Proceedings of the 1995 IEEE International Symposium on Circuits and Systems, vol. 3, pp. 2233–2236, 1995.

A High-Resolution Nonvolatile Analog Memory Cell

Chris Diorio, Sunit Mahajan, Paul Hasler, Bradley Minch, Carver Mead
California Institute of Technology
Pasadena, California 91125
(818) 395-6996
chris@pcmp.caltech.edu

Abstract— **A 3-transistor nonvolatile analog storage cell with 14 bits effective resolution and rail-to-rail buffered voltage output is presented. The memory, which consists of charge stored on a MOS transistor floating gate, is written by means of hot-electron injection and erased by means of gate oxide tunneling. The circuit allows simultaneous memory reading and writing; by writing the memory under feedback control, errors due to oxide mismatch or trapping can be nearly eliminated. Small size and low power consumption make the cell especially attractive for use in analog neural networks. The cell is fabricated in a 2 μm n-well silicon BiCMOS process available from MOSIS.**

I. INTRODUCTION

ONE IMPEDIMENT to the development of silicon neural networks is the difficulty in storing analog weight values on-chip. Prior efforts typically used capacitive storage with clocked refresh [1], or multi-bit digital storage [2]. Both approaches pay a large penalty in terms of cell size, complexity, resolution, and power consumption. We have developed a new, nonvolatile analog memory cell which is far simpler than either approach; we believe this cell is well suited for use in silicon neural networks.

Our goal was to design a simple memory cell with the following attributes:

1. Nonvolatile analog storage.
2. High resolution and large dynamic range.
3. Read/write circuitry located on-chip
4. Support for simultaneous reading and writing.
5. Low power consumption and compact size.
6. Compatibility with standard MOS processing.

Since floating gate transistors are naturally well suited to the storage of nonvolatile analog memories, we have concentrated on developing circuits and devices that apply them to learning applications. We have succeeded in designing a 3-transistor circuit using floating gate transistors that achieves all the above goals.

II. WRITING TO FLOATING GATES

Before presenting the memory cell, we will briefly review techniques for writing floating gate memories. Although the advantages of using floating gate transistors as memory elements are well known [3, 4], their application to silicon neural networks has been limited. The principal reason has been the lack of a suitable bidirectional mechanism for writing the memory. Since the gate of a floating gate transistor is completely insulated by silicon dioxide, writing the memory involves moving charge carriers through this oxide. Fowler–Nordheim tunneling [5] and hot-electron injection are both well known for moving electrons through SiO_2; using them to write very precise analog memories, however, has historically proven difficult.

We use unidirectional gate-oxide tunneling to remove electrons from a floating gate. Positive high voltages are applied to a lightly doped n-type well (impurity concentration ~$10^{16}/cm^3$) within which is located an n^+ implant and an adjoining polysilicon floating gate. Tunneling current versus oxide voltage for a typical 342 Å gate oxide in the 2 μm Orbit BiCMOS process is shown in Figure 1. Trap formation in this oxide is quite slow; Figure 2 shows the reduction in tunneling rate versus total charge thru the oxide. Although tunneling provides an acceptable means for removing electrons from a floating gate, finding a complementary process for adding them has proven nontrivial.

Bidirectional tunneling can be used to add or remove electrons from a floating gate. This solution, however, requires either dual polarity high voltages, or a single polarity high voltage and a means for pulling the floating gate to this voltage when adding electrons, and pulling it near ground when removing them. Both approaches are unattractive. The dual polarity solution has a negative voltage much lower than the substrate potential; the single polarity solution does not support simultaneous memory reading and writing.

Hot-electron injection in conventional n-channel MOSFETs provides another means for adding electrons to a floating gate [6]. Achieving reasonable injection rates, however, requires that both the drain and gate

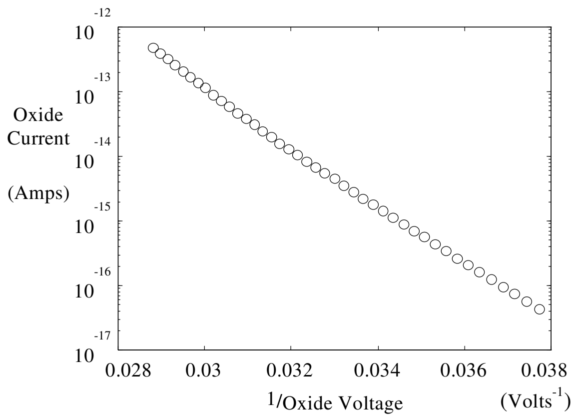

Figure 1: Tunneling current versus 1/(oxide voltage) for a 12 μm^2 gate oxide with 12 lineal μ of n^+ edge overlap

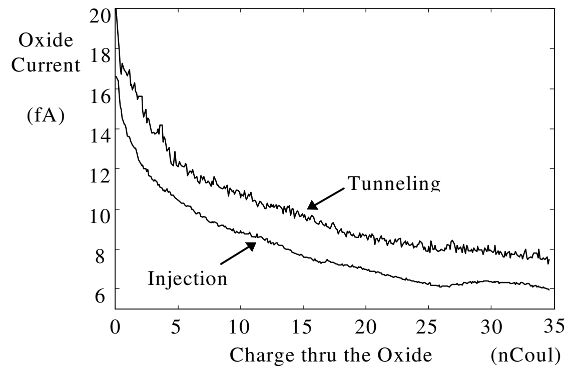

Figure 2: Tunneling and injection rates vs total charge thru the oxide. Tunneling junction V_{ox} = 29.5 V. Injection transistor V_{drain} = 3.25V, V_{gate} = 5.5V

voltages exceed 3.1V (the oxide potential barrier). High channel currents and high power consumption make this approach unattractive for learning networks.

We instead apply a bipolar transistor base implant (impurity concentration ~$10^{17}/cm^3$) to the channel of a conventional n-type MOSFET to enhance its injection rate. This transistor has a 6V threshold, allowing sub-threshold channel currents at gate voltages high enough to collect the injected electrons. Its drain breakdown voltage is ~7.25V, implying that for V_{drain}=5V the electric field in the depletion region surrounding the drain is very high. As indicated by the band diagram of Figure 3, by combining a large drain-to-channel electric field with high gate voltage in a subthreshold transistor, the probability of injecting electrons onto the gate is greatly increased. Typical injection efficiencies for this device are I_{inj}/I_{drain} = 0.01%. Oxide current versus drain voltage, for several values of gate voltage, are shown in Figure 4. Trapping rates are shown in Figure 2.

The data of Figure 4 show that over a wide range of drain voltage the injection rate is nearly constant. This effect is due to two competing processes. First, when drain voltage exceeds gate voltage, the electric field within the oxide opposes electron transport to the gate; injected electrons tend to return to the drain. Thus the injection efficiency decreases with increasing V_{dg}. Second, an increase in V_{dg} increases the field in the drain-to-channel depletion layer, exponentially increasing the number of electrons with sufficient energy to surmount the 3.1 eV oxide potential barrier. Over a wide range of drain voltages, the drop in efficiency is almost exactly compensated by the increasing population of hot electrons, yielding a nearly constant injection rate. By choosing to write the memory with an injection transistor drain voltage of 5V, the write rate is made insensitive to small variations in drain voltage.

III. THE MEMORY CELL

The memory cell is shown in Figure 5. Transistor Q1 is used for biasing. The amplifier formed by Q1 and Q2 drives the output node. Using subthreshold channel currents permits rail-to-rail output voltages and a power consumption measured in nW. Q3 is the pbase-channel transistor used for hot-electron injection. Achieving reasonable injection rates in this device requires keeping the floating gate between 5V and 6V; the circuit therefore requires a supply voltage of 6-7 Volts. Charge stored on the feedback capacitor C_i = 1 pF represents the analog memory. Capacitor C_p is the main parasitic element within the cell; it represents the coupling from capacitor C_i to ground. Also shown in Figure 5 is the feedback loop used to write the memory.

This paper assumes the following methodology for writing a memory cell:

1. The cell is erased before writing. A positive high voltage applied to the V_{tun} node removes electrons from the floating gate, causing V_{out} to approach ground. V_{tun} is then brought low, disabling the tunneling.

2. The desired memory voltage V_{in} is applied to the noninverting input of comparator A_1. Enabling the comparator output sets the drain of injector Q3 high, causing electron to be injected onto the floating gate. Electron injection causes V_{out} to slew upwards, at a rate set approximately by

$$\frac{dV_{out}}{dt} = \frac{I_{inj}}{C_i}$$

3. Once V_{out} exceeds V_{in}, the comparator lowers Q3's drain voltage to ground, leaving V_{out}=V_{in}. Disabling the comparator output preserves V_{out} at the desired value.

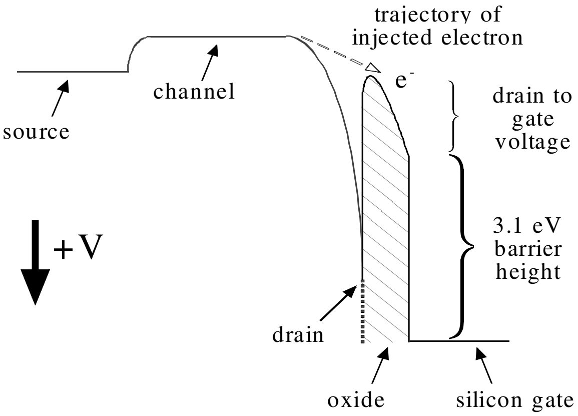

Figure 3: Energy band diagram for an n-type transistor with p-base channel implant

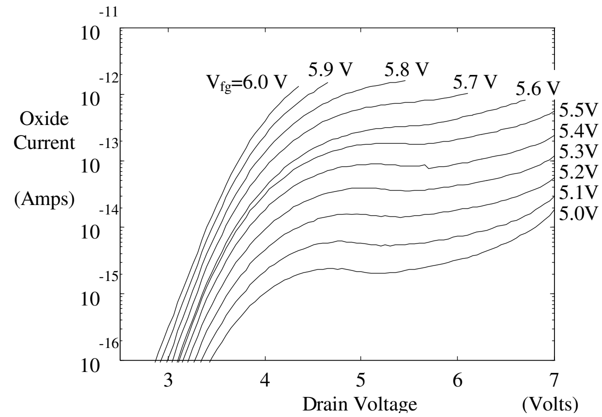

Figure 4: Oxide current vs drain voltage in a 4μ long, 6μ wide injection transistor, for several values of gate voltage

For the purposes of this paper, memory cells were tested under idealized conditions, using an off-chip voltage source to erase the memory and an LM324 amplifier to write the memory. On-chip applications currently in fabrication use conventional lightly-doped drain (LDD) high voltage transistors to select a memory cell for erasure, and a small number of multiplexed on-chip amplifiers with appropriate decoding to allow the sequential writing of memory cells.

IV. MEMORY CELL PERFORMANCE

By using a feedback loop to write the memory, write errors are kept small. Figure 6 shows the RMS value of the random write error vs V_{out}, for a write rate of 600 mV/sec. Figure 7 shows total RMS error and mean offset error versus memory write rate. The RMS error of Figure 7 is equivalent to a memory cell resolution of 14.7 to 15.4 effective bits.

Offset error depends upon the loop time constant, the injection-dependent loop slew rate, and memory cell parasitics. For low slew rates, the curve asymptotes to a ~5 mV offset. The principal reason for this offset is as follows: The negative feedback loop error signal is the injection current; it is set by the injection transistor drain voltage. However, the injection transistor parasitic drain-to-gate capacitance C_{dg} forms an alternate positive feedback pathway around the loop. This path gives the loop a hysteretic response, helping the comparator to completely switch once it begins slewing. It also causes an unavoidable offset error at the cell output. Defining ΔV_d to be the comparator output voltage swing, coupling from C_{dg} to the floating gate produces an output referenced offset error given by:

$$V_{ofs} \approx \frac{C_{dg}}{C_i} \Delta V_d$$

Power supply rejection is important to maintaining memory cell precision. Since the floating gate is positive supply (V_{dd}) referenced via transistor Q2, the cell is designed to drive V_{dd} referenced loads. Ideally, the cell output should track changes in V_{dd} with unity gain. Design parameters which achieve unity gain from V_{dd} to V_{out} can be derived to first order as follows:

1. Consider the feedback loop formed by capacitor C_i and transistors Q1 and Q2, with V_{dd} as the input and V_{out} the output. Assuming infinite loop gain, the closed loop transfer function is:

$$G_\infty = \left.\frac{v_{out}}{v_{dd}}\right|_{T=\infty} = \frac{C_i + C_p}{C_i} = 1 + \frac{C_p}{C_i} = 1 + \alpha$$

where $\alpha \equiv \frac{C_p}{C_i}$

2. We desire that the closed loop transfer function from V_{dd} to V_{out} be unity. This can be achieved by choosing a finite loop gain T to compensate α. The closed loop transfer function is:

$$\frac{v_{out}}{v_{dd}} = G_\infty \frac{T}{1+T} = \left(\frac{G_\infty}{1+\frac{1}{T}}\right) = \left(\frac{1+\alpha}{1+\frac{1}{T}}\right)$$

Choosing $\alpha = \frac{C_p}{C_i} = \frac{1}{T}$ gives $\frac{v_{out}}{v_{dd}} = 1$

3. Assuming Q1 and Q2 have identical drain conductances g_d, the low frequency loop gain T is:

$$T = \frac{g_m}{2 g_d} \frac{C_i}{C_i + C_p}$$

Where g_m is the transconductance of transistor Q2. Substituting $T = C_i/C_p$ gives the final result:

$$\frac{g_m}{2 g_d} = 1 + \frac{C_i}{C_p} \qquad (1)$$

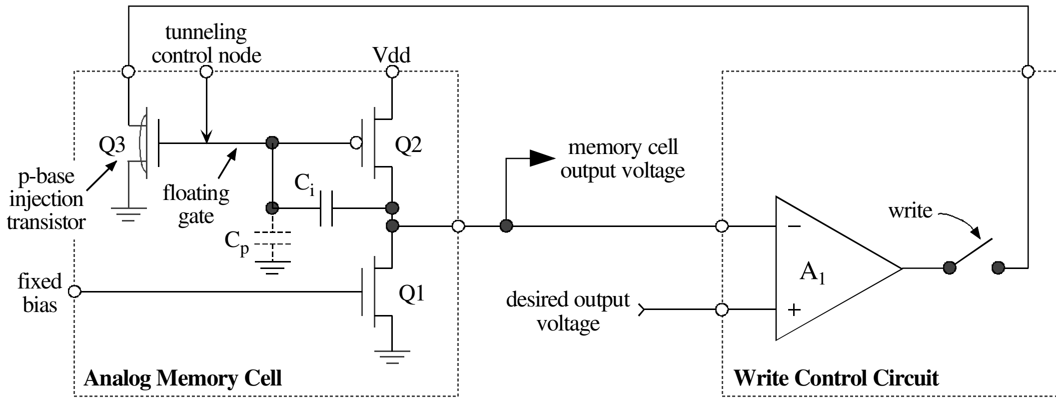

Figure 5: The analog memory cell (including parasitic capacitance C_p) with its write circuit

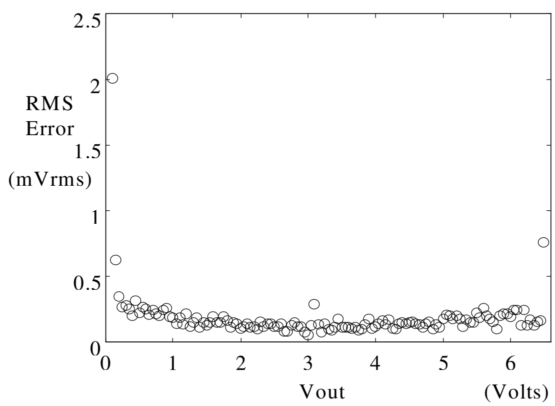

Figure 6: RMS write error vs Vout, neglecting the systematic offset, for a write rate of 600 mV/sec

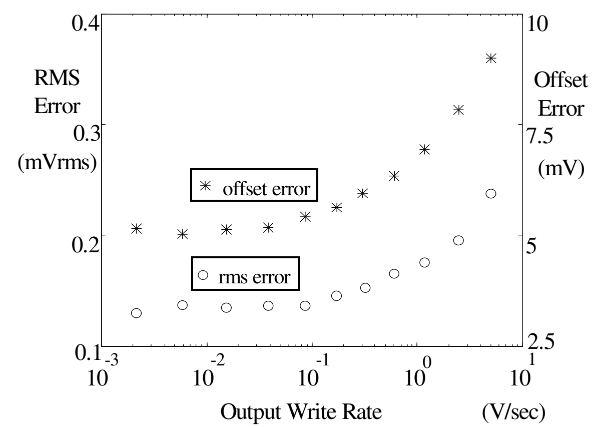

Figure 7: RMS and offset errors versus memory write rate

To first order, if the cell transistors and capacitors are sized according to (1), the memory will be insensitive to low frequency changes in V_{dd}. We have measured 74 dB supply rejection on the test chip, corresponding to 15 bit output precision for $\pm$ 5% variations in supply voltage. Of course, V_{out} remains sensitive to changes in bias voltage, but since bias is ground referenced, proper design practices can insure its stability.

V. CONCLUSION

We have developed a 3-transistor analog memory cell with 14-bit resolution, nonvolatile storage, rail-to-rail output range, nW power consumption, and 74 dB supply rejection. Its support for simultaneous memory reading and writing allows accurate feedback control of the write process. The cell can be fabricated in a standard 2 µm n-well BiCMOS process available from MOSIS. We believe this cell to be well suited for long-term learning in silicon neural networks, and as a means for storing precise analog voltages in MOS integrated circuits.

VI. ACKNOWLEDGEMENTS

This work was sponsored by a TRW graduate fellowship and by the Office of Naval Research, ARPA, and the Beckman Foundation.

V. REFERENCES

[1] B. Hochet, *et al.*,“Implementation of a Learning Kohonen Neuron Based on a New Multilevel Storage Technique,” *IEEE J. Solid-State Circuits*, vol. 26, no. 3, 1991, pp. 262-267.

[2] P. Hollis and J. Paulos, “A Neural Network Learning Algorithm Tailored for VLSI Implementation,” *IEEE Tran. Neural Networks*, vol. 5, no. 5, 1994, pp. 784-791.

[3] J. Lazzaro, *et al.*,“Systems Technologies for Silicon Auditory Models,” *IEEE Micro*, vol. 14, no. 3, 1994, pp. 7-15.

[4] T. Allen, *et al.*,“Writable Analog Reference Voltage Storage Device,” *U.S. Patent No. 5,166,562*, 1991

[5] M. Lenzlinger and E. H. Snow, “Fowler-Nordheim Tunneling into Thermally Grown SiO_2,” *J. Applied Physics*, vol. 40, no. 6, 1969, pp. 278-283.

[6] C. Hu, “Future CMOS Scaling and Reliability,” *Proc. of the IEEE,* vol. 81, no. 5, 1993, pp. 682-689.