Embedded Analog Nonvolatile Memory With Bidirectional and Linear Programmability

Yi-Da Wu, *Student Member, IEEE*, Kok-Choon Cheng, Chih-Cheng Lu, and Hsin Chen, *Member, IEEE*

***Abstract*—Nonvolatile storage of analog values with floating-gate transistors has been useful for many applications. While most proposed analog nonvolatile memory devices employ electron tunneling to modify floating-gate charges, this brief presents an embedded analog nonvolatile memory device that employs only hot-carrier injections to achieve bidirectional programmability. Without using electron tunneling, the proposed circuit not only avoids multiplexing high-voltage signals but also facilitates direct storage of new data. In addition, each memory cell incorporates a simple inverter to make the programming process nearly linear, facilitating bidirectional and linear adaptability for neuromorphic systems. A prototype array of the analog memory has been fabricated with the standard CMOS 0.35-μm technology. The chip includes a simple on-chip comparator to program the analog memory accurately and automatically by negative feedback. The effective resolution is more than 8 bits over a dynamic range of 1.4 V. The intercell disturbance during programming and the data retention ability are also examined.**

***Index Terms*—Analog nonvolatile memory, bidirectional programmability, floating-gate transistors, neuromorphic engineering.**

I. Introduction

FLOATING-GATE transistors are first employed to achieve nonvolatile storage of analog values in [1], facilitating parallel and large-scale analog computation in a neuromorphic system. The storage of audio data is also demonstrated in [2]. Following the development of the CMOS-compatible (embedded) floating-gate transistor [3], its application is further extended to a wide variety of analog circuits, including on-chip biasing [4], offset correction [5], and the autozeroing amplifier [6]. Moreover, by exploiting the continuous and bidirectional programmability, floating-gate transistors function as adaptive computing elements in neuromorphic systems [7]–[11].

The charges on a floating gate can be modified by the exposure to the ultraviolet light or the electron tunneling through silicon dioxide or the hot-carrier injection. The later two mechanisms become more favorable as the transistor size continues to shrink. However, both mechanisms are inherently nonlinear, making it challenging to store analog values precisely. Although the nonlinearity can be exploited to realize nonlinear synapse plasticity [12], many neuromorphic learning algorithms demand linear bidirectional adaptation [13].

For storing analog data accurately, one method is using pulsewidth modulation to control the programming process, but the programming speed is traded for accuracy [14]. Serrano *et al.* developed a predictive algorithm that is able to program analog memory accurately within eight iterations [15], [16]. However, the algorithm relies on precise sensing of the channel current and requires digital implementation with analog-to-digital overheads. Using a comparator to control the programming process automatically by negative feedback is thus a simpler solution [17]. However, all the designs in [14]–[18] remove floating-gate charges by Fowler–Nordheim tunneling, which normally requires a tunneling voltage higher than 8 V to achieve an acceptable erasing speed. To avoid the need for high-voltage logic gates, all the designs simply erase the memory content globally before new values can be stored. Therefore, none of the designs supports bidirectional linear adaptation for neuromorphic applications.

The analog memory in [19] realizes neuromorphic adaptation by confining electron tunneling to a tiny injection gate. The injected changes are then transferred to a large storage gate via a teraohm resistor. As a result, the voltage change can be small and approximately linear, but the large time constant of the charge transfer limits the programming speed significantly. A better alternative is fixing the tunneling voltage to keep the tunneling current constant [20]. The analog memory in [21] further exploits the virtual-short property of an amplifier to fix the floating-gate voltage, such that both electron tunneling and electron injection can remain constant to achieve linear adaptation. However, all these designs employ electron tunneling to program analog memory at least in one direction. High-voltage logic gates are thus needed.

This brief presents an embedded analog nonvolatile memory device employing both hot-hole and hot-electron injections to achieve bidirectional programmability. The proposed memory not only avoids high-voltage circuits but also has the following advantages. For data storage, the proposed memory is able to store a new value directly without the need for a global erase in advance. This helps reduce the programming time remarkably, particularly when the new value is close to the original one. Moreover, the programming process is controlled by a simple on-chip comparator in a feedback loop. The negative-feedback loop enables the memory cell to store analog data automatically and accurately, regardless of the programming nonlinearity. For neuromorphic applications, the memory cell further incorporates a simple inverter to facilitate linear and bidirectional adaptability. A prototype array of the proposed memory is designed and fabricated with the CMOS 0.35-μm technology provided by the Taiwan Semiconductor

Manuscript received May 21, 2011; revised November 9, 2011; accepted December 18, 2011. Date of current version February 23, 2012. This work was supported by the National Science Council under Project NSC100-2220-E-007-009. This paper was recommended by Associate Editor N. Krishnapura.

The authors are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: ydwu@ee.nthu.edu.tw; hchen@ee.nthu.edu.tw).

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/TCSII.2012.2184371

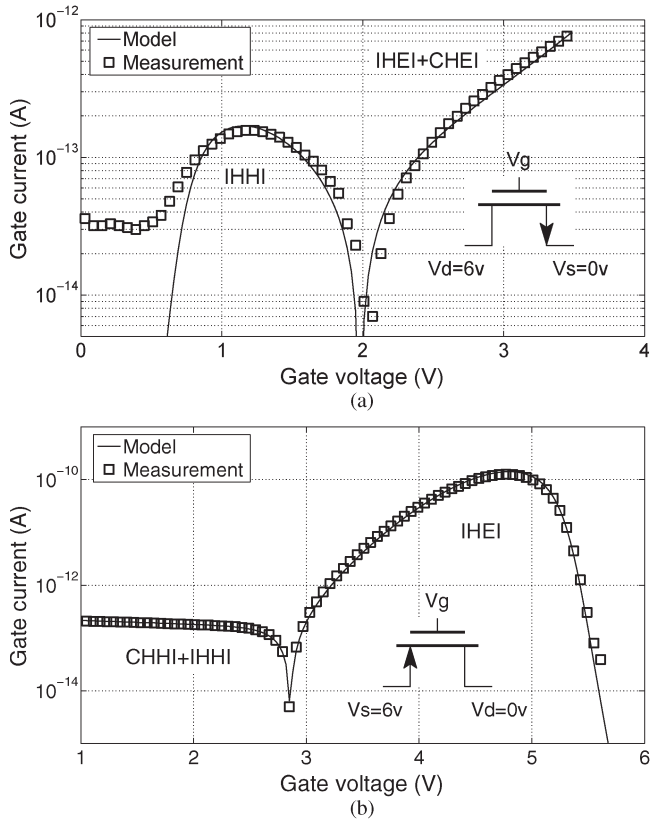

Fig. 1. Measured and modeled hot-carrier injection currents for the (a) n-type and (b) p-type transistors.

Manufacturing Co. (TSMC). Following the introduction, the models of the programming mechanisms are derived in Section II. The design of the proposed analog memory is then described in Section III, and the experimental results are presented and discussed in Section IV. Finally, Section V concludes the findings and potential applications.

II. Models of Programming Mechanisms

The proposed analog memory uses one nMOS and one pMOS to inject hot holes and hot electrons, respectively, to the floating gate. To model the injection currents, both nMOS and pMOS testkeys with a size of 5 μm/1 μm are first fabricated. With the drain-to-source voltage fixed at 6 V, injection currents $I_{\rm inj}$ versus gate voltages V_g are measured.

The square markers in Fig. 1(a) and (b) show the absolute value of $I_{\rm inj}$ averaged over five identical nMOS and pMOS transistors, respectively. For the nMOS with $V_g < 2V$, $I_{\rm inj}$ is negative and is mainly attributed to impact-ionized hot-hole injection (IHHI) into the gate [22]. This is verified by the fact that the gate current vanishes as the source terminal becomes floating. As V_g increases beyond 2 V, $I_{\rm inj}$ becomes positive because the relatively high V_g encourages both channel hot-electron injection (CHEI) and impact-ionized hot-electron injection (IHEI). Therefore, $|I_{\rm inj}|$ converges to zero when the hole injection is equal to the electron injection. For the pMOS with channel holes, the convergent point occurs at $V_g = 2.8$ V. $|I_{\rm inj}|$ preceding the self-convergent point involves both channel hot-hole injection (CHHI) and IHHI, whereas beyond this point, only IHEI is involved. It is notable that, although each transistor can generate both hot-hole and hot-electron injections,

TABLE I
Parameter Values for Hot-Carrier Injection Currents in the CMOS 0.35-μm Technology

PMOS	α_i	β_i
CHHI	6.37×10^{-12}	3.6185
IHHI	7.95×10^{-13}	6.5716
IHEI	0.93	-97.427
NMOS	α_i	β_i
CHEI	9.39×10^{-17}	50.028
IHHI	2.38×10^{-3}	-112.83
IHEI	7.835×10^{-18}	56.331

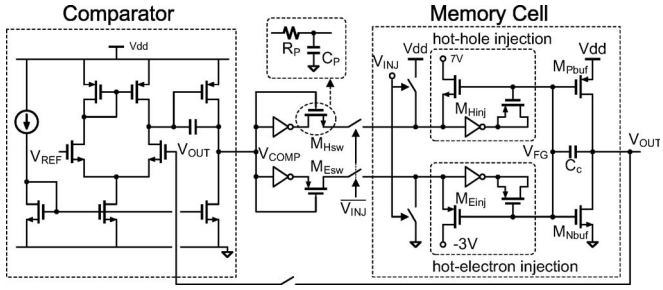

Fig. 2. Proposed memory cell and its programming circuit.

the nMOS is preferable than the pMOS to generate hot-hole injection to keep V_{GS} and the power consumption low. The same concern applies to the selection of pMOS for hot-electron injection.

The channel hot-carrier injection and the impact-ionized hot-carrier injection, respectively, can be modeled by [23]

$$I_{\rm CHHI/CHEI} = \alpha_1 I_S \left(|V_{DS}| - V_{\rm DSAT}\right)^2 \cdot e^{\frac{\beta_1}{|V_{DS}| - V_{\rm DSAT}}} \quad (1)$$

$$I_{\rm IHHI/IHEI} = \alpha_2 I_S \left(|V_{DS}| - V_{\rm DSAT}\right)^3 \cdot e^{\frac{\beta_2}{|V_{DS}| - V_{\rm DSAT}}} \quad (2)$$

where I_S represents the channel current, V_{DS} is the drain-to-source voltage, and $V_{\rm DSAT}$ is the saturation voltage. The process-dependent parameters α_i and β_i are extracted from the measurement results in Fig. 1 and are summarized in Table I.

Assuming that CHHI and IHHI occur with equal probability for the pMOS, its injection current can be written as (3). Similarly, the injection current for the nMOS is written as (4). These equations are given as follows:

$$I_{\rm inj,pMOS} = \frac{1}{2}(I_{\rm CHHI} + I_{\rm IHHI}) - I_{\rm IHEI} \quad (3)$$

$$I_{\rm inj,nMOS} = I_{\rm IHHI} - \frac{1}{2}(I_{\rm CHEI} + I_{\rm IHEI}). \quad (4)$$

The solid curves in Fig. 1 plot the modeled injection currents, agreeing with the measurement results satisfactorily. These models are incorporated as voltage-controlled current sources in the SPICE simulation to design the proposed memory circuit.

III. Circuit Description

A. Memory Cell

Fig. 2 shows the proposed analog memory cell and the programming circuit that is shared among memory cells. The analog information is stored on the floating gate $V_{\rm FG}$, and the injection transistors $M_{\rm Hinj}$ and $M_{\rm Einj}$ inject hot holes and hot electrons, respectively, to modify $V_{\rm FG}$. The inverter consisting of $M_{\rm Pbuf}$ and $M_{\rm Nbuf}$ then amplifies $V_{\rm FG}$ to produce the

memory output $V_{\rm OUT}$. With the inverter, $V_{\rm FG}$ remains nearly constant (1.24–1.19 V) as $V_{\rm OUT}$ changes over a wide dynamic range (1–2.4 V). This helps not only to reduce programming time but also to enhance programming linearity. The linearity comes from the fact that a constant V_{GS} results in a constant I_S and, thus, a constant injection current, as indicated by (1) and (2). Furthermore, $V_{\rm INJ}$ controls the selection of the memory cell for programming. As $V_{\rm INJ} = 0$ V and $V_{dd} = 3$ V, $M_{\rm Hinj}$ and $M_{\rm Einj}$ have $|V_{DS}| \geq 6$ V to induce injection currents when they are turned on by the comparator. The two sets of inverters and MOS capacitors are used to alleviate the charge feedthrough into $V_{\rm FG}$. As $V_{\rm INJ} = V_{dd}$, the source terminals of $M_{\rm Hinj}$ and $M_{\rm Einj}$ are connected to V_{dd} and 0 V, respectively, to disable the injection currents.

B. Programming Circuit

The programming circuit simply consists of a voltage comparator, two inverters, and two low-pass filters formed by $M_{\rm Hsw}$ and $M_{\rm Esw}$. In the programming mode, $V_{\rm OUT}$ is connected to the voltage comparator, which controls the source voltages of $M_{\rm Hinj}$ and $M_{\rm Einj}$. If $V_{\rm OUT}$ is greater than the target value $V_{\rm REF}$, $M_{\rm Hinj}$ is turned on to generate hot-hole injection to raise $V_{\rm FG}$ until $V_{\rm OUT} = V_{\rm REF}$. Similarly, if $V_{\rm OUT} < V_{\rm REF}$, $M_{\rm Einj}$ is turned on to inject hot electrons to increase $V_{\rm OUT}$.

The negative feedback enables $V_{\rm OUT}$ to track $V_{\rm REF}$ automatically and bidirectionally. As $V_{\rm OUT}$ is close to $V_{\rm REF}$, the phase margin of the feedback loop affects the stability remarkably. The feedback loop contains at least three poles, and the pole at $V_{\rm OUT}$ varies with the loading capacitor connected to $V_{\rm OUT}$. To ensure the stability, the dominant pole is determined by the compensation capacitor within the comparator. In addition, the comparator completely switches as $|V_{\rm OUT} - V_{\rm REF}| > 1$ mV. This voltage difference can be easily surpassed by the overshoots of the programming dynamics. $V_{\rm OUT}$ could thus still oscillate around $V_{\rm REF}$ even if the phase margin is sufficient. Although the oscillation can be prevented by incorporating a hysteretic comparator, the hysteretic window limits the comparator resolution directly. Therefore, we opt to connect $M_{\rm Hsw}$ and $M_{\rm Esw}$ to the source nodes of $M_{\rm Hinj}$ and $M_{\rm Einj}$, respectively, to form two low-pass filters to reduce the oscillation before settling.

C. Analog Memory Array

A prototype array of 4×4 analog memory cells is designed and fabricated with the CMOS 0.35-μm technology by the TSMC. Fig. 3 shows the photos of the chip and an individual memory cell. Each memory cell has an area of 170 μm $\times$ 70 μm, and the memory array has an area of 1330 μm $\times$ 1050 μm.

Fig. 4 illustrates the architecture of the memory array. In the programming mode, the signal WR_EN enables the comparator and the write decoder. The decoder inputs $W[3:0]$ then select and connect a specific cell with the comparator to form a negative feedback loop, which programs $V_{\rm OUT}$ to $V_{\rm REF}$ automatically. Afterward, another cell can be selected and programmed to a different $V_{\rm REF}$. In the readout mode, the signal RE_EN enables the read decoder, and $R[3:0]$ connect $V_{\rm OUT}$ of the selected cell to the external output.

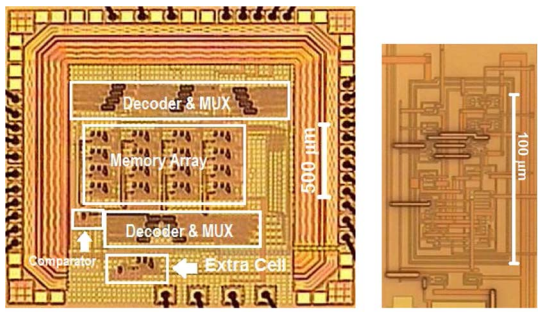

Fig. 3. Microphotograph of the (a) chip containing the memory array and (b) single memory cell.

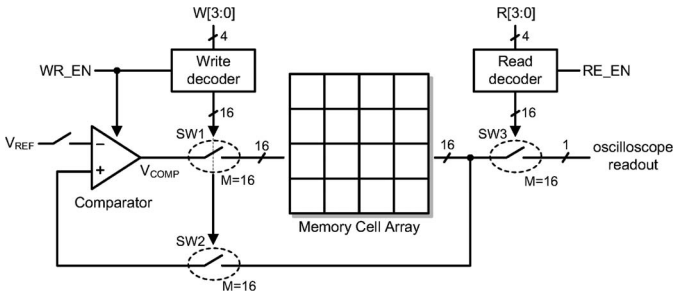

Fig. 4. Architecture of the memory array.

IV. Experimental Results

A. Programming Dynamics

The dynamics of programming $V_{\rm OUT}$ to different voltage levels were first investigated. In the beginning of each trial, $V_{\rm REF}$ was initialized to 1.5 V, so was $V_{\rm OUT}$ programmed to 1.5 V. As $V_{\rm REF}$ stepped to a target value between 1 and 2.4 V at $t = 0s$, the corresponding dynamics of $V_{\rm OUT}$ are shown in Fig. 5. For target values above 1.5 V, the IHEI occurred to increase $V_{\rm OUT}$, whereas for target values below 1.5 V, the IHHI occurred to reduce $V_{\rm OUT}$. Since the IHHI was smaller than the IHEI by two orders (see Fig. 1), the longest programming time was 120 ms for $V_{\rm OUT}$ to become 1 V. On the other hand, $V_{\rm OUT}$ reached all target values above 1.5 V within 1 ms. The programming speed for the IHHI could be improved by increasing the size of $M_{\rm Hinj}$ or $M_{\rm Nbuf}$. The former increases the hole-injection current, as indicated by (2), and the latter increases the inverter gain, subject to the gate capacitance of $M_{\rm Nbuf}$, and remains much smaller than C_c. However, the cell area or precision could be the tradeoff. Moreover, $V_{\rm OUT}$ reached all target values without significant oscillation, indicating that the filters formed by $M_{\rm Hsw}$ and $M_{\rm Esw}$ were effective.

The dynamics of programming $V_{\rm OUT}$ to 1 V also revealed the programming linearity of the proposed circuit clearly. Each 0.1-V change took around 25 ms when the IHHI occurred. By connecting the drain of $M_{\rm Einj}$ to -1.75 V to slow down the IHEI, the linearity of programming $V_{\rm OUT}$ to 2.4 V was also clearly demonstrated by the gray curve in Fig. 5. The linearity facilitated the alternative of programming $V_{\rm OUT}$ by controlling the pulsewidth of the WR_EN signal precisely. The bidirectional and linear programmability further fits the iterative learning required by neuromorphic systems. The learning rate can be adjusted by either the pulsewidth of the WR_EN or the V_{DS} of $M_{\rm Hinj}$ and $M_{\rm Einj}$.

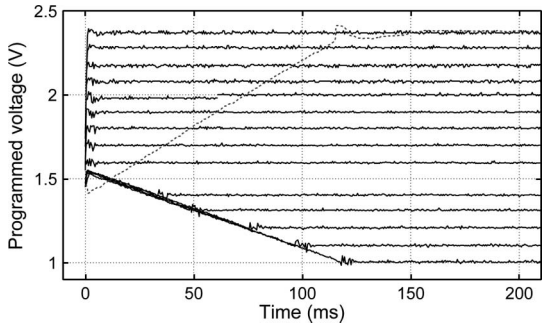

Fig. 5. Measured dynamics of $V_{\rm OUT}$ during programming.

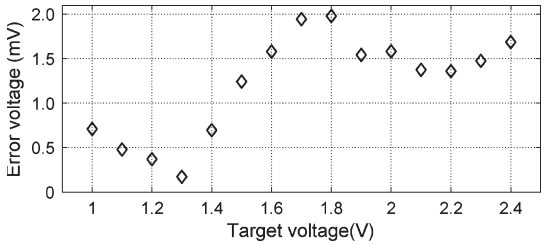

Fig. 6. Programming error over the programmable dynamic range.

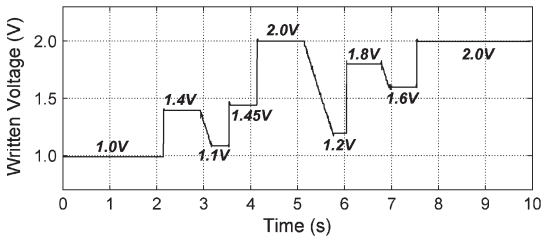

Fig. 7. Measured $V_{\rm OUT}$ when $V_{\rm REF}$ kept changing to different levels (noted beside the trace of $V_{\rm OUT}$) during 10 s.

B. Effective Resolution

After the programming was disabled, the difference between $V_{\rm OUT}$ and $V_{\rm REF}$ was measured as the programming error. Fig. 6 shows that the programming errors achieved were all smaller than 2 mV over the entire dynamic range. The maximum error corresponded to an effective resolution of 9.45 bits $(\log_2(1.4/0.002))$. The resolution had been comparable with [21] (10 bits) and sufficient for the neuromorphic systems proposed in [24] and [25]. Nevertheless, the main limitation on the resolution should come from the voltage comparator. The measurements across different chips with different comparators indicated that the worst resolution was greater than 8 bits. A high-precision comparator should be the key for improving the resolution further. For example, an off-chip comparator was employed to achieve 14-bit resolution in [17].

C. Bidirectional Programmability

To demonstrate the ability of $V_{\rm OUT}$ to track $V_{\rm REF}$ bidirectionally and directly, $V_{\rm REF}$ was stepped to eight different values during 10 s. The corresponding $V_{\rm OUT}$ was measured and is shown in Fig. 7. Obviously, $V_{\rm OUT}$ was able to track $V_{\rm REF}$ reliably without the need for erasure. This promising property came from the use of the IHHI instead of electron tunneling for increasing $V_{\rm FG}$, as well as the use of a single floating gate. Although similar behavior could be achieved by applying the IHEI to a pair of floating-gate transistors that form a differential amplifier [21], the common-mode voltage of the differential pair would change and alter the normal operation of the amplifier.

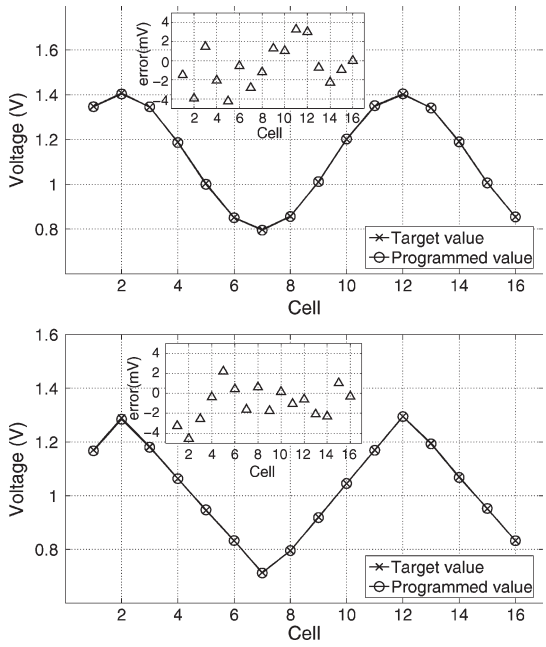

Fig. 8. (Top) Sinusoidal wave and (bottom) triangular wave stored in the memory cell array.

D. Programming the Memory Array

The ability of the memory array to store different analog waveforms, i.e., one sinusoidal and one triangular, was further examined. Each waveform was first sampled into 16 values. The sampled values were then programmed into the memory array sequentially. Fig. 8 depicts both the target and programmed values in the memory array, and the insets show the programming errors. No significant intercell disturbance was observed. This was because the proposed circuit did not need to multiplex high-voltage signals, and the likelihood to have disturbances became extremely low.

E. Data Retention Time

The ability of the proposed memory circuit to retain data for a long time is further investigated. According to the Arrhenius law [26], the data retention time can be modeled as

$$t_{\rm DR} = t_0 \cdot \exp\left(\frac{E_A}{kT_n}\right) \tag{5}$$

where T_n is the absolute temperature (in kelvins), E_A is the activation energy (in electronvolts), and k is the Boltzmann's constant. $t_{\rm DR}$ is the retention time at T_n, whereas t_0 is the retention time at the infinite temperature.

Based on the accelerated retention-time testing, the memory cell was initially programmed to 1.5 V and then baked under two different temperatures, namely, 378 and 398 K. The retained voltages across time were then measured and normalized with respect to the initial value (1.5 V). Fig. 9 plots the normalized retained voltages versus the baking time for the two experiments (the square and the diamond symbols). The exponential curves fitting to the measured data with minimal squared errors were also shown.

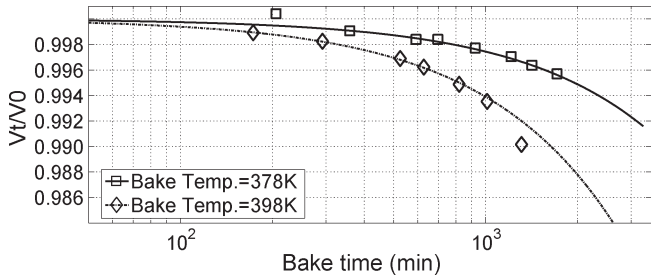

Fig. 9. Accelerated lifetime test of the proposed memory at 378 and 398 K.

The retention time $t_{\rm DR}$ for maintaining 7-bit resolution is then estimated as follows: Seven-bit resolution corresponds to a retained level of more than 99.27%. According to Fig. 9, the normalized retained voltage falls below the threshold after 3053 and 1280 min at 378 and 398 K, respectively. By substituting the two threshold points into (5), E_A is derived as 0.563 eV and t_0 as 9.4×10^{-5} min. The data retention time for 7-bit resolution at room temperature (298 K) is thus estimated to be 0.6 year. The retention time is shorter than that reported in [27] for at least two reasons. First, the gate oxide in our technology is much thinner than that in [27]. Second, hole injection could damage the gate oxide more severely than electron injection, increasing the likelihood for charges to leak away from the floating gate [28]. Nevertheless, the experiment in [22] demonstrates that the injection transistors can endure complete erasure and programming for more than 10^4 cycles. In addition, the retention time is already sufficient for general neuromorphic computation in VLSI.

V. Conclusion

A prototype array of analog nonvolatile memory with bidirectional programmability is designed and tested. The overall experimental results indicate that hot-hole injection instead of electron tunneling is a good alternative for achieving bidirectional programmability. For data storage, a simple on-chip comparator is shown capable of programming the analog memory automatically within 120 ms. The resolution is more than 8 bits over a dynamic range of 1.4 V. For neuromorphic application, the memory cell is shown to exhibit bidirectional and linear programmability with adjustable programming speed. The retention time is also proved to be sufficient for general neuromorphic applications.

Acknowledgment

The authors would like to thank the National Chip Implementation Center for chip fabrication.

References

[1] M. Holler, S. Tam, H. Castro, and R. Benson, "An electrically trainable artificial neural network (ETANN) with 10240 "floating gate" synapses," in *IEEE Int. Joint Conf. Neural Netw.*, Jun. 1989, pp. 191–196.

[2] H. Van Tran, T. Blyth, D. Sowards, L. Engh, B. S. Nataraj, T. Dunne, H. Wang, V. Sarin, T. Lam, H. Nazarian, and G. Hu, "A 2.5 V 256-level non-volatile analog storage device using EEPROM technology," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 1996, pp. 270–271.

[3] A. Thomsen and M. A. Brooke, "A floating-gate MOSFET with tunneling injector fabricated using a standard double-polysilicon CMOS process," *IEEE Electron Device Lett.*, vol. 12, no. 3, pp. 111–113, Mar. 1991.

[4] Y. L. Wong, M. H. Cohen, and P. A. Abshire, "A 750-MHz 6-b adaptive floating-gate quantizer in 0.35- μm CMOS," *IEEE Trans. Circuits Syst. I*, vol. 56, no. 7, pp. 1301–1312, Jul. 2009.

[5] Y. L. Wong, M. H. Cohen, and P. A. Abshire, "A 1.2-GHz comparator with adaptable offset in 0.35- μm CMOS," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 55, no. 9, pp. 2584–2594, Oct. 2008.

[6] P. Hasler, B. A. Minch, and C. Diorio, "An autozeroing floating-gate amplifier," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 74–82, Jan. 2001.

[7] C. Diorio, P. Hasler, B. A. Minch, and C. A. Mead, "A floating-gate MOS learning array with locally computed weight updates," *IEEE Trans. Electron Devices*, vol. 44, no. 12, pp. 2281–2289, Dec. 1997.

[8] P. Hasler, "Continuous-time feedback in floating-gate MOS circuits," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 56–64, Jan. 2001.

[9] P. Hasler and J. Dugger, "Correlation learning rule in floating-gate pFET synapses," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 65–73, Jan. 2001.

[10] C. Diorio, D. Hsu, and M. Figueroa, "Adaptive CMOS: From biological inspiration to systems-on-a-chip," *Proc. IEEE*, vol. 90, no. 3, pp. 345–357, Mar. 2002.

[11] P. Hasler and J. Dugger, "An analog floating-gate node for supervised learning," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 52, no. 5, pp. 834–845, May 2005.

[12] S. Ramakrishnan, P. E. Hasler, and C. Gordon, "Floating gate synapses with spike-time-dependent plasticity," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 3, pp. 244–252, Jun. 2011.

[13] G. Cauwenberghs and M. A. Bayoumi, *Learning on Silicon: Adaptive VLSI Neural Systems*. Norwell, MA: Kluwer Academic, 1999.

[14] K.-H. Kim, K. Lee, T.-S. Jung, and K.-D. Suh, "An 8-bit-resolution, 360-μs write time nonvolatile analog memory based on differentially balanced constant-tunneling-current scheme (DBCS)," *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1758–1762, Nov. 1998.

[15] G. Serrano, P. D. Smith, H. J. Lo, R. Chawla, T. S. Hall, C. M. Twigg, and P. Hasler, "Automatic rapid programming of large arrays of floating-gate elements," in *Proc. Int. Symp. on Circuits and Syst.*, 2004, pp. 373–376.

[16] A. Bandyopadhyay, G. J. Serrano, and P. Hasler, "Adaptive algorithm using hot-electron injection for programming analog computational memory elements within 0.2% of accuracy over 3.5 decades," *IEEE J. Solid-State Circuits*, vol. 41, no. 9, pp. 2107–2114, Sep. 2006.

[17] C. Diorio, S. Mahajan, P. Hasler, B. Minch, and C. Mead, "A high-resolution non-volatile analog memory cell," in *Proc. Int. Symp. on Circuits and Syst.*, May 1995, vol. 3, pp. 2233–2236.

[18] R. R. Harrison, J. A. Bragg, P. Hasler, B. A. Minch, and S. P. Deweerth, "A CMOS programmable analog memory-cell array using floating-gate circuits," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 48, no. 1, pp. 4–11, Jan. 2001.

[19] O. Fujita and Y. Amemiya, "A floating-gate analog memory device for neural networks," *IEEE Trans. Electron Devices*, vol. 40, no. 11, pp. 2029–2035, Nov. 1993.

[20] T. Shibata, H. Kosaka, H. Ishii, and T. Ohmi, "A neuron-MOS neural network using self-learning-compatible synapse circuits," *IEEE J. Solid-State Circuits*, vol. 30, no. 8, pp. 913–922, Aug. 1995.

[21] M. Figueroa, S. Bridges, and C. Diorio, "On-chip compensation of device-mismatch effects in analog VLSI neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2004, pp. 441–448.

[22] K. Lee and Y. King, "New single-poly EEPROM with cell size down to $8F^2$ for high density embedded nonvolatile memory applications," in *Proc. Symp. VLSI Tech.*, 2003, pp. 93–94.

[23] K.-H. Lee, S.-C. Wang, and Y.-C. King, "Self-convergent scheme for logic-process-based multilevel/analog memory," *IEEE Trans. Electron Devices*, vol. 52, no. 12, pp. 2676–2681, Dec. 2005.

[24] H. Chen and A. F. Murray, "Continuous restricted Boltzmann machine with an implementable training algorithm," *IEE Proc. Vis. Image Signal Process.*, vol. 150, no. 3, pp. 153–158, Jun. 2003.

[25] Y.-D. Wu, S.-J. Lin, and H. Chen, "A log-domain implementation of the diffusion network in very large scale integration," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2010, pp. 2487–2495.

[26] J. Park, M. Jo, E. M. Bourim, J. Yoon, D.-J. Seong, J. Lee, W. Lee, and H. Hwang, "Investigation of state stability of low-resistance state in resistive memory," *IEEE Electron Device Lett.*, pp. 485–487, May 2010.

[27] T. Morie, O. Fujita, and K. Uchimura, "Self learning analog neural network LSI with high-resolution nonvolatile analog memory and a partially serial weight-update architecture," *IEICE Trans. Electron.*, vol. E-80C, no. 7, pp. 990–995, Jul. 1997.

[28] S. Haddad, C. Chang, B. Swaminathan, and J. Lien, "Degradations due to hole trapping in Flash memory cells," *IEEE Electron Device Lett.*, vol. 10, no. 3, pp. 117–119, Mar. 1989.