A SIMULATION MODEL FOR FLOATING-GATE MOS SYNAPSE TRANSISTORS

Kambiz Rahimi, Chris Diorio, Cecilia Hernandez, M. Dean Brockhausen
kamr@ee.washington.edu, diorio@cs.washington.edu
University of Washington, Seattle, Washington

ABSTRACT

We propose an empirical simulation model for *p*-channel floating-gate MOS synapse transistors. Because our model requires only a transistor and controlled sources, and does not use the MOSFET's channel potential in its description, we can apply the model in any SPICE circuit simulator. The model parameters derive from simple oxide-current measurements. We present fit parameters from MOSFETs with 70Å oxides in a 0.35µm process, and verify our model by comparing simulations and measured data from a capacitive-feedback CMOS operational amplifier.

1. INTRODUCTION

Synapse transistors [1] are a new class of devices that are rapidly gaining acceptance as standard elements in MOS circuit design. Researchers have used synapse transistors in applications ranging from weight storage in a learning array [2], to trimming a digital-to-analog converter [3], to nulling input offsets in a capacitive-feedback operational amplifier [4], to unsupervised vector quantization [5]. Unfortunately, due to the lack of a simple and accurate simulation model for the synaptic devices, many of these circuits were designed using equation-based modeling and heuristics.

We show a layout and a band diagram for a synapse transistor in Fig. 1. The synapse is a four-terminal device comprising two *p*FETs: A floating-gate *p*FET for electron injection and readout, and a shorted *p*FET for electron tunneling. The charge on the floating gate represents a nonvolatile analog weight. We add electrons to the floating gate using impact-ionized hot-electron injection (IHEI) [6], and remove electrons by means of electron tunneling [7].

To design large-scale synaptic circuits, engineers need an accurate SPICE simulation model for the synaptic devices. This model should include the IHEI and tunneling currents, parasitic capacitances, and excess carriers generated by impact ionization. Hasler et al. have reported a synapse-transistor SPICE macromodel [8] that uses additional FET devices with parametric biases to model IHEI and tunneling. Unfortunately, this approach does not model the IHEI and tunneling processes over their full operating range, nor does it model the tunneling implant's nonlinear MOS capacitance.

To complicate matters, the physics-based equations that describe IHEI [9] are not amenable to a simple SPICE model, both for reasons of complexity and because they use channel potential as an explicit variable. We propose an alternative, empirical expression for the IHEI process that uses only drain, gate, and source potential as parameters, and fits the measured data over the entire range of device terminal voltages. We use this empirical expression, and simplified versions of known equations for impact ionization and tunneling, to construct a SPICE macromodel for the synapse.

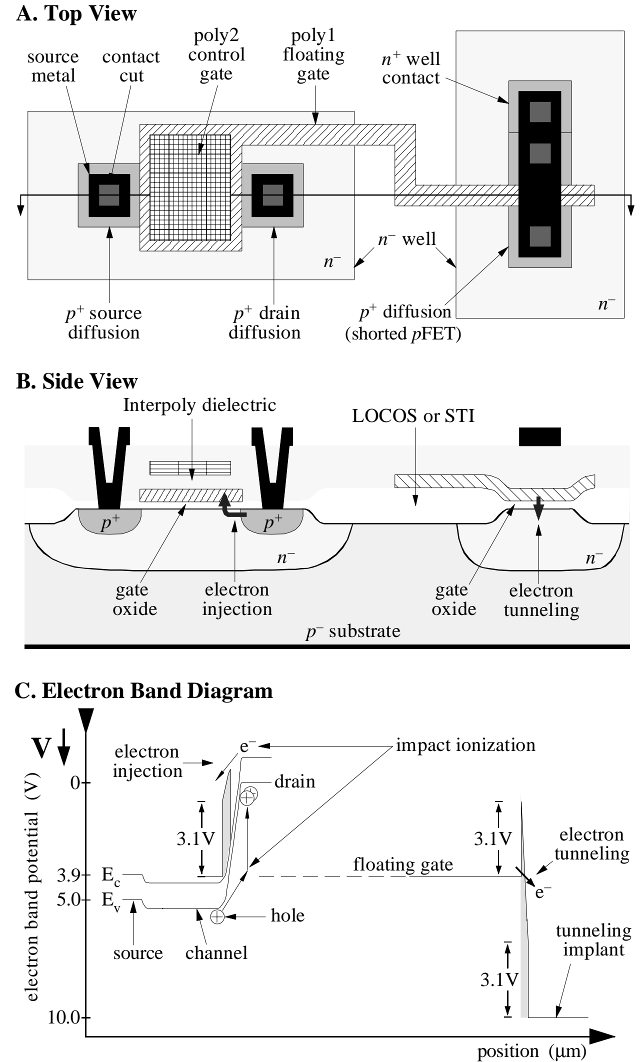


Fig. 1. A *p*FET synapse, showing the electron tunneling and injection locations. We aligned the three diagrams vertically, drew diagrams A and C to scale, exaggerated the vertical scale in diagram B, and assumed subthreshold operation ($I_s < 100\,$nA). Although the gate oxide's band diagram projects vertically, to better explain the IHEI process we rotated it by 90° and drew it in the channel direction. We decrease the synapse weight by tunneling electrons off the floating gate; we increase the weight by IHEI in the channel-to-drain depletion region. The poly2 control gate is optional, depending on process availability.

2. SYNAPSE TRANSISTORS

Synapse transistors are conventional transistors with three additional attributes: Nonvolatile analog weight storage, locally computed bidirectional weight updates, and simultaneous memory reading and writing. We and others construct our synapse transistors using floating-gate *p*-channel MOSFETs, where the charge stored on the floating gate represents the stored analog weight. Electron tunneling and IHEI allow bidirectional memory updates that depend on local terminal voltages. Because we can modify the memory during normal transistor operation, the synapse allows simultaneous memory reading and writing.

We use Fowler-Nordheim tunneling to increase the charge on the floating gate. A voltage difference between the tunneling junction (the shorted *p*FET in Fig. 1) and the floating gate causes the electrons to tunnel from the floating gate, through the *p*FET's gate oxide, to the n^- well. The magnitude of this tunneling current depends on the oxide voltage. We approximate this current by:

$$I_{tun} = -I_{tun0} WL \exp\left(-\frac{V_f}{V_{ox}}\right) \quad (1)$$

where I_{tun0} is a pre-exponential current, V_{ox} is the voltage across the oxide, and V_f is a constant that varies with oxide thickness [10]. *W* and *L* are the width and length of the tunneling *p*FET (in microns), respectively.

We use IHEI to decrease the charge on the floating gate. Channel holes, accelerated in the transistor's channel-to-drain depletion region, collide with the semiconductor lattice. When the channel-to-drain electric field is large, a fraction of these holes collide with sufficient energy to liberate additional electron–hole pairs [11]. The ionized electrons, promoted to their conduction band by the collision, are expelled from the drain by this same channel-to-drain field. Electrons expelled with more than 3.1 eV of kinetic energy can, if scattered upward into the gate oxide, overcome the 3.1V difference in electron affinities between the Si and SiO_2, inject into the oxide conduction band, and be collected by the floating gate.

We know of no simple model, in the existing literature, that describes IHEI in a format suitable for SPICE simulation. In particular, we know of no model that describes IHEI solely in terms of device terminal voltages. Consequently, we model IHEI with the following semi-empirical equation, in which we determine the fit constants experimentally:

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

where I_s is the source current, V_{gd} is the *p*FET's gate-to-drain voltage, and V_{sd} is its source-to-drain voltage. α, β, and δ are fit parameters. The parameter $\lambda = 1$; we include it in Eqn. (2) for units consistency.

Impact ionization in the channel-to-drain depletion region generates additional electron–hole pairs; the electrons tend to be collected by the well/source terminal, and the holes by the

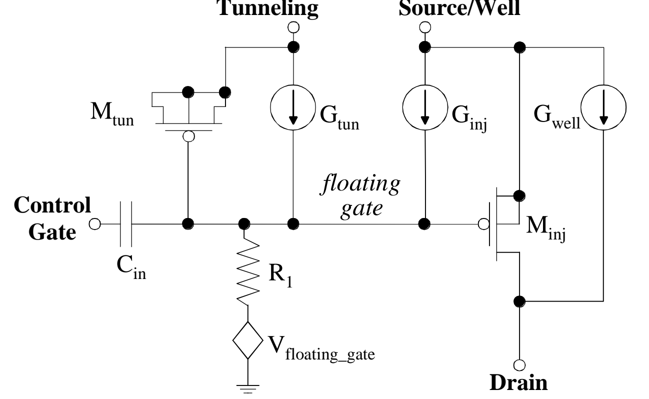


Fig. 2. Synapse-transistor circuit macromodel. M_{tun} models the tunneling-junction MOS capacitance; G_{tun} models the tunneling current. G_{inj} and G_{well} model the IHEI gate current and the impact-ionization current, respectively. R_1 and $V_{\text{floating_gate}}$ aid numerical convergence by providing an (artificial) dc path to ground.

*p*FET's drain. Consequently, our macromodel includes a current source from well to drain. We model the impact current using a semi-empirical expression based on NMOS substrate-current models [12, 13]:

$$I_b = \eta I_s \left(\gamma V_{sd} - \kappa V_{sg} + V_t\right) \exp\left(\frac{-\lambda}{\gamma V_{sd} - \kappa V_{sg} + V_t}\right) \quad (3)$$

where I_b and I_s are the well and source currents, and V_{sd} and V_{sg} are the source-to-drain and source-to-gate voltages, respectively. η, γ, κ and λ are fit parameters.

3. THE MODEL

Fig. 1 shows that the IHEI gate current depends on a synapse *p*FET's channel-to-drain potential. The key observation that allowed us to develop our SPICE macromodel is that we can describe the IHEI current in terms of gate-to-drain potential rather than channel-to-drain potential. The reasoning is as follows: IHEI happens only when a synapse *p*FET has a large channel-to-drain potential, so during IHEI the drain must be pinched off. In pinchoff, the surface potential at the drain end of the channel follows the gate potential linearly. Consequently, gate-to-drain potential and channel-to-drain potential are related by a scale factor that we can absorb into the fit constants of our IHEI equation. This observation allows us to describe IHEI in terms of the synapse *p*FET's terminal voltages, and to model it using a voltage-controlled current source.

Fig. 2 shows our synapse-transistor macromodel. The tunneling implant is a MOS capacitor; we model the tunneling current by a voltage-dependent current source G_{tun}. We model the injection current by a voltage-dependent current source G_{inj}, and the impact-ionization current by a voltage-dependent current source G_{well}. These current sources directly implement Eqns. (1)–(3). We add a voltage-dependent voltage source $V_{\text{floating_gate}}$ from ground to the floating gate, through an arbitrarily valued resistor R_1. There is no current

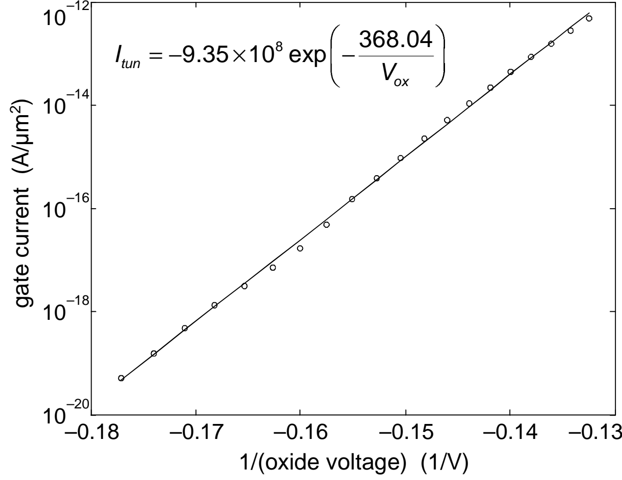


Fig. 3. Tunneling (gate) current I_g versus $-1/V_{ox}$, for a synapse fabricated in a 0.35 μm CMOS process. V_{ox} is the potential between the tunneling junction and the floating gate. We normalized the gate current to the tunneling-junction (gate oxide) area.

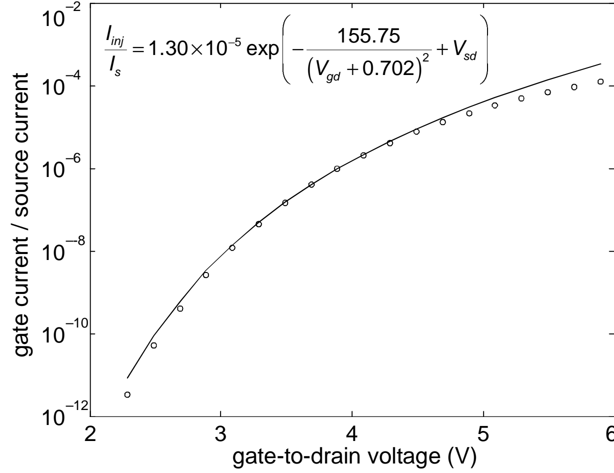


Fig. 4. IHEI efficiency (gate current I_g divided by source current I_s), versus gate-to-drain potential V_{gd}, for a synapse fabricated in a 0.35 μm CMOS processes. We held the transistor's source current at 10nA during this experiment.

through R_1, because $V_{floating_gate}$ tracks the floating-gate voltage itself. This artificial DC path to ground aids numerical convergence in some SPICE simulators.

We extract our equation parameters by measuring oxide currents and well currents in synaptic test structures. In Figs. 3–6, we show measured data from a 0.35μm CMOS process. We include fits from Eqns. (1)–(3). We have found that these equations accurately model synaptic currents in CMOS processes ranging from 2μm down to 0.25μm.

4. DEMONSTRATING THE MODEL

We fabricated a 0.35μm CMOS autonulling amplifier as a demonstration vehicle for our SPICE macromodel. We show the circuit in Fig. 7. Its operation is similar to an amplifier described by Hasler et al. in [4].

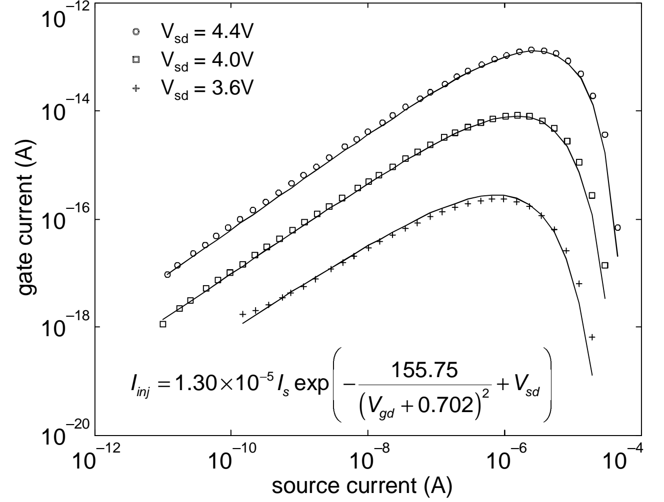


Fig. 5. IHEI gate current versus source current, for three different source-to-drain voltages. The synapse for this experiment was on the same chip we used for the experiment of Fig. 4.

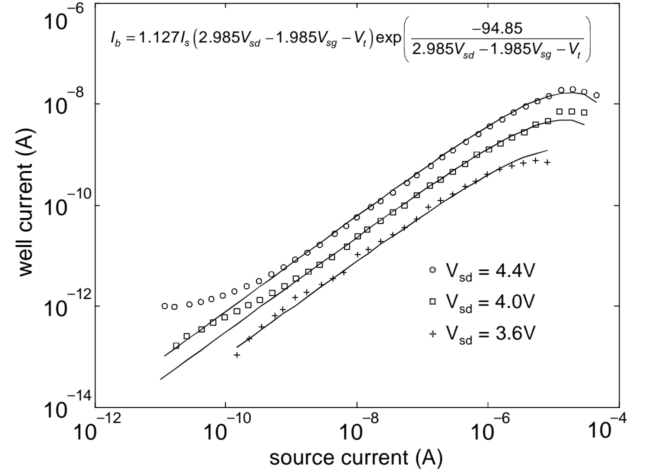
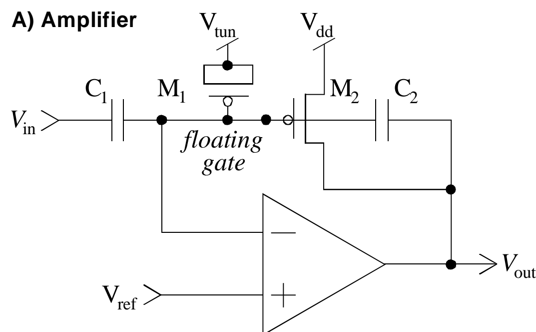
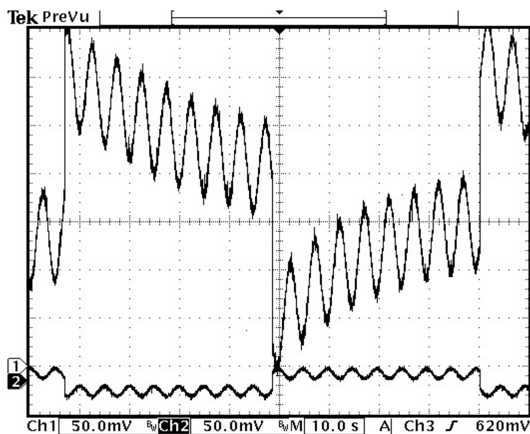


Fig. 6. Well current versus source current, measured during the experiment of Fig. 5. The well current comprises primarily electrons generated by impact ionization at the synapse's drain. V_t is the transistor's threshold voltage (positive valued).

The circuit uses capacitive feedback, with a synapse transistor to adjust the voltage on the op-amp's (floating) inverting input. Capacitors C_1 and C_2 set the closed-loop gain. We apply a fixed high voltage to V_{tun}, causing a small electron current to tunnel off the floating gate. The op-amp compensates by lowering V_{out}, causing transistor M_2 to inject electrons back onto the floating gate. V_{out} stabilizes when the tunneling and injection currents are equal and opposite, and the floating-gate voltage is equal (in a DC sense) to V_{ref}. V_{tun} sets both the quiescent value of the op-amp's output and its adaptation rate: If we raise V_{tun}, we lower V_{out} and increase the adaptation rate. We describe the low-frequency response using the term "adaptation" rather than "time constant", because tunneling and injection are nonlinear processes, so the adaptation does not follow typical time-constant dynamics.

We simulated this circuit using our floating-gate SPICE macromodel; we show measured and simulated waveforms in Fig. 7. We used the BSIM3 transistor models provided by the foundry. The circuit gain is roughly 10.





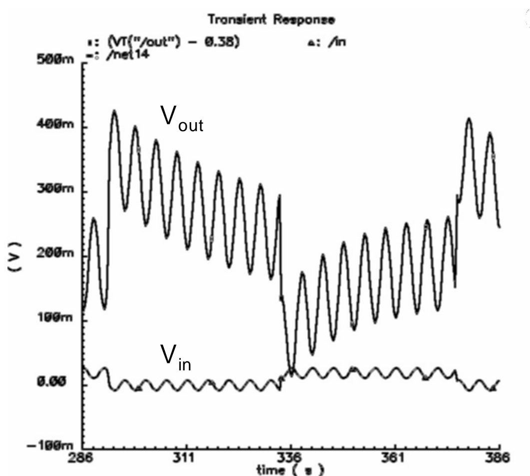


Fig. 7. (A) Autonulling amplifier. The frequency response is band-pass, with the low-frequency corner set by V_{tun}. (B) Measured response to a 0.2Hz, 15mV sinewave superimposed on a 0.012Hz, 19mV squarewave. The amplifier passes the 0.2Hz sinewave, and attenuates the 0.012Hz squarewave. The input is ground-centered; the output has a 380mV DC offset (adjustable using V_{tun}, as described in Section 4). (C) A SPICE transient simulation of the same circuits, using our floating-gate macromodel.

5. CONCLUSION

We have described a SPICE macromodel for a *p*FET synapse transistor that accurately mimics the synapse operation over a wide range of device terminal voltages. Our macromodel uses semi-empirical equations to model the tunneling and IHEI processes implicit in *p*FET synapses, and is compatible with standard SPICE circuit simulators. We can use this macromodel to simulate large-scale synaptic systems.

The gate oxides in our chosen 0.35µm process are roughly 70Å thick. Reducing the oxide thickness causes the floating gates to leak. This problem is not unique to synapse transistors—it affects all nonvolatile floating-gate memory devices. Because we want our synaptic systems to retain their nonvolatile analog memory, we anticipate that our synapses will use the 70Å oxide available in most dual-gate-oxide CMOS processes (i.e. the synapse oxide thickness will not shrink with process scaling). Consequently, we do not include a leakage term in our SPICE macromodel. Also, because we will continue using 3.3V transistors with 70Å oxides for our synapses, we do not anticipate changing the fit parameters we have shown in this paper for the foreseeable future.

6. REFERENCES

[1] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, “Floating-gate MOS synapse transistors,” in T. S. Lande (ed.), *Neuromorphic Systems Engineering: Neural Networks in Silicon,* Boston, MA: Kluwer Academic Publishers, pp. 315–337, 1998.

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

[3] M. Figueroa, J. Hyde, T. Humes, and C. Diorio, “A floating-gate trimmable high-resolution DAC in standard 0.25µm CMOS,” *Proc. IEEE Nonvolatile Semiconductor Memory Workshop,* Monterey, CA, pp. 46–47, 2001.

[4] P. Hasler, B. A. Minch, and C. Diorio, “An autozeroing floating-gate amplifier," *IEEE Trans. Circuits and Systems II*, vol.48, no. 1, pp. 74–82, 2001.

[5] D. Hsu, M. Figueroa, and C. Diorio, “A silicon primitive for competitive learning,” *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. Dietterich, and V. Tresp, eds., MIT Press, pp. 713–719, 2001.

[6] P. Hasler, *Foundations of Learning in Analog VLSI,* Ph.D. thesis, Department of Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 1997.

[7] M. Lenzlinger and E. H. Snow, “Fowler-Nordheim tunneling into thermally grown SiO2,” *J. of Applied Phys*., vol. 40, no. 6, pp. 278–283, 1969.

[8] A. Low and P. Hasler, “Cadence-based simulation of floating-gate circuits using the EKV model,” *Proc. IEEE Midwest Symposium on Circuits and Systems*, Las Cruces, NM, 1999.

[9] P. Hasler, A. Andreou, C. Diorio, B. A. Minch, and C. Mead, “Impact ionization and hot-electron injection derived consistently from Boltzmann transport,” *VLSI Design,* vol. 8, no. 14, pp. 455–461, 1998.

[10] Y. Xu, *Electron transport through thin film amorphous silicon—A tunneling study*, Ph.D. Thesis, Stanford University, 1992.

[11] T. C. Ong; P. K. Ko; and C. Hu, “Hot-carrier current modeling and device degradation in surface-channel p-MOSFETs” *IEEE Trans. Electron Devices*, vol. 37, no. 7, pp. 1658–1666, 1990.

[12] J. S. Kolhatkar and A. K. Dutta, “A new substrate current model for sub micron MOSFETs” *IEEE Tran. Electron Devices*, vol. 47, no. 4, pp. 861–863, 2000.

[13] N. D. Arora and M. S. Sharma, “MOSFET substrate current model for circuit simulation,” *IEEE Trans. Electron Devices*, vol. 8, no. 6, pp. 1392–1398, 1991.