A Logic-Compatible Embedded Flash Memory Featuring a Multi-Story High Voltage Switch and a Selective Refresh Scheme

Seung-Hwan Song, Ki Chul Chun, and Chris H. Kim

Dept. of ECE, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA (Email: songx278@umn.edu)

Abstract

A logic-compatible embedded flash memory that uses no special devices other than standard core and IO transistors is demonstrated in a low-power standard logic process having a 5nm tunnel oxide. An overstress-free high voltage switch expands the cell V_{TH} window by >170% while a 5T embedded flash memory cell with a selective row refresh scheme is employed for improved endurance.

Introduction

Low-cost and moderate-density Embedded Non-Volatile Memories (eNVMs) are expected to play a critical role in managing circuit variability and reliability effects in future microprocessors. For example, adaptive self-healing techniques or on-line repair schemes would require a moderate amount of eNVM for storing diagnostic data during periods when the chip is not powered. Table I compares different types of embedded flash memory technologies. Dual-poly flashes have been used in many system-on-chip applications for high-density nonvolatile storage, but this technology requires significant modification to the standard logic process as special thick oxide layer and floating gate formation process are required in the High Voltage Switch (HVS) and storage cell [1, 2]. A single-poly based embedded flash was previously demonstrated using special drain extended devices and cascoding techniques [3]; however, the operating margin was limited even with a dual cell configuration due to gate oxide reliability concerns in the HVS. In this paper, we present a logic-compatible embedded flash memory that uses no special devices other than the standard core and IO transistors. The proposed design employs an overstress-free multi-story HVS for expanding the cell V_{TH} margin and a 5T cell with a selective row refresh capability for effectively compensating the charge loss by the stress induced leakage current (SILC). To the best of our knowledge, this is the first demonstration of a truly logic-compatible embedded flash memory including fully functional peripheral circuits using a tunnel oxide thickness of only 5nm.

Flash Cell Operation

The proposed 5T embedded flash memory architecture and cell bias conditions are shown in Figs. 1 and 2, respectively. All five transistors (M_1, M_2, M_3, S_1, S_2) in the cell are implemented using standard 2.5V IO transistors with a T_{ox} of 5nm. The width of M_1 is 8 times wider than that of M_2 and M_3 achieving a high coupling ratio for effective erase and program operation. PMOS transistors are utilized for M_1 and M_2 to achieve high speed programming by biasing the devices in a non-depletion mode. The n-wells used as control gates (i.e. PWL and WWL) are shared in the WL direction attaining a tight bit-line pitch. During erase operation, a pulse is applied to the selected WWL while PWL is biased at 0V. The large gate capacitance of the upsized M_1 generates a high electric field in the gate oxide of M_2 removing electrons from FG through Fowler-Nordheim tunneling; therefore, the single WL erase operation is achieved without a negative boosted WL. During program operation, a high voltage is applied to the cells to be programmed while self-boosting [4] inhibits program in the unselected bitlines by turning off pass transistors S_1 and S_2. During read, the pass transistors are activated and the BL voltage levels are compared to a reference voltage, VREF. Note that a separate erase device (M_2) along with the self-boosting technique allows the column peripheral circuits to be built using low voltage core devices which improves operating speed and reduces power consumption.

High Voltage Switch for Enhanced Signal Margin

The proposed multi-story HVS in Fig. 3 is also implemented using standard IO transistors. Four boosted supply levels (VPP1-4) are generated from charge pumps with the highest level VPP4 being 3 to 4 times the nominal IO voltage. When SEL switches from low to high, nodes A, B and D are discharged to VPP3, VPP2 and VPP1, respectively, while node C and E are charged to VPP3 and VPP2. As a result, WWL is connected to VPP4 through the stacked PMOS transistors in the output stage. On the contrary, when SEL switches from high to low, WWL is connected to GND or VRD through the NMOS stack. An internally generated pulse from the address transition detector changes the state of the 3 stacked latches. The pulse width is kept short to minimize static power consumption. All transistors in the multi-story HVS operate within the nominal supply range of 2.5V. Deep n-well layers were used sparingly to minimize area overhead while keeping drain to body voltages of all transistors in the HVS to less than 5V, which is just half of the junction breakdown voltage for this process. Compared to a cascode circuit [3], the internal voltage levels of the proposed HVS are more robust as they are determined by the VPP levels rather than a transistor V_t drop. Using a VPP4 level that is 25% higher compared to [3], the memory cells V_{TH}'s can be programmed to >1.6V in 10μs or erased to <-0.3V in 1ms (Fig. 5) without causing oxide reliability problems in the HVS. As a result, the cell V_{TH} window of the proposed design can be improved by >170% without introducing a dual cell architecture which incurs a 2X cell area penalty [3]. Furthermore, program speed of our single-ended cell is ~1000X faster than a dual cell, because the speed of the latter is limited by the slow erase operation occurring in one of the dual cells.

Selective Refresh Scheme for Increased P/E Cycles

Unlike in stand-alone flashes where a strict retention time specification must be met, an intermediate refresh is permissible for embedded applications that have to support a higher number of P/E cycles throughout the product lifetime. In this work, a refresh scheme is proposed for compensating the charge loss and improving the cell retention characteristics. As shown in Fig. 4 (above), both trap annihilation and charge loss process contribute to the cell V_{TH} degradation during retention mode [5] with the latter being accelerated by the traps generated during P/E cycling. To take advantage of the fact that the cell V_{TH} degradation due to charge loss decreases after sufficient trap annihilation, we refresh flash cells that fall into the tail zone (Fig. 4 below/left). This enhances the post-refresh retention characteristics. Details of the proposed row-by-row selective refresh scheme are given in Fig. 4 (below). Unlike [6] where the entire memory array is refreshed, the proposed refresh operation is applied to the "weak" WL's only which prevents the good WL's from being stressed unnecessarily. Weak WL's are identified through two read verify operations using reference voltage levels, V_E and V_P. Once the number of cells in the tail zone exceeds a target value (N_t) determined by the ECC capability, the weak WL is first erased after the data is temporarily stored in a column buffer. Subsequently, the original or ECC-repaired data is re-programmed into the WL.

Test Chip Results

A 2kb embedded flash memory was implemented in a 65nm low power standard CMOS logic process for concept demonstration. Fig. 5 shows the measured cell V_{TH} distributions with cell endurance characteristic as well as the P/E pre-cycle dependency on cell retention. A noticeable interface trap-up is observed for larger P/E cycles. Fig. 6 shows a significant number of tail cells for heavily pre-cycled WL's. No apparent spatial correlation is observed within the same WL. Fig. 7 shows the enhanced cell V_{TH} margin and retention time after refresh operation attributed to the considerable number of interface traps being annihilated prior to the refresh operation. Finally, the die photograph of the fabricated embedded flash test chip is shown in Fig. 8.

References

[1] H. Kojima et al., IEDM, 2007. [2] C. Deml et al., ISSCC, 2007. [3] J. Raszka et al., ISSCC, 2004. [4] K. Suh et al., ISSCC, 1995. [5] J. Lee et al., *TDMR*, vol. 4, no. 1, Mar. 2004. [6] A. Umezawa et al., VLSI Circuits Symp., 1993.

Table I. Embedded flash comparison

Embedded Flash	Dual-Poly [1]	Previous Single-Poly [3]	This Work
Process	90nm Logic CMOS	130nm Logic CMOS	65nm Logic CMOS
Cell Transistor	Floating Gate NMOS	3.3V I/O Device	2.5V I/O Device
High Voltage Switch	Special HV NMOS	3.3V I/O Device, Drain Extended NMOS	2.5V I/O Device (WL)
Tunnel Oxide	10nm	7nm	5nm
Cell V_{TH} Window	5.6V	0.7V	>1.9V
Single WL Program	Yes	No	Yes
Erase Time (Access Unit)	2ms (Block)	10ms (Block)	1ms (WL)
Program Time (Access Unit)	20μs (Word)	10ms (Block)	10μs (WL)
Read Time (Access Unit)	N. A.	10μs (Block)	10ns (WL)
Unit Cell Size	0.44μm^2	700μm^2 (estimated)	8.62μm^2
Capacity	3.5Mb	2kb	2kb
Application	High Density Code/Data Storage	Low Density Code/Data Storage	Moderate Density Code/Data Storage

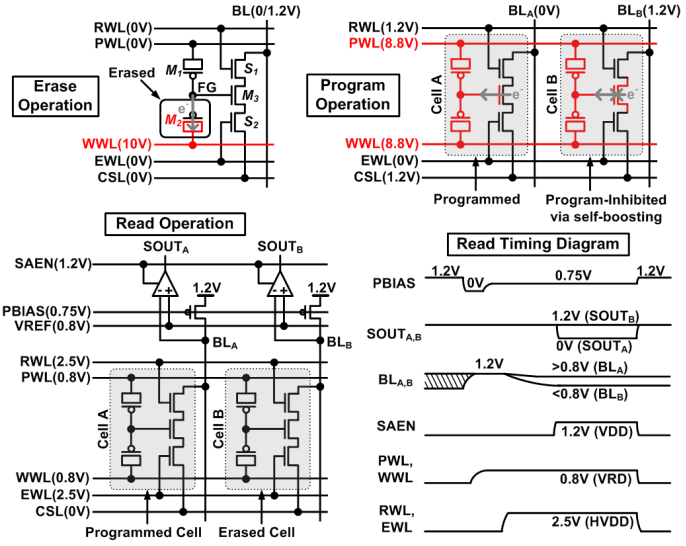

Fig. 2. Erase, program, and read bias condition and timing diagram.

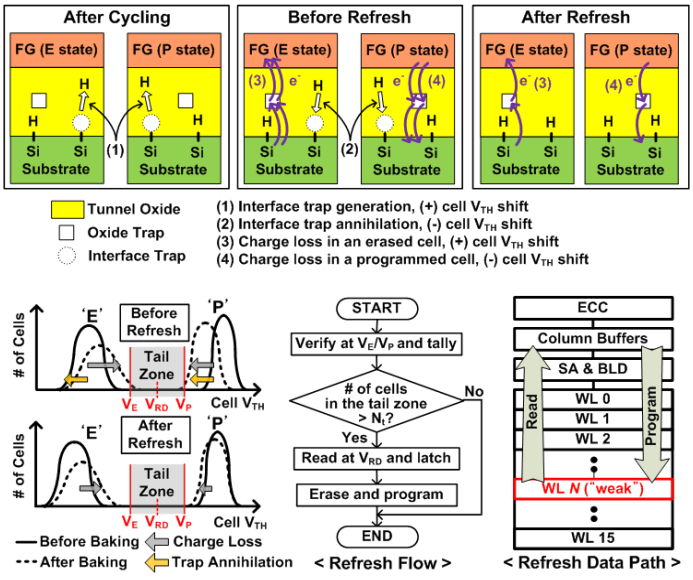

Fig. 4. Physical model of cell retention and proposed refresh scheme.

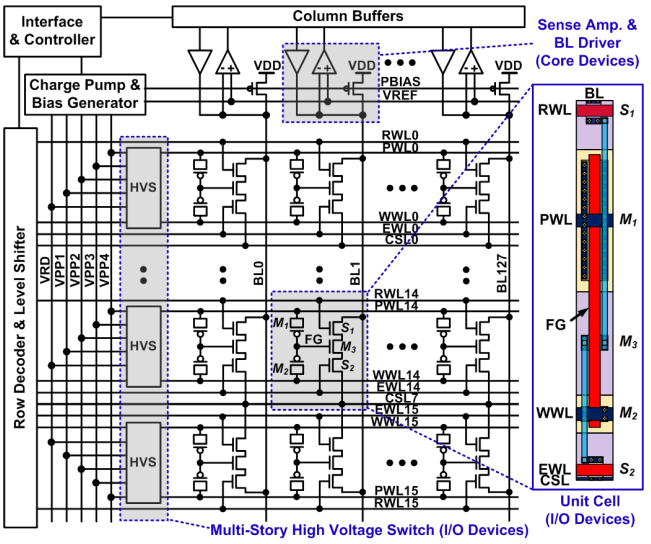

Fig. 1. Proposed logic-compatible embedded flash memory.

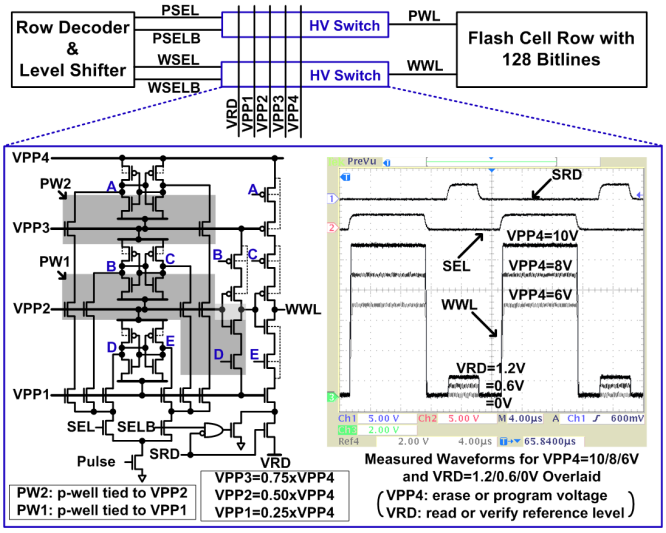

Fig. 3. Proposed multi-story high voltage switch works reliably up to 10V which significantly expands the cell V_{TH} window.

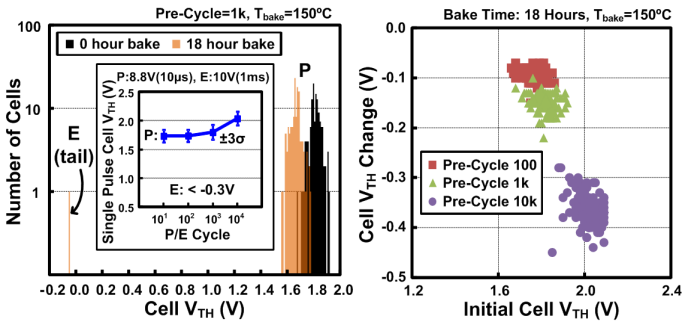

Fig. 5. Measured retention and endurance characteristic (left). Cell retention strongly depends on the number of P/E pre-cycles (right).

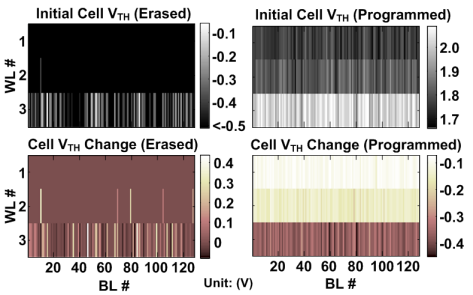

Fig. 6. Spatial cell V_{TH} map after an 18 hour bake at 150°C. WL 1, 2, 3 are pre-cycled 100, 1k, 10k times, respectively, prior to baking.

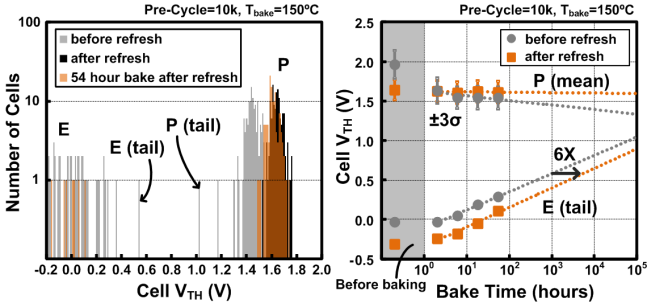

Fig. 7. Measured retention characteristic shows the enhanced cell V_{TH} margin and retention time after refresh operation.

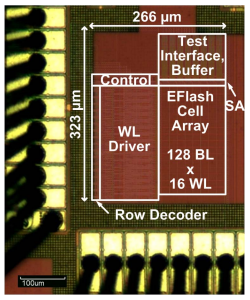

Fig. 8. Die photograph.