A Floating-Gate Analog Memory with Bidirectional Sigmoid Updates in a Standard Digital Process

Junjie Lu and Jeremy Holleman
The University of Tennessee, Knoxville, TN, USA
jlu9, jhollema@utk.edu

***Abstract*—A floating-gate current-output analog memory is implemented in a 0.13-µm digital CMOS process. The proposed memory cell achieves random-accessible and bidirectional updates with a sigmoid update rule. A novel writing scheme is proposed to obtain tunneling selectivity without on-chip high-voltage switches or charge pumps, and reduces interconnections and pin count. Parameters of empirical models for floating gate charge modification are extracted from measurements. Measurement and simulation results show that the proposed memory consumes 45 nW of power, has a 7-bit programming resolution, 53.8 dB dynamic range and 86.5 dB writing isolation.**

I. INTRODUCTION

A floating-gate analog memory uses the charge trapped on the isolated gate to store analog variables in a non-volatile way. It has been widely used in analog reconfigurable, adaptive and neuromorphic systems, such as electronic potentiometer [1], pattern classifier [2], silicon learning networks [3], and adaptive filter [4].

Without direct electrical connections, the stored value of the memory is updated by depositing electrons to the floating gate (FG) by hot-electron injection, or removing them by Fowler–Nordheim tunneling [5]. Compared to injection, tunneling selectivity is harder to obtain because it often involves controlling a high voltage (HV) on chip. Therefore, many previous works [2], [3] use tunneling as the global erase, and injection to program individual memory to its target value. However, in an online adaptive system, a bidirectional update is preferable because the stored values need to vary with the inputs. Previous works have proposed approaches to achieve selective tunneling. In [6], the selected memory is tunneled by pulling up the tunneling voltage and pulling down the gate voltage simultaneously. This approach requires a number of tunneling control pins equal to the number of rows in the memory array, which is not desirable for large-scale systems. In [1], a HV switch is built with lightly-doped-drain nFETs. This device is not compatible with standard digital processes and consumes static power because it cannot be completely turned off. In [4], a charge pump is used to generate a local HV for the selected memory. A simple charge pump provides limited voltage boost, while a more complex one consumes larger area and/or requires multi-phase clocks.

Another important performance metric of analog memory is the update rule. The dynamic of the single-transistor FG memory [6] leads to exponential and value-dependent update, which, in general, affects the stability of the adaptation [4]. A linear update can be obtained by fixing the FG node voltage during update with a capacitive feedback loop around a differential [1] or single-ended amplifier [4].

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Army Research Office (ARO) agreement number W911NF-12-1-0017. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, the Department of the Army, or the U.S. Government

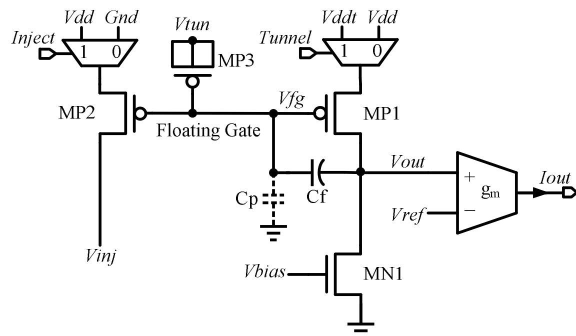


Figure 1. Schematic of the proposed floating-gate analog memory cell.

In this paper, we present our floating-gate current-output analog memory which allows random-accessible control of both tunneling and injection. It avoids the use of charge pump, minimizes interconnection and pin count, and is compatible with standard digital process. The update rule is sigmoid-shaped, which is a smooth, monotonic and bounded function preferred by many adaptive and neuromorphic applications. Implemented in a commercially available 0.13-µm digital CMOS process using thick-oxide IO FETs, the memory cell achieves small area and low power consumption, and is suitable for integration into systems that exploit the high-density digital logic available in modern CMOS technology.

II. FLOATING-GATE ANALOG MEMORY CELL

A. Circuit Description

The schematic of the proposed FG analog memory cell is shown in Fig. 1. The gate of MP1-MP3 and the top plate of C_f form the FG. The stored charge can be modified by the injection transistor MP2 and the tunneling transistor MP3. MP1 together with the current source MN1 forms a single-ended inverting amplifier. The amplifier has C_f as its feedback capacitor and the FG voltage V_{fg} as its input. The two MUXs at the sources of MP1 and MP2 control the tunneling and injection of the FG, which will be discussed later. The trans-conductor g_m converts voltage V_{out} to output current I_{out}. V_{ref} determines the nominal voltage of V_{out} during operation; it is set close to *Vdd* to avoid unwanted injection from MP1.

The negative feedback loop comprising the inverting amplifier and C_f keeps the FG voltage V_{fg} constant, ensuring a linear update of V_{out}. Tunneling or injection to the FG node

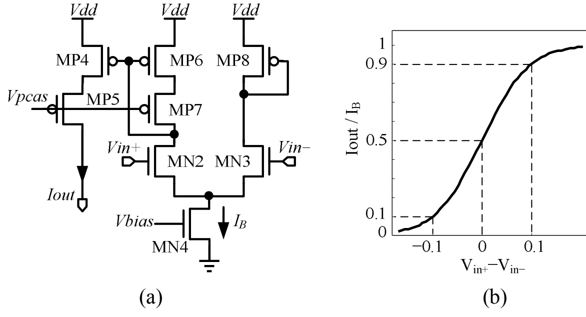


Figure 2. (a) Schematic of the trans-conductor and (b) its transfer curve.

changes the charge stored in C_f, therefore changes the output of the amplifier by $\Delta V_{out} = \Delta Q / C_f$. The loop gain L around the feedback loop is given by

$$L = A_v \frac{C_f}{C_f + C_p} \ , \quad (1)$$

where A_v is the open loop gain of the amplifier, and C_p is the parasitic capacitance on the FG as in Fig. 1. L attenuates the voltage variation of V_{fg} due to the swing of V_{out}, and suppresses the noise from the ground supply. To have a value-independent update as well as good ground noise rejection, C_f is made much larger than C_p, and MN1/MP1 are made long to increase their drain resistances.

The trans-conductor is implemented with a differential pair MN2/MN3 and a cascode current mirror MP4-MP7, depicted in Fig. 2(a). MP8 is added to improve the drain voltage matching between MN2 and MN3. The differential input $\Delta V_{in} = V_{in+} - V_{in-}$ steers the tail current I_B, and therefore converts the voltage input ΔV_{in} to current output I_{out}. Biased in deep sub-threshold region, the trans-conductor exhibits a transfer curve resembling a *tanh* function, plotted in Fig. 2(b). The mild nonlinearity is smooth, monotonic and bounded. From Fig. 2(b), a ΔV_{in} of 0.2V is enough to cause a change of I_{out} from $0.1 I_B$ to $0.9 I_B$, this reduced swing requirement further improves the update linearity, and makes the selective tunneling possible, as will be discussed in the following part.

B. Floating gate charge modification modeling

The proposed analog memory uses Fowler–Nordheim tunneling to remove the electrons from the FG and decrease the memory value. The applied electric field across MP3's gate oxide reduces its effective thickness, increasing the probability of electrons' tunneling through it. The tunneling current I_{tun} can be expressed by the empirical model [7] as

$$I_{tun} = I_{tun0} \exp\left(-\frac{V_f}{V_{ox}}\right), \quad (2)$$

where V_{ox} is the voltage across the tunneling transistor gate oxide, I_{tun0} and V_f are process dependent constants determined by measurements. Fig. 3 shows the measured tunneling current I_{tun} versus the oxide voltage V_{ox}, the fitted curve gives I_{tun0}=323.57 A and V_f = 248 V.

Hot-electron injection is employed to increase the stored value of the memory. During injection, a high source-to-drain voltage is applied to the injection transistor MP2. The large lateral field accelerates the carrier holes near the drain pinch-off region and generates hot electrons due to collision. The injection current I_{inj} depends on the source current and the drain-to-source voltage of MP2. A simplified empirical model derived from [7] approximates I_{inj} as

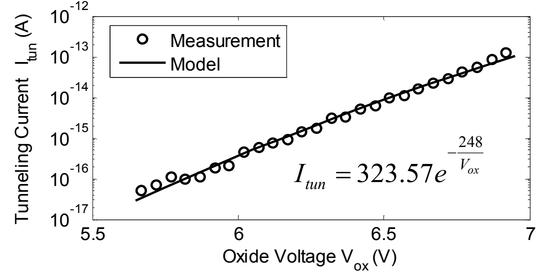


Figure 3. Tunneling current I_{tun} versus oxide voltage V_{ox}. V_{ox} is the voltage difference between V_{tun} and V_{fg}.

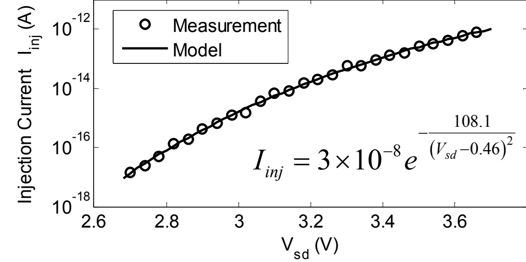


Figure 4. Injection current I_{inj} versus drain-to-source voltage of the injection transistor V_{sd}, I_s is constant and about 96 nA.

$$I_{inj} = \alpha I_s \exp\left(\frac{\beta}{V_{sd} + \delta}\right)^2, \quad (3)$$

where I_s is the injection transistor's source current, V_{sd} is the drain-to-source voltage, and α, β, δ are fit constants. In our memory cell, I_s is set by the biasing current of MP1 and the aspect ratios between MP1 and MP2. Fig. 4 shows the measured I_{inj} versus V_{sd}, and the fitted model.

The extracted models above can be used in the future designs as well as to improve programming convergence, as will be described in Section IV.

C. Selective and Value-independent Update Scheme

The proposed tunneling scheme exploits the steep change of tunneling current with regard to V_{ox} to achieve a good isolation between selected and unselected memories. As discussed, multiplexing or generating the high voltage supply V_{tun} will inevitably lead to complex or standard-process-incompatible circuits. Instead we simply connect the V_{tun} of all the memory cells to an off-chip voltage and manipulate V_{fg} to achieve selective tunneling. As $V_{ox} = V_{tun} - V_{fg}$, reducing V_{fg} is equivalent to increasing V_{tun} on its effect on the tunneling current. The operation of this scheme can be described by Fig. 5, showing the memory cell omitting components irrelevant to tunneling process. To show how V_{ox} is changed, typical nodal voltages are annotated. The negative feedback keeps the FG voltage at

$$V_{fg} = Vdd - V_{GS,P} \approx Vdd - 0.4, \quad (4)$$

where $V_{GS,P}$ is the gate to source voltage of MP1. Equation (4) holds for any Vdd values as long as MN1 and MP1 are in saturation region. Therefore, reducing supply voltage of the selected memory effectively reduces V_{fg} and increases V_{ox}. In our design, the power supply is switched from 3 V Vdd to a 1 V $Vddt$ so that V_{ox} is increased from 4.4 V to 6.4 V.

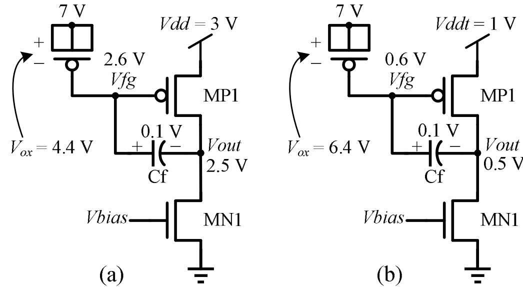
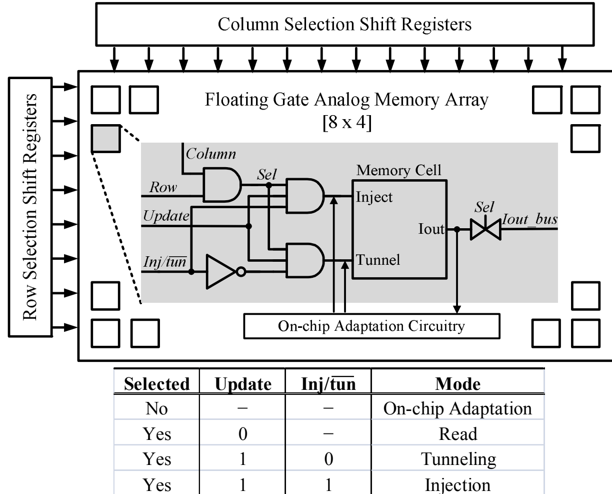


Figure 5. Simplified schematics and typical nodal voltages of memory cells (a) not selected. (b) selected for tunneling.



Selected	Update	Inj/$\overline{tun}$	Mode
No	–	–	On-chip Adaptation
Yes	0	–	Read
Yes	1	0	Tunneling
Yes	1	1	Injection

Figure 6. Block diagram of the FG analog memory array, and a table showing control signal settings for different operation modes of the cells.

According to (2), the tunneling current is 4.8 fA at V_{ox} = 6.4 V, while it's 1.1e−22 A at V_{ox} = 4.4 V, indicating an isolation over 7 orders of magnitude. In practice, the leakage at lower V_{ox} may be degraded by direct tunneling, which is a weaker function of the applied field [5], and parasitic coupling. Isolation of 83.54 dB is observed in measurement. The condition that MN1 stays in saturation during tunneling can be satisfied by choosing a proper V_{ref} and using the proposed trans-conductor to reduce the V_{out} swing. From Fig. 2(b), during normal operation, V_{out} needs to go as low as 2.3 V for zero output current when V_{ref} is 2.5 V. In this worst case, V_{out} is 0.3 V during tunneling, which is more than enough to saturate MN1 biased in sub-threshold. In a voltage-output memory, another inverting amplifier in place of the trans-conductor can be used to reduce V_{out} swing.

Injection selectivity is achieved by switching the source voltage of the injection transistor MP2. The source of MP2 in the unselected memory is connected to ground while the one in the selected cell is connected to *Vdd*, enabling injection. The *inj* terminals of all the memory cells are connected to an off-chip voltage, facilitating fine-tuning of injection rate. I_s is a constant for all memory cells since it's a mirrored version of the biasing current in MN1. Therefore, the injection is also selective and value-independent.

III. FLOATING-GATE MEMORY ARRAY

32 proposed FG analog memory cells are connected to form a memory array. They are organized in two dimensions

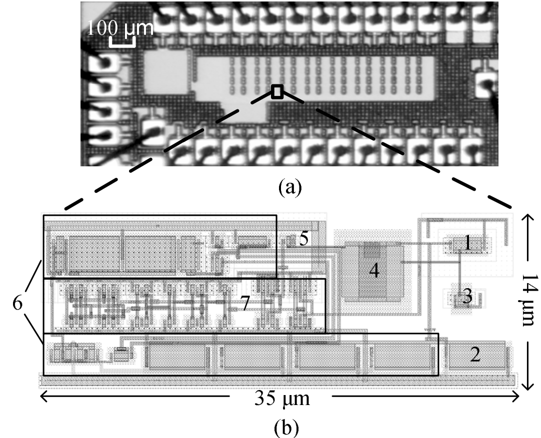


Figure 7. (a) Chip micrograph of the memory array together with on-chip adaptation circuitry and (b) layout view of a single memory cell. (1) MP1. (2) MN1. (3) MP3. (4) Cf. (5) MP2. (6) g_m. (7) Logics and MUXs.

and can be randomly accessed (selected) for read and write operations by setting both *column* and *row* inputs to high. The block diagram is shown in Fig. 6 with the cell symbolized. The cells are augmented by digital logics controlling their operation modes. The list of digital control combinations and their corresponding operation modes is shown in Fig. 6.

Once selected, a transmission gate connects the output of that cell to off-chip through *Iout_bus* for read-out during programming. The $Inj/\overline{tun}$ signal sets the direction of memory writing. The magnitude of writing is controlled by the pulse width of *Update* signal. When a cell is not selected, it maintains its value and can be read or written by on-chip circuits to implement adaptive algorithms. The proposed architecture is scalable because all signals and inter-connections are shared among the cells, and the pin count does not increase with the size of the array.

IV. MESUREMENT RESULTS

The proposed FG memory array has been fabricated in a 0.13-µm standard digital CMOS process using thick-oxide IO FETs. The die micrograph is shown in Fig. 7. Due to extensive metal fills in this process, details of the circuits cannot be seen. So the Virtuoso layout view is also presented.

The area of a single memory cell is 35x14 μm^2. It operates at 3 V power supply and consumes 15 nA with an output range of 0-10 nA. The biasing current is tunable and allows the designer to balance between range, speed and power consumption. The performance summary is listed in Table I.

The test setup is built around a National Instrument data acquisition (DAQ) card and a host PC. The programming procedure is controlled by a Labview program in the host PC and can be described as follows. The memory cell is first selected by configuring the shift registers. The output of the selected memory is then converted to voltage by a trans-impedance amplifier (TIA) and sampled by the DAQ card. The measured output and the target value are compared and the needed *Update* pulse width is predicted according to the models (2) or (3). Finally the DAQ card digital outputs are configured to generate the desirable $Inj/\overline{tun}$ and *Update* signals. This procedure is repeated until the error is below the error preset. The predictive programming procedure converges

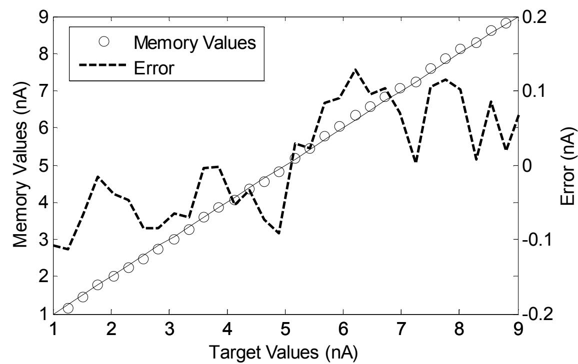


Figure 8. Programming accuracy of 30 linearly spaced values.

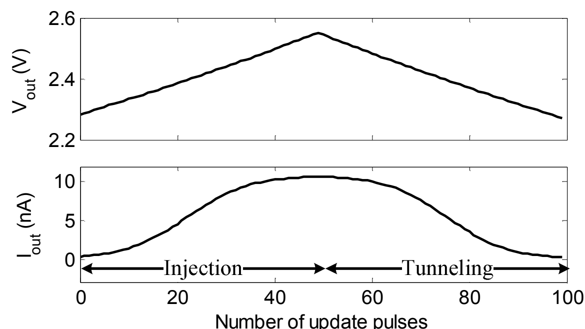


Figure 9. Ramping of the memory value, showing the update rule.

fast and the average number of iterations required to achieve a 0.5% error is 5-6. Fig. 8 demonstrates 30 memory cells programmed to values between 1 and 9 nA. The standard deviation of the programming error is 76 pA, indicating a 7-bit programming resolution. This resolution is limited by the noise and resolution of the TIA and DAQ. The memory output noise is 20.5 pA_{rms} over 10 KHz bandwidth from simulation, indicating a 53.8 dB dynamic range. Larger dynamic range and higher resolution can be obtained by increasing the biasing current.

To show the update rule, a memory is first ramped up then ramped down with fixed pulse width. The corresponding V_{out} and I_{out} are plotted in Fig. 9. Both injection and tunneling is linear to V_{out}, and the current output has a smooth sigmoid update rule. During the same test, the stored values of the other 31 unselected cells are monitored to measure the writing crosstalk. The crosstalk from the injection and tunneling of the selected cell to the unselected ones are plotted in Fig. 10. There is no observable injection crosstalk because the I_s can be completely turned off in the unselected cells. Tunneling crosstalk is very small, comparable to the noise floor of the measurement system. By averaging the values among 31 cells, a 471 fA tunneling crosstalk is approximated with a 10 nA writing magnitude in the selected cell, corresponding to an 86.5 dB isolation. The retention of the proposed memory cells were tested by continuously monitoring their outputs for 2 days at room temperature. During these 48 hours, no observable leakage was seen after the initial relaxation period during which the electrons trapped in the oxide are released. This is sufficient for general adaptive and neuromorphic applications. A longer test time at an elevated temperature can better characterize the retention of the memory and will be performed in the future.

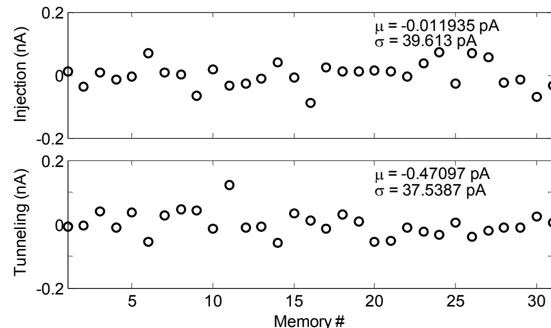


Figure 10. Value changes due to crosstalk among the 31 unselected cells when a selected cell is injected and tunneled with a magnitude of 10 nA.

TABLE I. SUMMARY OF PERFORMANCES

Parameter	*Value*
Technology	1P8M 0.13-μm CMOS
Area	35x14 μm^2
Power supply	3 V
Power consumption	45nW
Output range	0 - 10 nA
Programming resolution	7 bits
Dynamic range	53.8 dB
Programming isolation	86.5 dB

V. CONCLUSIONS

We have presented a floating-gate current-output analog memory in a 0.13-μm standard digital CMOS process. The novel update scheme allows random-accessible control of both tunneling and injection without the needs for high-voltage switches, charge pumps or complex routing. The update dynamics is sigmoid, suitable for many adaptive and neuromorphic applications. FG model parameters have been extracted to facilitate predictive programming. Measurement and simulation shows that with 45 nW power consumption, the proposed memory achieves 7-bit programming resolution, 53.8 dB dynamic range and 86.5 dB writing isolation.

REFERENCES

[1] 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.

[2] S. Peng, P. Hasler and D. V. Anderson, "An analog programmable multidimensional radial basis function based classifier," *IEEE Trans Circuits and Syst. I, Reg Papers,* vol. 54, no. 10, pp. 2148-2158, Oct. 2007.

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

[4] M. Figueroa, S. Bridges, D. Hsu and C. Diorio, "A 19.2 GOPS mixed-signal filter with floating-gate adaptation," *IEEE J. Solid-State Circuits,* vol. 39, no. 7, pp. 1196-1201, July 2004.

[5] J. J. Sanchez and T. A. DeMassa, "Review of carrier injection in the silicon/silicon-dioxide system," *IEE Proc. G–Circuits, Devices Systems,* vol. 138, no. 3, pp. 377-389, Jun. 1991.

[6] C. Diorio, "A p-channel MOS synapse transistor with self-convergent memory writes," *IEEE Trans. Electron Dev.,* vol. 47, no. 2, pp. 464-472, Feb. 2000.

[7] K. Rahimi, C. Diorio, C. Hernandez and M. D. Brockhausen, "A simulation model for floating-gate MOS synapse transistors," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2002, vol. 2, pp.532-535.