, 2012; revised January 28, 2013; accepted January 29, 2013. Date of publication March 07, 2013; date of current version April 19, 2013. This paper was approved by Associate Editor Hideto Hidaka.

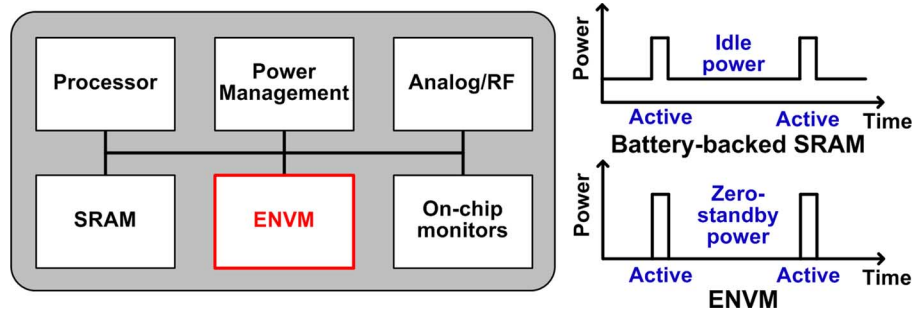
S.-H. Song is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: songx278@umn.edu).

K. C. Chun was with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. He is currently with the Memory Division, Samsung Electronics, Hwasung-si, 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.2013.2247691



Eflash	1T Dual-Poly [4]	2T Dual-Poly [6]	3T Dual-Poly [7]	Split Gate [10]	Single-Poly [16]	10T Single-Poly [19, 20]	3T Single-Poly [22]	5T Single-Poly (This work)
Unit Cell Schematic								
Process	90nm Eflash	90nm Eflash	0.4μm Eflash	90nm Eflash	0.13μm Logic	0.25μm/0.18μm/90nm/65nm Logic	65nm Logic	65nm Logic
Process Overhead	Floating Gate	Floating Gate	Floating Gate	Split Gate	None	None	None	None
Tunnel Oxide	10nm (Dedicated)	7nm (Dedicated)	N. A.	N. A.	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	CHE Injection	FN Tunneling	FN Tunneling	SS Injection	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Erase Method	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling	FN Tunneling
Write Voltage	±10V	16V	22V	14V	8V	8V	8V	5 to 10V
Write Power	High	Low~Medium	Low	Medium	Low	Low	Low	Low
Erase Disturb	None	None	None	None	None	(Write Disturb) Unselected WL's	Unsel. WL's	None
Measured Retention	>1k hours @ 250°C, EP 1k	>48 hours @ 250°C	N. A.	>1k hours @ 150°C, EP 10k	>1k hours @ 150°C, EP 1k	>6500 hours @ 85°C, EP 100k	N. A.	>486 hours @ 27°C, EP 10k
Cell Size	0.44μm ² (54F ²)	N. A.	4.36μm ² (27F ²)	N. A.	700μm ² (est.) (41000F ²)	N. A. *Similar 10T Cell [21]: ~220μm ²	N. A.	8.62μm ² (2111F ²)

Fig. 2. Single-poly and dual-poly eflash memory cells. Single-poly eflash can be built in a standard CMOS logic process and has a lower writing voltage owing to the flexible device sizing, making it attractive for moderate-density (e.g., few kilo-bytes) on-chip non-volatile storage. Note that single-poly eNVM is not a replacement for dual-poly eNVM or emerging memory (e.g., STT-MRAM, RRAM) due to their large cell size. Rather, it is targeted for applications that can benefit from having a moderate-density secure on-chip nonvolatile solution that can be built in a generic logic process (e.g., reliability and aging related failure mitigation, self-healing, zero-standby power).

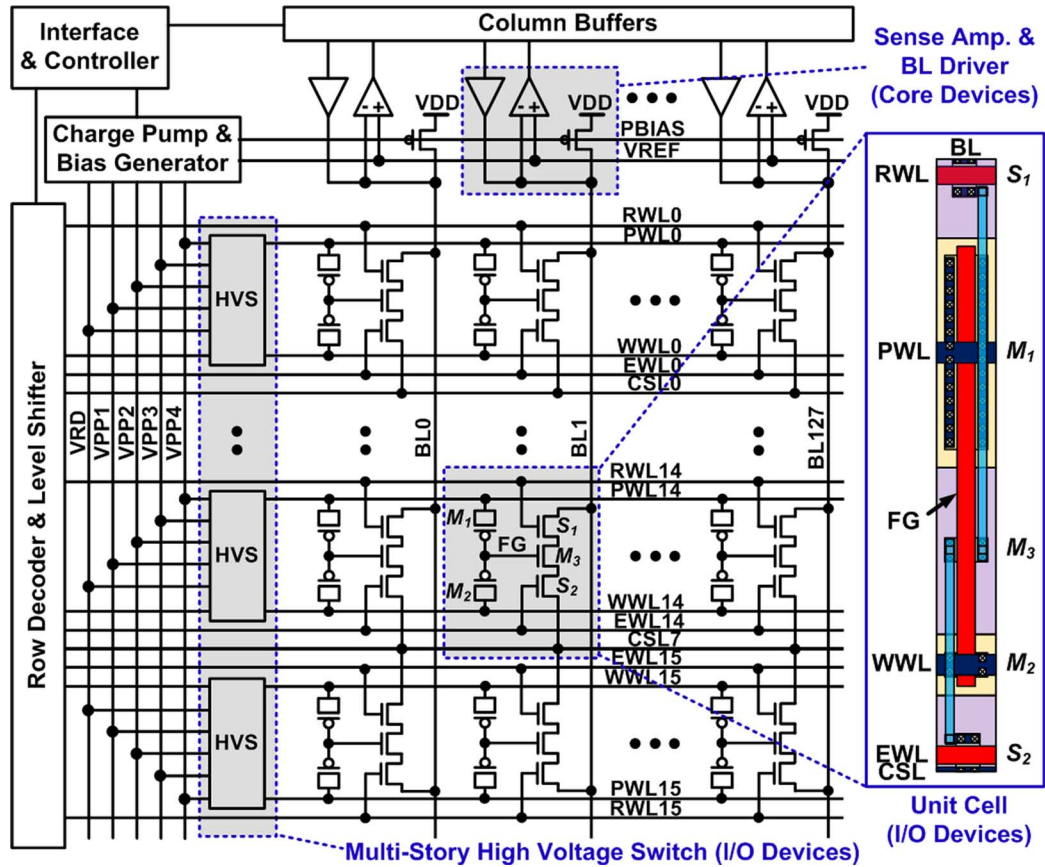


Fig. 3. Array architecture and unit cell layout of the proposed logic-compatible eflash memory. Only standard I/O and core devices are used.

to the requirement of floating gate cell transistors and high voltage devices. Moreover, achieving a high coupling ratio between CG and FG for effective program and erase operation involves process optimization well beyond what is needed for developing a standard CMOS logic process. Also, the program and erase voltage levels are typically greater than 10 V which increases the complexity of the high voltage generation and switching circuitry. 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 between CG (i.e., PWL and WWL in Fig. 2) and FG can be easily obtained by upsizing the width of the coupling transistor (M_1 in Fig. 2). 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 kilobytes) non-volatile storage in cases where a dedicated eflash process is not available [13]–[24]. 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 [16]–[21] to enhance the cell sensing margin at the expense of larger cell area, while several single cell eflash were proposed for a compact cell area [13]–[15], [22]–[24]. In this paper, we present a new 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. 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 eflash memory employs an overstress-free multi-story HVS capable of expanding the cell threshold voltage (V_{TH}) window. A selective row-by-row refresh scheme was also developed which improves the overall cell endurance limit. To the best of our knowledge, this is the first demonstration of a truly logic-compatible embedded flash memory, including fully functional peripheral circuits, based on standard logic I/O transistors with a tunnel oxide thickness of only 5 nm.

The remainder of this paper is organized as follows. Section II describes the memory architecture and cell operation as well as the high voltage switch and selective refresh scheme. Measurement results from test chips fabricated in a 65 nm low power CMOS process are presented in Section III. Section IV compares the proposed 5T single-poly eflash to the prior single-poly eflash, and conclusions are given in Section V.

II. PROPOSED LOGIC-COMPATIBLE SINGLE-POLY EFLASH

A. Overall Memory Architecture

The proposed single-poly 5T eflash memory architecture and the unit cell layout are shown in Fig. 3. All five transistors (M_1 , M_2 , M_3 , S_1 , S_2) in the unit cell are implemented using standard 2.5 V I/O transistors with a tunnel oxide thickness (T_{OX}) of 5 nm. Here, M_1 is the coupling device, M_2 is the

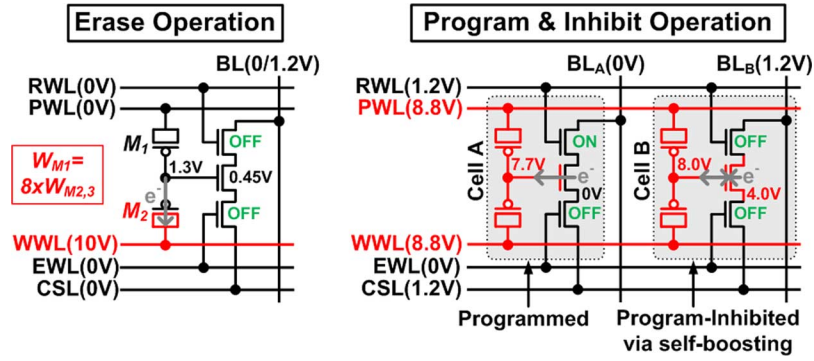


Fig. 4. Bias conditions in the selected WL for erase, program, and program inhibit operations of the proposed eflash. A single WL write operation ensures that unselected WL's are protected from the high voltage levels.

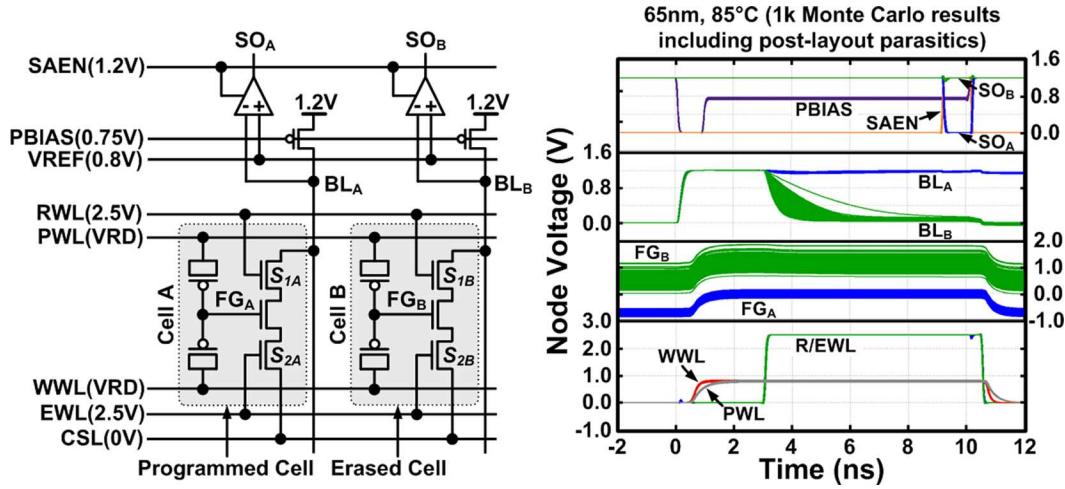


Fig. 5. Bias condition and simulated timing diagram during read mode. Waveforms are from 1 k Monte Carlo runs using a post-layout extracted netlist.

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. The gate terminals of the three transistors 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 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 used as CG (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 1.2 V thin T_{OX} core transistors while the high voltage switch in the WL driver is implemented using standard 2.5 V 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.

B. Proposed 5T Eflash Cell Operation

Cell bias conditions for erase and program operation of the proposed eflash cell are shown in Fig. 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 0 V. 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. 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 13 V so it can reliably support a 10 V erase operation. During program mode, a high voltage is applied to both the PWL and WWL of the selected row while self-boosting of the localized electron channel of the read device (M_3 in Fig. 3) prevents the cells of the unselected columns from being programmed by turning off the two select transistors (S_1 and S_2 in Fig. 3) in the unselected columns [25], [26]. The coupling ratio of PWL and WWL to FG during program operation is approximately 0.9. A row-by-row 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 [20], [22]. 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. 5. The extracted WL/BL parasitic capacitances and resistances are included in this Monte Carlo simulation. For the read

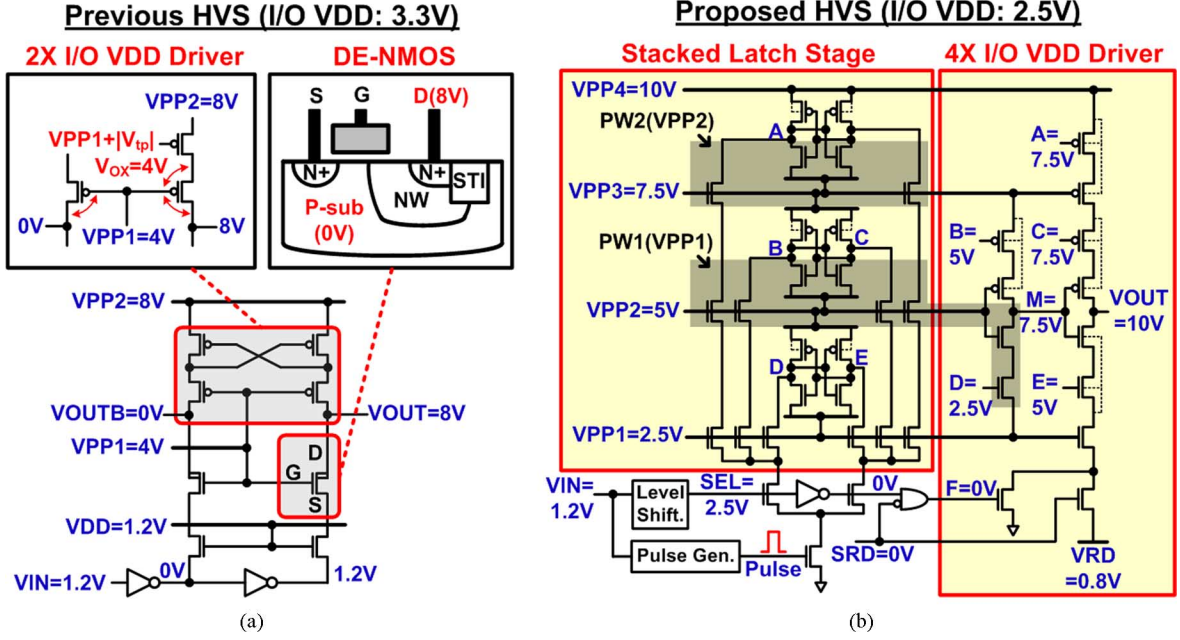


Fig. 6. (a) Previous HVS circuit has a limited output voltage ($V_{OUT} < 8V$) even with a $0.7V$ voltage overstress [16]. Moreover, the internal node voltage in the PMOS cascode is sensitive to threshold voltage variations. (b) Proposed HVS increases V_{OUT} up to $10V$ and provides robust internal voltage levels by utilizing multi-story latches and additional VPP supplies.

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, VREF, using sense amplifiers to produce digital output signals SO_A and SO_B . VREF 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 II-D). Note that the WL/BL lengths of the 2 kb eflash array in this work are $120\mu m$ and $200\mu m$, respectively (see Fig. 16), 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 [27].

C. Multi-Story High-Voltage WL Driver

Fig. 6 compares a prior HVS [16] 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. 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 (V_{OUT}) levels

by utilizing multi-story latches with additional VPP supplies. The four boosted supply levels (V_{PP1} - V_{PP4}) can be generated from an on-chip charge pump [16], [28], [29] with the highest voltage V_{PP4} 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 [16] 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 (V_{IN}) 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 Fig. 7. During program operation, PWL/WWL pulses are applied to the selected row consisting of 128 cells (Fig. 7(a)). Simulated current and voltage waveforms of the four boosted supplies are shown when PWL/WWL signal levels switch at $t = 3\mu s$ and $t = 5\mu s$ (Fig. 7(b) - (c)). Parasitics were extracted from the array layout for accuracy. When SEL switches from low to high for a program operation (Fig. 7(b)), nodes A, B, D, and F are discharged to V_{PP3} , V_{PP2} , V_{PP1} , and $0V$, respectively, while node C and E are pulled up to V_{PP3} and V_{PP2} . As

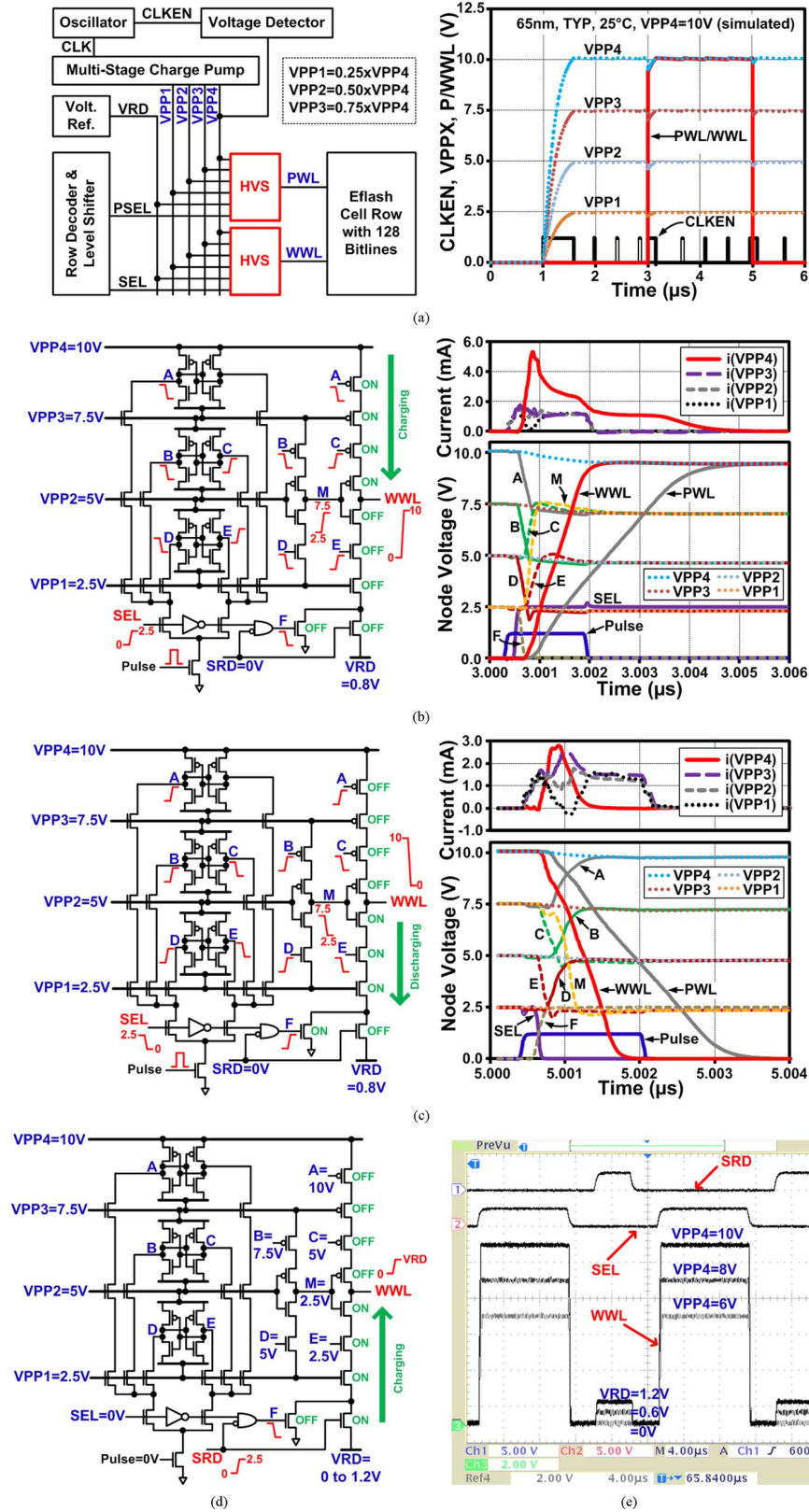


Fig. 7. Basic operation of the proposed HVS. (a) VPP levels simulated using a voltage doubler based on-chip charge pump [16], [28], [29]. (b), (c) Current and voltage waveforms during low-to-high and high-to-low transitions of WWL. (d) Low-to-high transition during read mode. (e) Measured waveforms of the proposed HVS for three different read and write voltage levels. Since we did not include the charge pump in the test chip, off-chip voltage supplies were used during chip measurements ($V_{RD} = 0/0.6/1.2$ V, $V_{PP4} = 6/8/10$ V, $V_{PP3} = 0.75 \times V_{PP4}$, $V_{PP2} = 0.5 \times V_{PP4}$, $V_{PP1} = 0.25 \times V_{PP4}$).

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 6 ns

which is significantly shorter than the usual program pulse width ($\sim 10 \mu s$). When SEL switches from high to low (Fig. 7(c)), the opposite transition occurs wherein WWL switches to 0 V

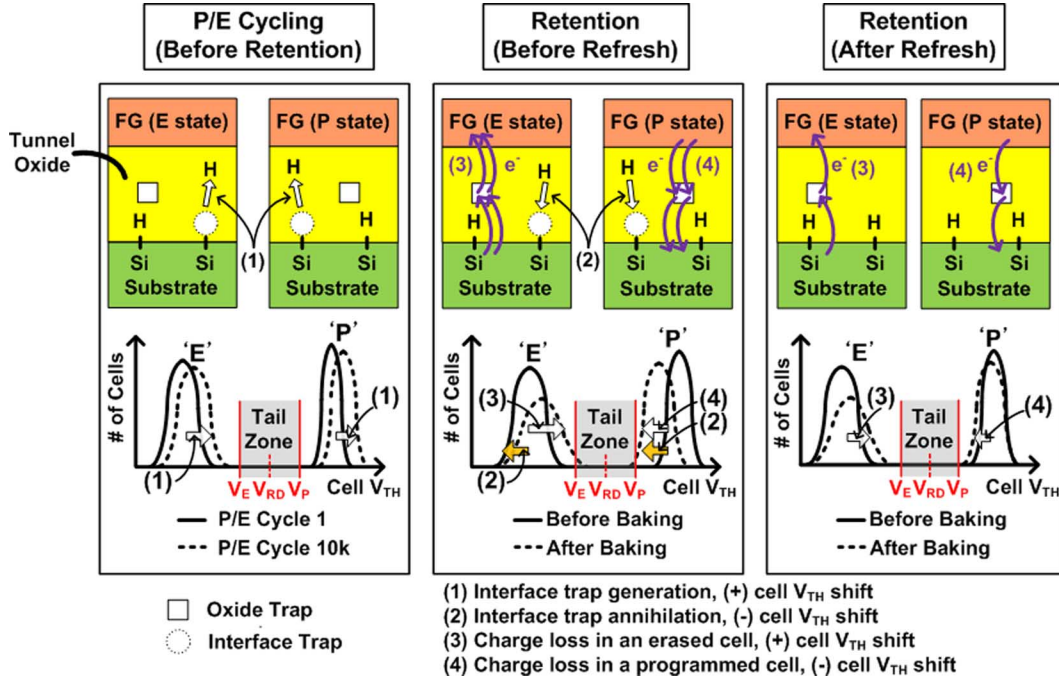


Fig. 8. Physical model explaining endurance and retention characteristics considering oxide and interface trap creation, interface trap annihilation, and trap-assisted charge loss [30]–[36]. The oxide and interface traps generated during P/E cycling contribute to a positive cell V_{TH} shift. During retention, some of the interface traps are restored, causing a negative cell V_{TH} shift, while the remaining oxide and interface traps facilitate the trap-assisted charge loss process. After a refresh operation, the charge loss rate becomes slower, improving the cell retention characteristic.

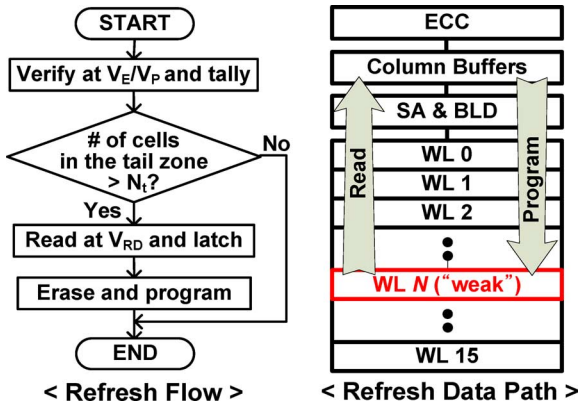


Fig. 9. Proposed selective WL refresh scheme. “Weak” WL’s are identified and selectively refreshed to avoid unnecessary P/E cycles in the good cells.

through the stacked NMOS transistors. When SRD is activated for read operation (Fig. 7(d)), the bottom NMOS in the output stage turns on making WWL switch to the read voltage V_{RD} . 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 (Fig. 5). Measured waveforms of the proposed HVS for three different read and write voltage levels are shown in Fig. 7(e). A charge pump circuit [16], [28], [29] was implemented for obtaining the simulation results in Fig. 7(a) – (d), 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 ($= 10$ V) compared to prior designs (e.g., 8 V in [16]), the cell V_{TH} window can be improved by more than 170% using the proposed HVS. The

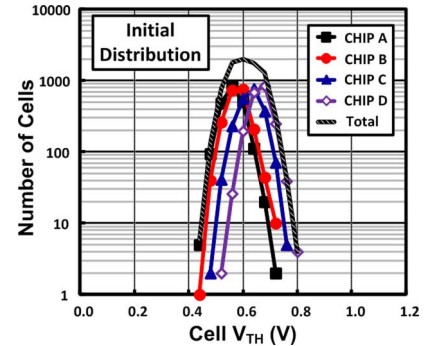


Fig. 10. Measured initial cell V_{TH} distributions of four 2 kb eflash memory chips show cell-to-cell and chip-to-chip variations. The initial cell distribution ranges from 0.44 to 0.80 V with an average value of 0.61 V.

V_{TH} of the eflash cell can be programmed to higher than 1.6 V in 10 μ s or erased to lower than -0.3 V in 1 ms without resulting in oxide reliability problems in the HVS as verified in Fig. 12(a). The proposed HVS operates reliably for a wide range of read voltages (from 0 to 1.6 V) and write voltages (from 5 to 10 V).

D. Selective WL Refresh Scheme

Previous literatures have reported that when a flash cell is programmed and erased repetitively, trap sites are created inside the tunnel oxide or oxide-silicon interface causing instability in the cell V_{TH} [30]–[36]. 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. 8 (left). According to the interface trap annihilation model [31], [35], interface traps created as a

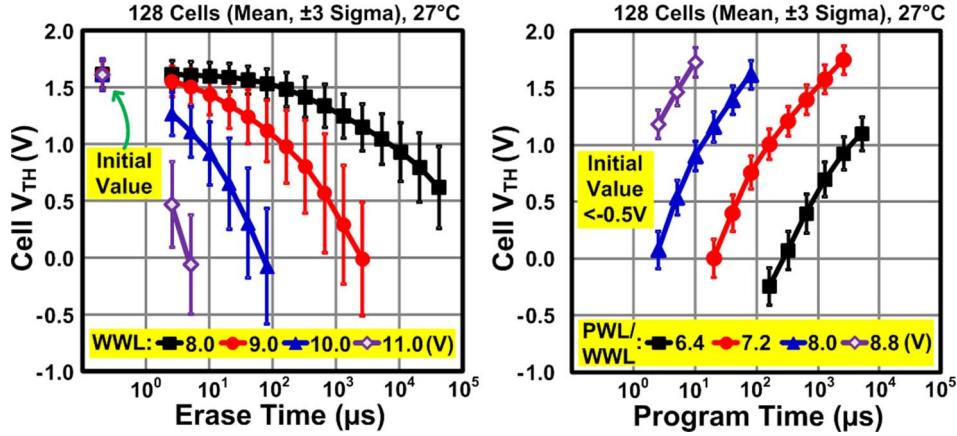


Fig. 11. Measured cell V_{TH} for different P/E voltages and pulse durations. Note that WWL and PWL were supplied by an off-chip voltage source to eliminate any non-ideal effects.

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. 8 (center). The oxide and interface trap sites are believed to facilitate the Trap-Assist-Tunneling (TAT) phenomena [33]–[35], which in turn accelerates the charge loss from the silicon substrate and from the floating gate resulting in positive and negative cell V_{TH} shifts for the erased and programmed cells, respectively.

Unlike stand-alone 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 [37], an intermediate refresh is conceivable for embedded 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 [31], [35], the trap assisted charge loss becomes smaller after a refresh operation as described in Fig. 8 (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 [37]–[40].

The proposed selective WL refresh scheme illustrated in Fig. 9 identifies “weak” WL’s by keeping track of the number of cells falling into the tail zone (Fig. 8). 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 [41] 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 (5 nm) logic process. This is in contrast to the previous design built using dedicated thick oxide floating gate devices [41] 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. 14 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.

III. TEST CHIP MEASUREMENT RESULTS

A. Initial Cell V_{TH} , and Writing Speed

To demonstrate the proposed circuit techniques, a 2 kb eflash memory was implemented in a 65 nm low power logic process. To measure the cell V_{TH} , we simultaneously swept the PWL and WWL voltage levels while checking whether the sensed data has flipped. Fig. 10 shows the initial cell V_{TH} distributions

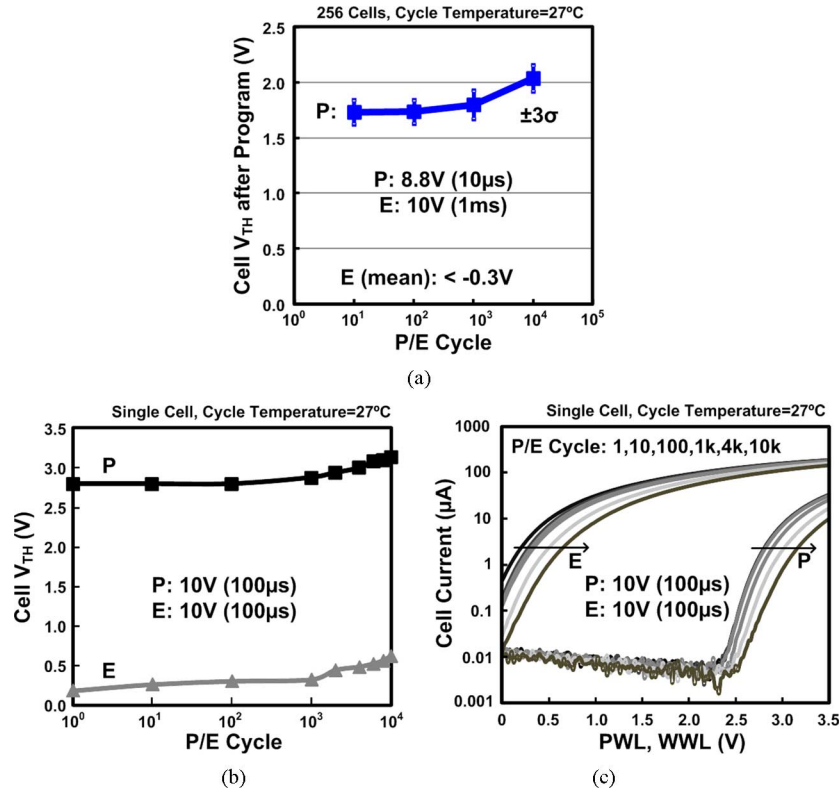


Fig. 12. Measured endurance characteristic. (a) Programmed cell V_{TH} shows a positive shift for P/E cycles greater than 1 k. (b) Erased cell V_{TH} shows a similar positive shift for P/E cycles greater than 1 k. (c) Cell current measured from a single cell exhibits severe sub-threshold slope degradation beyond 1 k P/E cycles, implying that a considerable number of interface traps have been generated [32].

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 8 kb cells has an average value of 0.61 V and a 3-sigma value of 0.18 V.

The measured erase and program speeds for different voltage levels are shown in Fig. 11. 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: 9 V, Program: 8.8 V). 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 our proposed single cell configuration [13]–[15], [22]–[24] is roughly $1000\times$ faster than that of a dual cell configuration [16]–[21] where the erase operation in one of the cells always limits the performance.

B. Endurance and Retention

Fig. 12 shows the measured endurance characteristics of the proposed eflash cells. P/E pre-cycling up to 10 k cycles was performed at room temperature (27°C). Fig. 12(a) shows results for an 8.8 V/10 μ s program pulse and a 10 V/1 ms erase pulse, while Fig. 12(b) and (c) are for 10 V/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 1 k, the programmed cell V_{TH} starts to shift in the positive direction as shown in Fig. 12(a). The erased cell

V_{TH} shows a similar positive shift for P/E cycles greater than 1 k as shown in Fig. 12(b). The cell current measured from a single cell test structure (Fig. 12(c)) shows a severe degradation in the sub-threshold slope with increased P/E cycles, 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 [32]. All the graphs show that a cell V_{TH} window greater than 1.9 V is achieved for up to 10 k P/E cycles.

Fig. 13 shows the measured retention characteristic of the proposed eflash cells. Fig. 13(a) and (b) show that a sufficient sensing margin is maintained at a 150°C bake temperature for the 1 k and 10 k pre-cycled cells, respectively. Fig. 13(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 10 k. 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 16 M flash memory array did not show any spatial correlation [33]. Fig. 13(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 exhibit a larger V_{TH} shift. The cell V_{TH} shifts for each of P/E pre-cycle cases do not show a strong correlation with the initial cell V_{TH} values. Fig. 14 shows the evolution of the tail cell V_{TH} for different baking temperatures. For P/E pre-cycling, an 8.8 V/10 μ s program pulse and an 8.8 V/10 ms erase pulse were repetitively applied to three chips at room temperature (27°C). Then, the three programmed chips

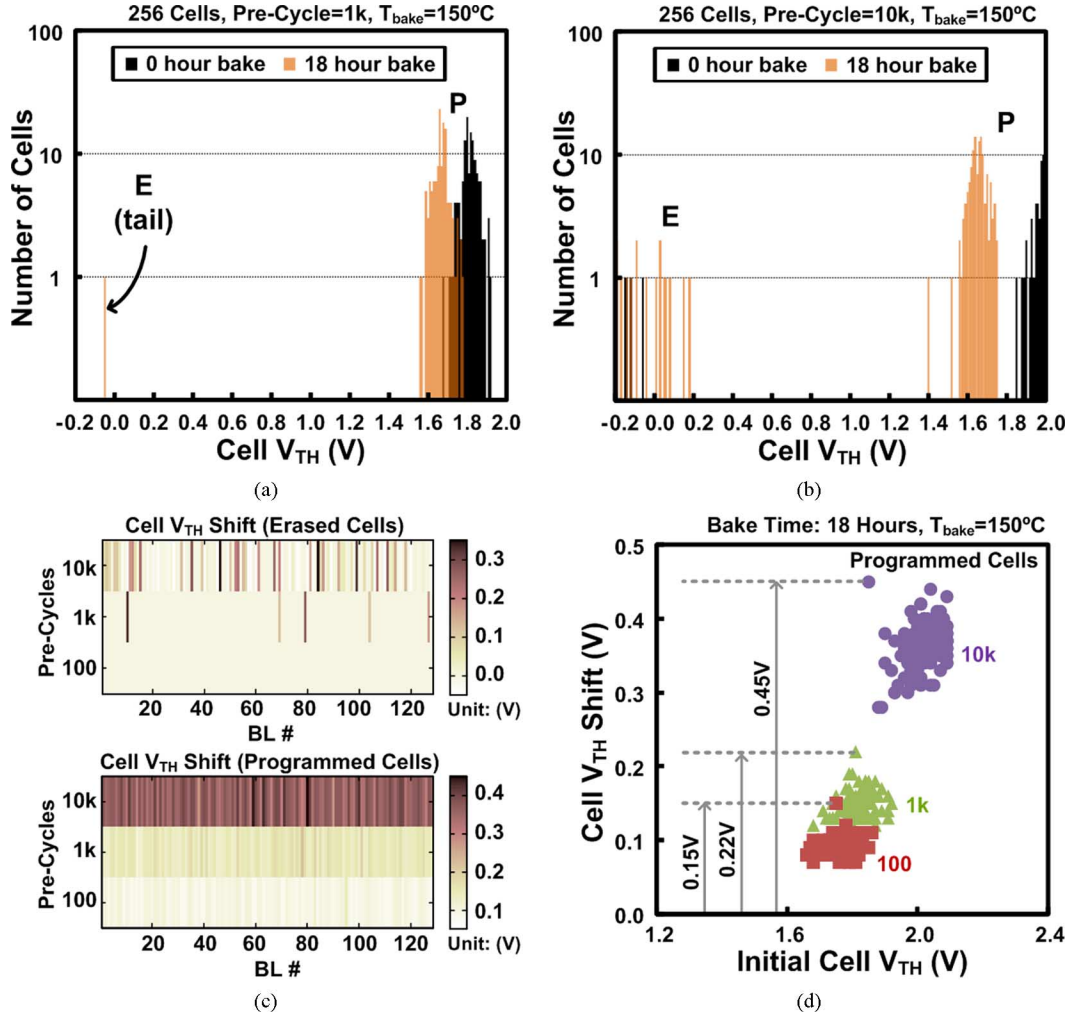


Fig. 13. Measured retention characteristic. (a, b) Cell V_{TH} distributions for 1 k and 10 k P/E pre-cycled cells at a 150°C bake temperature. (c) Spatial bit maps show the cell V_{TH} shift of erased and programmed cells after 100/1 k/10 k P/E pre-cycles. (d) Cell V_{TH} shift vs. initial cell V_{TH} level for 100/1 k/10 k pre-cycles.

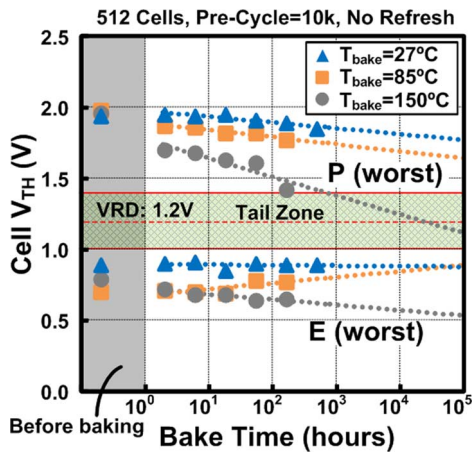


Fig. 14. Measured retention characteristic as a function of baking temperature. Three chips were baked at 27/85/150°C, respectively. The effects of charge loss and interface trap annihilation are canceled out for the erased cells, while a negative cell V_{TH} shift is observed in the programmed cells [35].

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. 8. 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 as explained in Fig. 8 [35].

C. Effectiveness of Refresh Operation

Fig. 15 shows the cell retention characteristics of 256 cells with 10 k 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 8.8 V/10 μ s program pulse and a 10 V/1 ms erase pulse. All cells experience the same program and erase pulses during pre-cycling. A read reference voltage of 1.0 V 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 (1.13 V \rightarrow 1.38 V) compared to the pre-refresh 54 hour bake results as shown in Fig. 15(a). Projections based on the retention time of 10 k pre-cycled cells (Fig. 15(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.

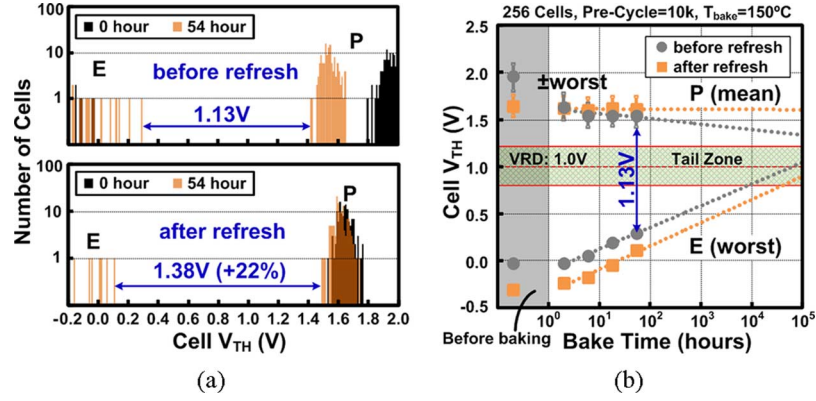


Fig. 15. (a) Cell V_{TH} distribution and (b) retention characteristics before and after refresh at a retention temperature of 150°C. Post-refresh retention characteristic shows an enhanced cell V_{TH} margin and an increased retention time.

TABLE I
SINGLE-POLY EFLASH COMPARISON

Single-Poly EFLASH	VLSI 2000 [13]	CICC 2001 [14]	ISSCC 2004 [16]	IEICE 2007 [17]	NVSMW 2008 [23]	This Work
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

IV. COMPARISON WITH OTHER SINGLE-POLY EFLASH

Table I 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 [13]–[15]. To utilize the higher program efficiency of the FN tunneling program method, Raszka *et al.* utilized 3.3 V I/O devices with a 7 nm tunnel oxide for the eflash cell [16]. The typical current for simultaneously programming 2 kb cells via FN tunneling was reported as around 1 μ A [16]. The HVS in their work uses special DE-NMOS and cascading stages. However, the cell V_{TH} window was limited to 0.7 V even though devices in the HVS experience a voltage overstress. Yamamoto *et al.* also utilized 3.3 V I/O devices with a 7.6 nm tunnel oxide in the memory cell [17], [18]. The cell V_{TH} window in this work was around 1.8 V. These two prior eflash memories employ dual cell architectures to boost the sensing margin but this slows down the program speed by $\sim 1000\times$ as explained in

Section III-A (Fig. 11). Moreover, each unit cell included a dedicated HVS to resolve the write disturbance issue in the unselected WL's (Fig. 2) [19]–[22]; however this significantly increases the memory footprint compared to other single-poly eflash memories shown in Table I. Later, Roizin *et al.* presented a single-poly eflash with a single cell architecture and FN tunneling programming using 3.3 V I/O devices having a 7 nm tunnel oxide in a 0.18 μ m logic process [23].

In contrast, the proposed 5T eflash was successfully implemented in a 65 nm low power standard CMOS logic process where the I/O devices have a 5 nm tunnel oxide. Despite the HVS being overstress free, the proposed multi-story high voltage WL driver achieves a cell V_{TH} window greater than 1.9 V as shown in Section III-B (Fig. 12) by allowing the WL voltage to be raised to 3 or 4 times the I/O voltage during erase and program operation. The proposed 5T eflash employs single cell architecture for fast program operation and all 128 cells connected to a single WL are accessed simultaneously

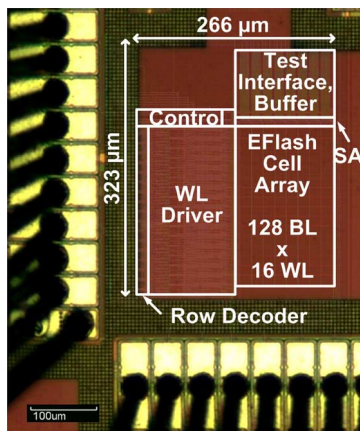


Fig. 16. Die photograph of 2 kb eflash test chip.

improving overall throughput and simplifying the refresh sequence as shown in Section II-D (Fig. 9). Unselected WL's are protected from the high erase and program voltage through the single WL erase and program architecture. This feature helps improve overall memory endurance. 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 [16]–[18]. Compared to prior single cell implementations [13], [14], [23], our proposed cell is 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 65 nm eflash test chip is shown in Fig. 16.

V. CONCLUSION

Although single-poly eflash are not suitable for high-density NVM applications due to their large cell size, they can be useful in a range of applications where a few kilo bytes 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 work, we proposed and experimentally demonstrated a logic-compatible eflash memory in a 65 nm 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 10 k P/E cycles.

REFERENCES

- [1] S. Hanson *et al.*, "A low-voltage processor for sensing applications with picowatt standby mode," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1145–1155, Apr. 2009.
- [2] H. Hidaka, "Evolution of embedded flash memory technology for MCU," in *IEEE Int. Conf. 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 90 nm 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 μm 2.125 MB 23.5 ns embedded flash with 2 GB/s read throughput for automotive microcontrollers," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2007, pp. 478–479.
- [6] Y. Lee *et al.*, "2T-FN eNVM with 90 nm logic process for smart card," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSIMW)*, 2008, pp. 26–27.
- [7] T. Ikehashi *et al.*, "A 60 ns access 32 kByte 3-transistor flash for low power embedded applications," in *IEEE Symp. 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 (NVSIMW)*, 2006, pp. 15–16.
- [9] M. Fliesler, D. Still, and J. Hwang, "A 15 ns 4 Mb NVSRAM in 0.13 μm SONOS technology," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSIMW)*, 2008, pp. 83–86.
- [10] J. Yater *et al.*, "16 Mb split gate flash memory with improved process window," in *Proc. IEEE Int. Memory Workshop (IMW)*, 2009, pp. 1–2.
- [11] Y. Yano, "Take the expressway to go greener," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 24–30.
- [12] 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. (ISSCC) Dig. Tech. Papers*, 2013, pp. 212–213.
- [13] R. McPartland and R. Singh, "1.25 volt, low cost, embedded flash memory for low density applications," in *IEEE Symp. VLSI Circuits Dig.*, 2000, pp. 158–161.
- [14] S. Shukuri, K. Yanagisawa, and 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.
- [15] M. Yamaoka *et al.*, "A system LSI memory redundancy technique using an IE-flash (inverse-gate-electrode flash) programming circuit," *IEEE J. Solid-State Circuits*, vol. 37, no. 5, pp. 599–604, May 2002.
- [16] 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. (ISSCC) Dig. Tech. Papers*, 2004, pp. 46–47.
- [17] 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.
- [18] Y. Yamamoto *et al.*, "Nonvolatile semiconductor memory device," US Patent 7,755,941, Jul. 13, 2010.
- [19] Y. Ma *et al.*, "Floating-gate nonvolatile memory with ultrathin 5 nm tunnel oxide," *IEEE Trans. Electron Devices*, vol. 55, no. 12, pp. 3476–3481, Dec. 2008.
- [20] A. Pesavento, F. Bernard, and J. Hyde, "PFET nonvolatile memory," US Patent 7,221,596, May 22, 2007.
- [21] 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.
- [22] H. Chen *et al.*, "Single polysilicon layer non-volatile memory and operating method thereof," US Patent 8,199,578, Jun. 12, 2012.
- [23] Y. Roizin *et al.*, "C-flash: An ultra-low power single poly logic NVM," in *Proc. IEEE Non-Volatile Semiconductor Memory Workshop (NVSIMW)*, 2008, pp. 90–92.
- [24] 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. VLSI Circuits Dig.*, 2012, pp. 130–131.
- [25] T. Jung *et al.*, "A 117 mm² 3.3 V only 128 Mb multilevel NAND flash memory for mass storage applications," *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1575–1583, Nov. 1996.
- [26] K. Suh *et al.*, "A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, no. 11, pp. 1149–1156, Nov. 1995.
- [27] M. Jefremow *et al.*, "Bitline-capacitance-cancellation sensing scheme with 11 ns read latency and maximum read throughput of 2.9 GB/s in 65 nm embedded flash for automotive," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 428–429.

- [28] 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.
- [29] 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.
- [30] 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.
- [31] A. Hdiy, G. Salace, and C. Petit, "Relaxation of interface states and positive charge in thin gate oxide after Fowler-Nordheim stress," *AIP J. Appl. Phys.*, vol. 73, no. 7, pp. 3569–3570, Apr. 1993.
- [32] Y. Park and D. Schroder, "Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM," *IEEE Trans. Electron Devices*, vol. 45, no. 6, pp. 1361–1368, Jun. 1998.
- [33] Y. Manabe *et al.*, "Detailed observation of small leak current in flash memories with thin tunnel oxides," *IEEE Trans. Semicond. Manufact.*, vol. 12, no. 2, pp. 170–174, May 1999.
- [34] R. Bez *et al.*, "Introduction to flash memory," *Proc. IEEE*, vol. 91, no. 4, Apr. 2003.
- [35] J. Lee *et al.*, "Effects of interface trap generation and annihilation on the data retention characteristics of flash memory cells," *IEEE Trans. Device Mat. Rel.*, vol. 4, no. 1, pp. 110–117, Mar. 2004.
- [36] N. Mielke *et al.*, "Flash EEPROM threshold instabilities due to charge trapping during program/erase cycling," *IEEE Trans. Device Mat. Rel.*, vol. 4, no. 3, pp. 335–344, Sep. 2004.
- [37] Y. Pan *et al.*, "Quasi-nonvolatile SSD: Trading flash memory non-volatility to improve storage system performance for enterprise applications," in *Proc. IEEE Int. Symp. High Performance Comput. Architecture (HPCA)*, 2012, pp. 1–10.
- [38] 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.
- [39] C. Miccoli *et al.*, "Assessment of distributed-cycling schemes on 45 nm NOR flash memory arrays," in *Proc. IEEE Int. Reliability Physics Symp. (IRPS)*, 2012, pp. 2A.1.1–2A.1.7.
- [40] S. Tanakamaru, Y. Yanagihara, and K. Takeuchi, "Over-10x-extended-lifetime 76%-reduced-error Solid-State Drives (SSDs) with error-prediction LDPC architecture and error-recovery scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 2012, pp. 424–425.
- [41] A. Umezawa *et al.*, "A new self-data-refresh scheme for a sector erasable 16 Mb flash EEPROM," in *IEEE Symp. VLSI Circuits Dig.*, 1993, pp. 99–100.



Seung-Hwan Song (S'12) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree in electrical engineering at University of Minnesota, Minneapolis, MN, USA.

In 2006, he joined Samsung Advanced Institute of Technology, Yongin, Korea, where he performed research on vertical/MLC NAND flash memory. In 2008, he transferred to Memory Division, Samsung Electronics, Hwasung, Korea, where he was

involved in the development of flash controller and software algorithm. During his Ph.D. study, he was an intern at Broadcom Corporation, Edina, MN, USA, where he was involved in high voltage circuit designs for the embedded OTP memory IP in 2012. His research interests include low power memory, power management, and mixed signal circuits. He is an inventor or co-inventor of more than 35 issued international patents.

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



Ki Chul Chun (M'11) 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. study at the University of Minnesota, he rejoined Samsung Electronics in 2012, where he has worked in 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 in 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 at 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 at the University of Minnesota, Minneapolis, MN, USA, in 2004 where he is currently an Associate Professor. His research interests include digital, mixed-signal, and memory circuit design in silicon and non-silicon (such as organic TFT and spin) technologies.

Prof. Kim was the recipient of an NSF CAREER Award, a McKnight Foundation Land-Grant Professorship, a 3 M 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 is an author/coauthor of 100+ journal and conference papers and has served as a technical program committee chair for the 2010 International Symposium on Low Power Electronics and Design (ISLPED).