Analog Integrated Circuits and Signal Processing, 13, 153–166 (1997)

A Complementary Pair of Four-Terminal Silicon Synapses

CHRIS DIORIO, PAUL HASLER, BRADLEY A. MINCH, CARVER MEAD

chris@pcmp.caltech.edu

Physics of Computation Laboratory, California Institute of Technology, Pasadena, CA 91125, USA

Received April 9, 1996; Accepted June 28, 1996

Abstract. We have developed a complementary pair of *p*FET and *n*FET floating-gate silicon MOS transistors for analog learning applications. The memory storage is nonvolatile; hot-electron injection and electron tunneling permit bidirectional memory updates. Because these updates depend on both the stored memory value and the transistor terminal voltages, the synapses can implement a learning function. We have derived a memory-update rule for both devices, and have shown that the synapse learning follows a simple power law. Unlike conventional EEPROMs, the synapses allow simultaneous memory reading and writing. Synapse transistor arrays can therefore compute both the array output, and local memory updates, in parallel. We have fabricated prototype synaptic arrays; because the tunneling and injection processes are exponential in the transistor terminal voltages, the write and erase isolation between array synapses is better than 0.01 %. The synapses are small, and typically are operated at subthreshold current levels; they will permit the development of dense, low-power silicon learning systems.

Key Words: synapse transistor, silicon learning, floating-gate MOSFET

1. Introduction

Our goal is to develop silicon learning systems. One impediment to achieving this goal has been the lack of a simple circuit element combining nonvolatile analog memory storage with locally computed memory updates. Existing circuits [1, 2] are typically large and complex; the nonvolatile floating-gate devices, such as EEPROM transistors, typically are optimized for binary-valued data storage [3], and do not compute their own memory updates. Although floating-gate transistors can provide nonvolatile analog memory storage [4, 5], because writing the memory entails the difficult process of moving electrons through SiO_2, these devices have not seen wide use as memory elements in silicon learning systems.

We have fabricated *synapse transistors* that not only possess nonvolatile analog memory storage, and compute locally their own memory updates, but also permit simultaneous memory reading and writing, and compute locally the product of the stored memory value and the applied input. To ensure nonvolatile memory, we employ standard floating-gate transistors; in addition, we adapt the physical processes that write the memory to perform a learning function. Although the SiO_2 electron transport still is difficult, and does require high voltages, because our devices integrate both memory storage and local computation within a single device, we expect them to find wide application in silicon learning systems.

We call our devices *silicon synapses* because, like a neural synapse, they compute the product of the stored analog memory value and the applied input. Also like the neural synapse, they can learn from the input signal, without interrupting the ongoing computation. Although we do not believe that a single device can model completely the complex behavior of a neural synapse, our single-transistor synapses do implement a learning function. With them, we intend to build autonomous learning systems in which both the system outputs, and the memory updates, are computed locally and in parallel.

We have described previously [6, 7, 8] the four-terminal *n*FET synapse discussed here. We have also described an analog memory cell that employs the *n*FET device [9], and an auto-zeroing amplifier that employs the *p*FET device [10]. We here present the four-terminal *n*FET synapse in greater detail than we did previously, and for the first time present the four-terminal *p*FET synapse. We have also described a three-terminal *n*FET synapse [11]. Although the four-terminal synapses require slightly more layout area than does this three-terminal device, the additional terminal permits greater control over the write and erase processes.

2. The Synapses

The *n*FET and *p*FET synapses each possess a poly1 floating gate, a poly2 control gate, and a lightly doped *n*-well tunneling implant. Both synapses use hot-electron injection [12] to add electrons to their floating gates, and Fowler–Nordheim (FN) tunneling [13] to remove the electrons. The *n*FET synapse differs from a conventional *n*-type MOSFET by its use of a moderately-doped channel implant. This implant facilitates hot-electron injection. The *p*FET synapse, by contrast, achieves a sufficient hot-electron gate-current using a conventional *p*-type MOSFET; no special channel implant is required. Both synapses have been fabricated in the 2µm *n*-well Orbit BiCMOS process available from MOSIS.

In both synapses, the memory is stored as floating-gate charge. Either channel current or channel conductance can be selected as the synapse output. Inputs typically are applied to the poly2 control gate, which capacitively couples to the poly1 floating gate. From the control gate's perspective, altering the floating-gate charge shifts the transistor's threshold voltage V_t, enabling the synapse output to vary despite a fixed-amplitude control-gate input.

We typically operate the synapses in their subthreshold regime [14], and typically select either drain current or source current as the synapse output. We have chosen subthreshold operation for three reasons. First, because the power consumption of a subthreshold MOSFET is typically less than 1µW, our learning systems will operate at low power. Second, because the channel current in a subthreshold MOSFET is an exponential function of the gate voltage, only small quantities of oxide charge are required for learning. Third, the channel current in a subthreshold floating-gate MOSFET is the product of the stored memory value and the applied input:

$$I_s = I_o\, e^{\frac{\kappa Q_{fg}}{C_T U_t}}\, e^{\frac{\kappa C_{in} V_{in}}{C_T U_t}} = I_o\, e^{\frac{\kappa Q_{fg}}{Q_T}}\, e^{\frac{\kappa' V_{in}}{U_t}} \quad (1)$$

$$= I_m\, e^{\frac{\kappa' V_{in}}{U_t}} \quad (2)$$

where I_s is the source current, I_o is the pre-exponential current, κ is the floating-gate to channel-surface coupling coefficient, Q_{fg} is the floating-gate charge, C_T is the total capacitance seen by the floating gate, U_t is the thermal voltage kT/q, C_{in} is the input (poly1 to poly2) coupling capacitance, V_{in} is the signal voltage applied to the poly2 input, $Q_T \equiv C_T U_t$, and $\kappa' \equiv \kappa C_{in}/C_T$.

The quantity I_m is the stored memory; its value changes with synapse use. The synapse output is the product of I_m and the exponentiated gate input. Because the tunneling and injection gate currents vary with the synapse terminal voltages and channel current, I_m varies with the terminal voltages, which are imposed on the device, and with the channel current, which is the synapse output. Consequently, the synapses exhibit a type of learning by which their future output depends on both the applied input and the present output.

2.1. The nFET Synapse

Top and side views of the *n*FET synapse are shown in Fig. 1. Its principal features are the following:

- Electrons tunnel from the floating gate, through the 350Å gate oxide, to the tunneling implant. A high voltage applied to the tunneling implant provides the oxide E-field required for tunneling. To prevent reverse-bias *pn*-junction breakdown, the tunneling implant is a lightly doped n^- well. Because tunneling removes electrons from the floating gate, from the control gate's perspective tunneling reduces the transistor's threshold voltage V_t.
- Electron tunneling is enhanced where the poly1 floating gate overlaps the heavily doped well contact, for two reasons. First, the gate cannot deplete the n^+ contact, whereas it does deplete the n^- well. Thus, the oxide E-field is higher over the n^+. Second, enhancement at the gate edge further augments the oxide field.
- Electrons inject from the channel-to-drain space-charge layer to the floating gate. To facilitate injection, we apply a *p*-type bipolar-transistor base

implant to the MOS transistor channel. As a result, the channel-to-drain depletion region approximates a one-sided step junction, increasing the injection likelihood. The channel implant also raises the transistor threshold voltage V_t, favoring the collection of the injected electrons by the floating gate. Because injection adds electrons to the floating gate, from the control gate's perspective injection increases the transistor's threshold voltage V_t.

- Oxide uniformity and purity determine the initial matching between synapses, as well as the learning-rate degradations due to oxide trapping. We therefore use the thermally grown gate oxide for all SiO_2 carrier transport.

2.2. The pFET Synapse

Top and side views of the *p*FET synapse are shown in Fig. 2. Its principal features are the following:

- Electrons tunnel from the floating gate to the tunneling implant through 350 Å gate oxide. The tunneling implant is identical to that used in the *n*FET synapse. As in the *n*FET synapse, tunneling removes electrons from the floating gate. However, because the *p*FET and *n*FET synapses are complementary, from the control gate's perspective tunneling has the opposite effect on the *p*FET synapse—it increases, rather than decreases, the transistor's threshold voltage V_t.
- Hole impact ionization, at the channel–drain junction, generates the electrons for oxide injection. Channel holes, accelerated in the channel-to-drain E-field, collide with the semiconductor lattice to produce additional electron–hole pairs. The liberated electron, promoted to its conduction band by the collision, is expelled rapidly from the drain region by this same channel-to-drain E-field.
- Impact-generated electrons that acquire more than 3.2 eV of kinetic energy can, if scattered upward into the gate oxide, inject from the channel-to-drain space-charge layer onto the floating gate. As in the *n*FET synapse, injection adds electrons to the floating gate; because the device is a *p*FET, however, from the control gate's perspective injection reduces the transistor's threshold voltage V_t.
- Like the *n*FET synapse, the *p*FET synapse uses gate oxide for all SiO_2 carrier transport.

3. The Gate Current Equation

We intend to use our synapses to build silicon learning systems. Because the learning behavior of any such system is determined in part by the tunneling and injection processes that alter the stored memory, we have investigated these processes over the subthreshold operating range.

3.1. The Tunneling Process

The tunneling process, for the *n*FET and *p*FET synapses, is shown in the energy-band diagrams [15] of Figs. 1 and 2, respectively. In the FN-tunneling process, the potential difference between the tunneling implant and the floating gate reduces the effective oxide thickness, facilitating electron tunneling from the floating gate, through the SiO_2 barrier, into the oxide conduction band. These electrons are then swept by the oxide E-field over to the tunneling implant. We apply positive high voltages to the tunneling implant to promote electron tunneling.

3.2. The Tunneling Equation

The data of Fig. 3 show the tunneling gate current versus the oxide voltage, where we define *oxide voltage* to be the potential difference between the tunneling implant and the floating gate. We fit these data with a modified FN fit, which employs a built-in potential, $\mathrm{V_{bi}}$, to account for oxide traps:

$$I_g = \xi\left(V_{ox} + \mathrm{V_{bi}}\right)^2 e^{-\frac{\mathrm{V_o}}{V_{ox}+\mathrm{V_{bi}}}} \tag{3}$$

where I_g is the gate current, V_{ox} is the oxide voltage, and ξ, $\mathrm{V_{bi}}$, and $\mathrm{V_o}$ are constants. For comparison, we also show the conventional FN fit [13, 16]:

$$I_g = \varphi V_{ox}^2 e^{-\frac{\mathrm{V_f}}{V_{ox}}} \tag{4}$$

where $\mathrm{V_f}$=928 V is consistent with a recent survey [17] of SiO_2 tunneling, given the synapse transistor's 350 Å gate oxide, and φ is a fit parameter.

The data of Fig. 3 are normalized to the gate-to-n^+ edge length, in lineal microns. The reason is that the floating gate induces a depletion region in the lightly doped n^- tunneling implant, reducing the effective oxide voltage, and therefore also the tunneling current. Because the gate cannot appreciably deplete the n^+ drain contact, the oxide field is higher where the

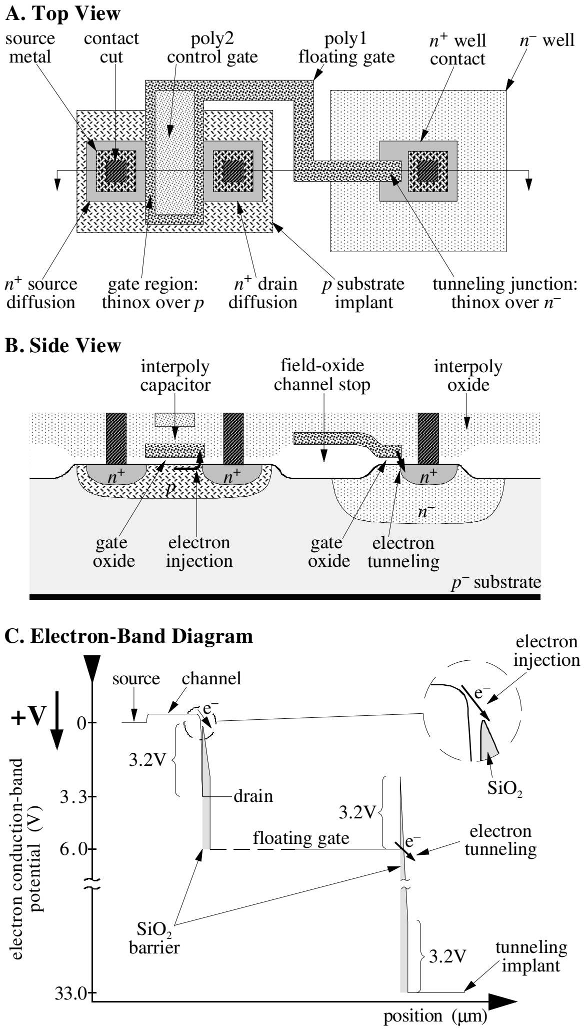

Fig. 1. The nFET synapse, showing the electron tunneling and injection locations. The three diagrams are aligned vertically. Diagrams *A* and *C* are drawn to scale; the vertical scale in diagram *B* has been exaggerated for clarity. In the 2 μm Orbit process, the synapse length is 44 μm, and the width is 17 μm. All voltages in the conduction-band diagram are referenced to the source potential, and subthreshold ($I_s < 100$ nA) operation is assumed. Although in the band diagram the gate oxide actually projects into the plane of the page, for convenience it has been rotated 90° and drawn in the channel direction. When compared with a conventional nFET, the synapse transistor's additional p-type substrate doping quadruples the MOS gate-to-channel capacitance. With a 25 fF interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.2. To facilitate testing, we enlarged the interpoly capacitor to 1 pF in the test device, thereby increasing the coupling to 0.8.

self-aligned floating gate overlaps the n^+. Because the tunneling current is exponential in the oxide voltage, tunneling in the synapse transistors is primarily an edge phenomenon.

3.3. *The Hot-Electron Injection Process*

The hot-electron injection process [18], for both the nFET and pFET synapses, is shown in the energy-band diagrams of Figs. 1 and 2, respectively. Electrons inject from the transistor channel, over the 3.2 V Si–SiO_2 work-function barrier, into the oxide conduction band. These electrons then are swept by the oxide E-field over to the floating gate. Successful injection, for both the nFET and pFET synapses, requires that the following three conditions be satisfied: (1) the electrons must possess the 3.2 eV required to surmount the Si–SiO_2 barrier, (2) the electrons must scatter upward into the gate oxide, and (3) the oxide E-field must be in the proper direction to transport the electrons to the floating gate.

nFET Injection: In a conventional n-type MOSFET, requirements (1) and (2) are readily satisfied. We merely operate the transistor in its subthreshold regime, with a drain-to-source voltage greater than about 3 V. Because the subthreshold channel-conduction band is flat, the channel-to-drain transition is steep, implying a large electric field. Channel electrons are accelerated rapidly in this field; a fraction of them acquire the 3.2 eV required for hot-electron injection. A fraction of these 3.2 eV electrons naturally scatter, by means of collisions with the semiconductor lattice, upward into the gate oxide.

It is principally requirement (3) that prevents injection in a conventional nFET. Subthreshold operation typically implies gate voltages < 0.8 V. With the transistor drain at 3 V, and the gate at 0.8 V, the drain-to-gate electric field opposes transport of the injected electrons to the floating gate. The electrons are instead returned to the transistor drain.

To promote the transport of injected electrons to the floating gate, we increase the synapse transistor's bulk channel doping. The additional dopant increases the channel surface–acceptor concentration, raising the transistor's threshold voltage from 0.8 V to 6 V. With the synapse drain at 3 V, and the gate at 6 V, the channel current is subthreshold, but now the oxide E-field sweeps injected electrons over to the floating gate, rather than returning them to the silicon surface.

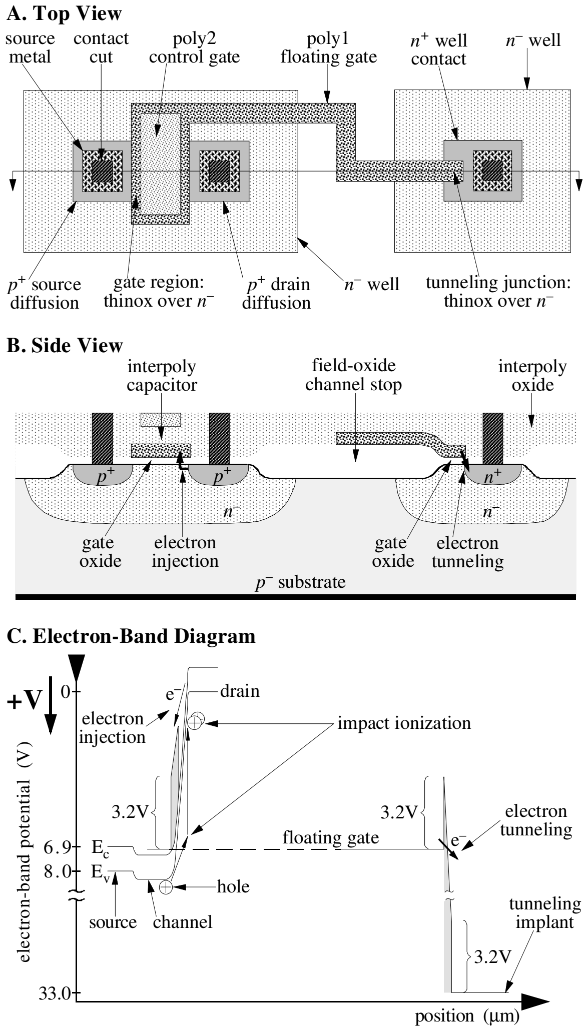

Fig. 2. The *p*FET synapse, showing the electron tunneling and injection locations. The well contact is not shown. Like those in Fig. 1, the three diagrams are aligned vertically, diagrams *A* and *C* are drawn to scale, the vertical scale in diagram *B* has been exaggerated, the voltages in the band diagram are referenced to the source potential, and subthreshold ($I_s < 100\,\text{nA}$) operation is assumed. In the 2 μm Orbit process, the synapse length is 56 μm, and the width is 16 μm. Whereas the tunneling process is identical to the *n*FET synapse, the injection process is different. Because the charge carriers are holes, we generate electrons by means of hole impact ionization at the drain. Refer to the text for a description of the impact-ionized electron-injection process. With a 20 fF interpoly capacitor as shown, the coupling coefficient between the poly2 control gate and the poly1 floating gate is only 0.25. We enlarged the interpoly capacitor to 1 pF in the test device, thereby increasing the coupling to 0.8.

pFET Injection: Because the *p*FET channel current comprises holes, the *p*FET hot-electron injection process is different from that in the *n*FET. We accelerate channel holes in the channel-to-drain depletion region of a subthreshold *p*FET. A fraction of these holes collide with the semiconductor lattice at energies sufficient to liberate additional electron–hole pairs. The ionized electrons, promoted to their conduction band by the collision, are expelled from the drain by the channel-to-drain E-field. If these ionized electrons are expelled with more than 3.2 eV of kinetic energy, they can inject onto the floating gate.

Like in the *n*FET synapse, in the *p*FET synapse injection requirements (1) and (2) are easily satisfied. We merely operate the transistor in its subthreshold regime, with a drain-to-source voltage greater than about 6 V. The higher drain-voltage requirement, when compared with the *n*FET synapse, is a consequence of the two-step injection process.

Because, in a subthreshold *p*FET, the gate-to-source voltage typically is less than 1 V, if the drain-to-source voltage exceeds 6 V, the gate voltage must exceed the drain voltage by at least 5 V. The oxide E-field supports strongly the transport of injected electrons to the floating gate, and requirement (3) is always satisfied. Unlike conventional *n*FET transistors, conventional *p*FET transistors naturally inject electrons onto their floating gates (at sufficient drain-to-source voltages); we do not need to add a special channel implant to facilitate injection.

3.4. *The Injection Equation*

The data of Fig. 4 show injection efficiency (gate current divided by source current), versus drain-to-channel potential, for both the *n*FET and *p*FET synapses. The data are plotted as efficiency because, for both devices, the gate current is linearly proportional to the source current over the entire subthreshold range. Because the hot-electron injection probability varies with channel potential, we reference all terminal voltages to the channel. We can re-reference our results to the source terminal using the relationship between source and channel potential in a subthreshold MOS transistor [19, 20]:

$$\Psi \approx \kappa V_{fg} + \Psi_o \tag{5}$$

where Ψ is the channel-surface potential, V_{fg} is the floating-gate voltage, κ is the gate-to-channel–surface

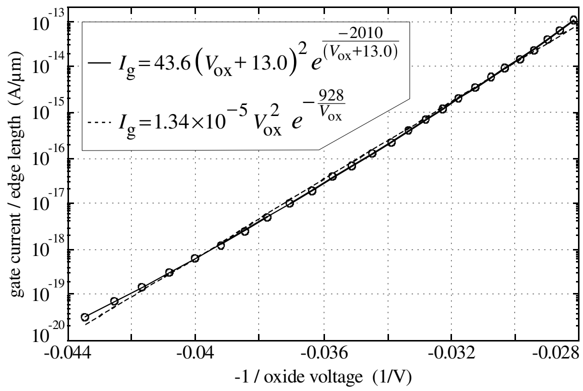

Fig. 3. Tunneling gate current versus oxide voltage, normalized to the tunneling-junction edge length in lineal microns. V_{ox} is defined as the potential difference between the tunneling junction and the floating gate. The modified Fowler–Nordheim fit (solid line) employs a built-in voltage to account for oxide traps; the conventional Fowler–Nordheim fit (dashed line) is shown for comparison.

coupling coefficient, and Ψ_o is derived from the MOS process parameters.

For both devices the injection efficiency is independent, to first-order, of the floating-gate–to-channel voltage, as long as $V_{fg} > V_d$ (where V_{fg} and V_d are the floating gate and drain voltages, respectively). In the *p*FET synapse, this condition is always satisfied. In the *n*FET synapse, this condition is not necessarily satisfied; the data of Fig. 4 show what happens when we sweep the *n*FET drain from voltages much less than V_{fg}, to voltages much greater than V_{fg}. As V_d approaches V_{fg}, the oxide voltage becomes small, and the gate current drops.

We fit the injection data of Fig. 4 empirically; we are currently analyzing the relevant electron-transport physics to derive equivalent analytic results. For the *n*FET synapse, we chose not to fit the region where $V_d > V_{fg}$ because, at such high drain voltages, the gate currents are too large for use in a practical learning system. For both synapses, then,

$$I_g = \eta I_s e^{-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{6}$$

where V_{dc} is the drain-to-channel potential and η, V_β, and V_η are measurable device parameters.

3.5. The Gate-Current Equation

Because the tunneling and injection gate currents are in opposite directions, we obtain the final gate-current equation, for both synapses, by subtracting Eqn. (6) from Eqn. (3):

$$I_g = \xi (V_{ox} + V_{bi})^2 e^{-\frac{V_o}{V_{ox}+V_{bi}}} - \eta I_s e^{-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{7}$$

The principal difference between the *n*FET and *p*FET synapses is the sign of the learning. In the *n*FET synapse, tunneling increases the channel current, whereas injection decreases it; in the *p*FET synapse, tunneling decreases the channel current, whereas injection increases it.

3.6. Impact Ionization

We equate a synapse's weight value with its source current. However, because for both synapses the activation energy for impact-ionization is less than the barrier energy for injection, a channel-to-drain E-field that generates injection electrons must also liberate additional electron–hole pairs [21]. For both synapses, the drain current can therefore exceed the source current. If we choose drain current, rather than source current, as the synapse output, we can rewrite the gate-current equation in terms of drain current using a (modified) lucky-electron [22] formulation:

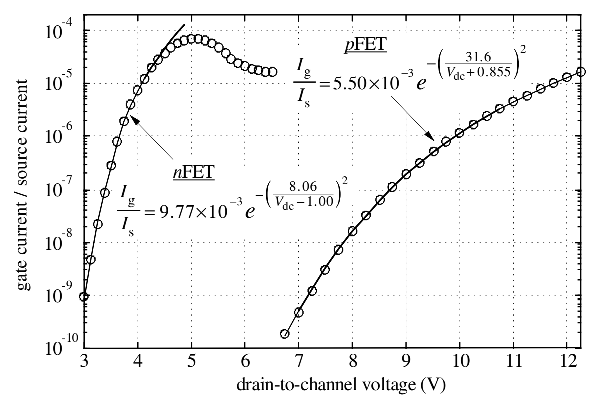

Fig. 4. Injection efficiency versus drain-to-channel voltage, for both the *n*FET and *p*FET synapses. The gate-to-channel voltages were held fixed during the experiments. For the *n*FET, $V_{gc} = 5.66$ V; for the *p*FET, $V_{gc} = 1.95$ V. In the *n*FET synapse, when the drain voltage exceeds the floating-gate voltage, the oxide E-field tends to return the injected electrons to the silicon surface, rather than transporting them to the floating gate. As a result, for drain-to-channel voltages near $V_{gc} = 5.66$ V, the *n*FET data deviate from the fit.

$$I_d = I_s\left(1+\varepsilon e^{-\sqrt{\frac{V_m}{V_{ds}-V_\varepsilon}}}\right) \quad (8)$$

where I_d is the drain current and ε, V_m, and V_ε are measurable device parameters. In Fig. 5 we plot impact ionization data for both synapses.

4. Synaptic Arrays

A synaptic array, with a synapse transistor at each node, can form the basis of a silicon learning system. We fabricated simplified 2×2 arrays to investigate synapse isolation during tunneling and injection, and to measure the synapse learning rates. Because a 2×2 array uses the same row–column addressing employed by larger arrays, it allows us to characterize completely the synapse isolation and learning.

4.1. The nFET Array

The *n*FET array is shown in Fig. 6. We chose, from among the many possible ways of using the array, to select source current as the synapse output, and to turn off the synapses while tunneling. We applied the voltages shown in Table 1 to read, tunnel, or inject synapse {1,1} selectively, while ideally leaving the other synapses unchanged.

Because the synapse tunneling and drain terminals are connected within a row, but not within a column, the crosstalk between column synapses is negligible. We reduce the crosstalk between row synapses by using oversize, 1 pF gate capacitors that provide 80% voltage coupling from the control gate to the floating gate. Because the tunneling gate current increases exponentially with the oxide voltage V_{ox}, (V_{ox}, in turn, decreases linearly with the floating-gate voltage), and because the hot-electron gate current increases linearly with the channel current I_s, (I_s, in turn, increases exponentially with the floating-gate voltage), the isolation between row synapses increases exponentially with their floating-gate voltage differential. By using 5V control-gate inputs, we achieve about a 4V differential between the floating gates of the selected and deselected synapses; the resulting crosstalk between row synapses is $<0.01\%$ for all operations.

To obtain the data in Fig. 7, we initially set all synapses to $I_s = 100$ pA. We tunneled the {1,1} synapse up to 100 nA, and then injected it back down to 100 pA, while measuring the source currents of the other three synapses. As expected, the row 2 synapses were unaffected by either the tunneling or the injection. Coupling to the {1,2} synapse also was small.

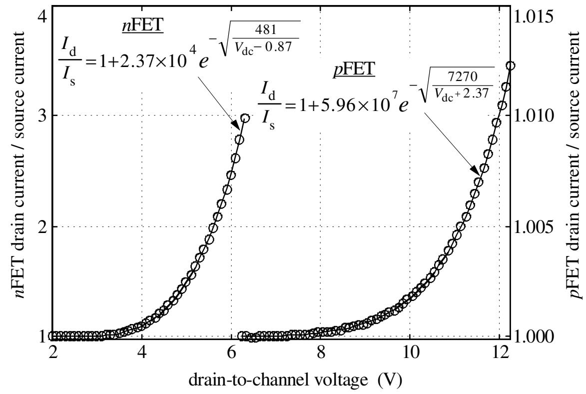

Fig. 5. Impact ionization versus drain-to-channel potential, for both the *n*FET and *p*FET synapses. Impact ionization in the *n*FET is markedly more efficient than in the *p*FET, for two reasons. First, as a consequence of its bulk *p*-type substrate implant, the *n*FET synapse experiences a higher drain-to-channel electric field than does the *p*FET, thereby increasing the ionization likelihood. Second, the impact-ionization process is naturally more efficient for electrons (the *n*FET charge carriers) than it is for holes (the *p*FET charge carriers).

To obtain the data in Fig. 8, we first set all four synapses to $I_s = 100$ nA. We injected the {1,1} synapse down to 100 pA, and then tunneled it back up to 100 nA. As in the experiment of Fig. 7, crosstalk to the other synapses was negligible. Our large gate capacitors minimize crosstalk, at the expense of synapse size and learning rate. We intend to fabricate future synapses with smaller gate capacitors.

4.2. The pFET Array

The *p*FET array is shown in Fig. 9. We ground the *p*-type substrate, apply +12 V to the *n*-type well, and reference all terminal voltages to the well potential.

We again chose source current as the synapse output, but here, unlike in the *n*FET array, we leave the *p*FET synapses turned on while tunneling. We applied the voltages shown in Table 2 to read, tunnel, or inject synapse {1,1} selectively, while ideally leaving the other synapses unchanged.

To obtain the data in Fig. 10, we initially set all synapses to $I_s = 100$ pA. We injected the {1,1} synapse

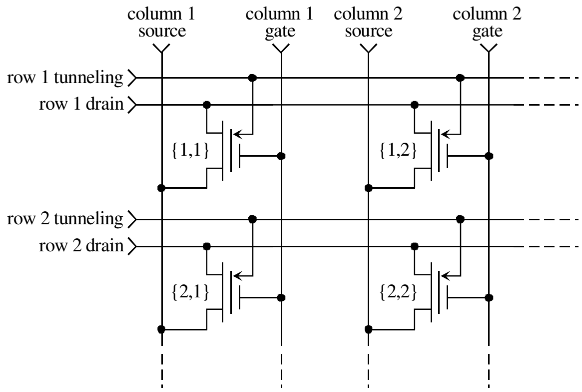

Fig. 6. A 2×2 array of *n*FET synapses. Because the row synapses share common tunneling and drain wires, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

up to 100 nA, and then tunneled it back down to 100 pA. To obtain the data in Fig. 11, we performed the opposite experiment. As was the case in the *n*FET array, in the *p*FET array the crosstalk between column synapses was negligible, and the crosstalk between row synapses was small.

When we inject the {1,1} synapse we apply −4 V, rather than 0 V, to the {1,2} synapse control gate. We do so because a *p*FET synapse can experience hot-electron injection by a mechanism different from that described in Section 3. If the floating-gate voltage exceeds the well potential, and the drain-to-channel voltage is large, electrons can inject onto the floating gate by means of a non-destructive avalanche-breakdown phenomenon [12] at the MOS surface. This alternate injection mechanism will be the subject of a future paper.

5. The Synapse Learning Rule

We repeated the experiments of Figs. 7 and 10, for several tunneling and injection voltages; in Figs. 12–15 we plot, for the *n*FET and *p*FET synapses, the temporal derivative of the source current as a function of the source current. If we equate a weight w with the source current I_s, these data show the synapses' weight-update rates. Starting from the gate-current equation, Eqn. (7), we now derive a learning rule that fits these data.

Table 1. The terminal voltages we applied to the array of Fig. 6, in order to obtain the data of Figs. 7 and 8.

	col 1 gate	col 1 source	col 2 gate	col 2 source	row 1 drain	row 1 tun	row 2 drain	row 2 tun
read	+5	0	0	0	+1	0	0	0
tunnel	0	0	+5	0	0	+31	0	0
inject	+5	0	0	0	3.15	0	0	0

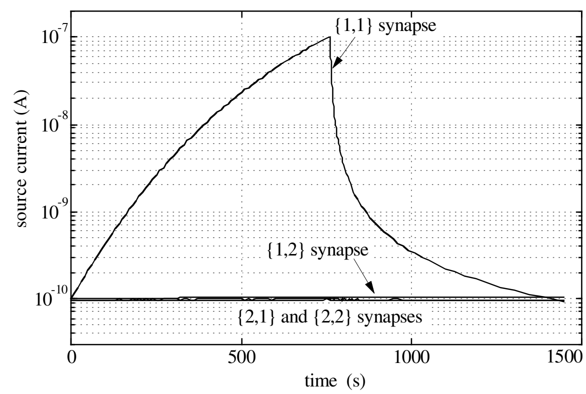

Fig. 7. Isolation in a 2×2 array of *n*FET synapses. Source current is the synapse output. The {1,1} synapse first is tunneled up to 100 nA, then is injected back down to 100 pA. The tunneling voltage, referenced to the substrate potential, is V_{tun}=31 V; the injection voltage is V_{ds}=3.15 V. Crosstalk to the {1,2} synapse, defined as the fractional change in the {1.2} synapse divided by the fractional change in the {1,1} synapse, is 0.006 % when tunneling, and is 0.002 % when injecting.

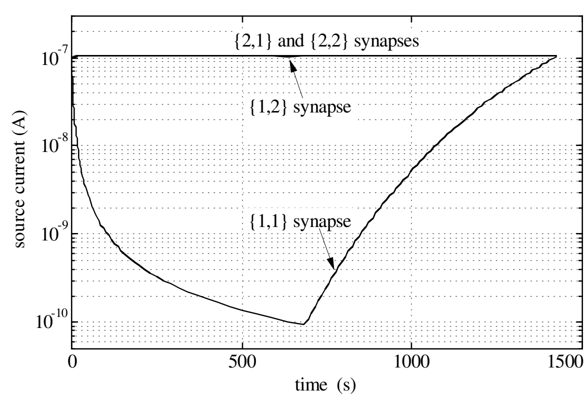

Fig. 8. The same experiment as in Fig. 7, but here the {1,1} synapse first is injected down to 100 pA, then is tunneled back up to 100 nA. Crosstalk to the {1,2} synapse is 0.001 % when injecting, and is 0.002 % when tunneling.

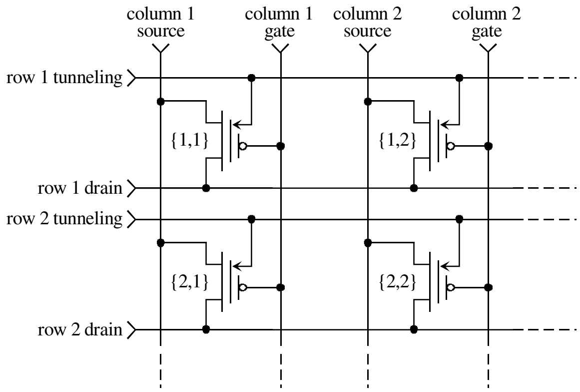

Fig. 9. A 2×2 array of *p*FET synapses. The well connections are not shown. As in the *n*FET array, because the row synapses share common tunneling and drain wires, tunneling or injection at one row synapse can cause undesired tunneling or injection at other row synapses.

5.1. Tunneling

We begin by taking the temporal derivative of Eqn. (1):

$$\frac{\partial I_s}{\partial t}=\frac{\kappa}{Q_T}I_o e^{\frac{\kappa' V_{in}}{U_t}} e^{\frac{\kappa Q_{fg}}{Q_T}}\frac{\partial Q_{fg}}{\partial t}=\frac{\kappa}{Q_T}I_s I_g \qquad (9)$$

In *Appendix A.1,* we substitute for the tunneling gate current using Eqn. (3), redefine I_s as the synapse weight w, and solve for the tunneling weight-update rule:

$$\frac{\partial w}{\partial t}\approx\frac{\kappa\xi'}{Q_T}w^{(1-\sigma)} \qquad (10)$$

where $$\sigma \equiv \frac{V_o U_t}{\kappa\left(V_{tun}+V_{bi}\right)^2} \qquad (11)$$

Equation (10) fits accurately the tunneling weight-update data for both synapses. In the *n*FET synapse, $0.12<\sigma<0.22$; in the *p*FET, $0.01<\sigma<0.05$.

5.2. Injection

We begin with Eqn. (9):

$$\frac{\partial I_s}{\partial t}=\frac{\kappa}{Q_T}I_s I_g \qquad (12)$$

In *Appendix A.2,* we substitute for the injection gate current using Eqn. (6), replace I_s with w, and

Table 2. The terminal voltages we applied to the array of Fig. 9, in order to obtain the data of Figs. 10 and 11.

	col 1 gate	col 1 source	col 2 gate	col 2 source	row 1 drain	row 1 tun	row 2 drain	row 2 tun
read	-5	0	0	0	-5	0	0	0
tunnel	-5	0	0	0	-5	+28	0	0
inject	-5	0	-4	0	-9.3	0	0	0

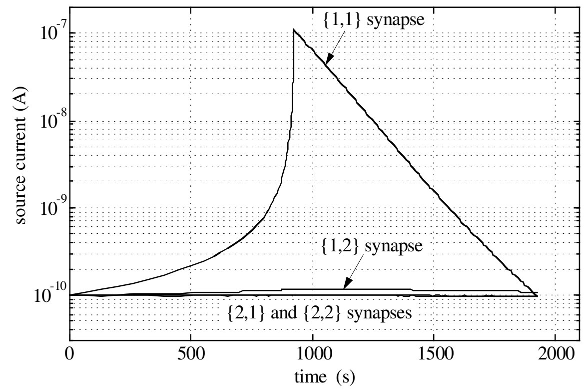

Fig. 10. Isolation in a 2×2 array of *p*FET synapses. Source current is the synapse output. The {1,1} synapse first is injected up to 100 nA, then is tunneled back down to 100 pA. The injection voltage is $V_{ds}=-9.3$ V; the tunneling voltage, referenced to the well potential, is $V_{tun}=28$ V. Crosstalk to the {1,2} synapse, defined as the fractional change in the {1.2} synapse divided by the fractional change in the {1,1} synapse, is 0.016 % when injecting, and is 0.007 % when tunneling.

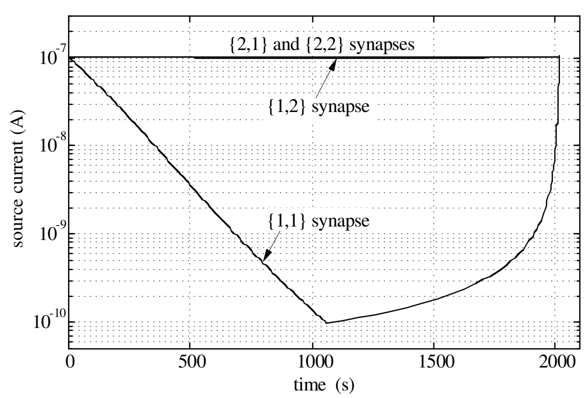

Fig. 11. The same experiment as in Fig. 10, but here the {1,1} synapse first is tunneled down to 100 pA, then is injected back up to 100 nA. Crosstalk to the {1,2} synapse is 0.005 % when injecting, and is 0.004 % when tunneling.

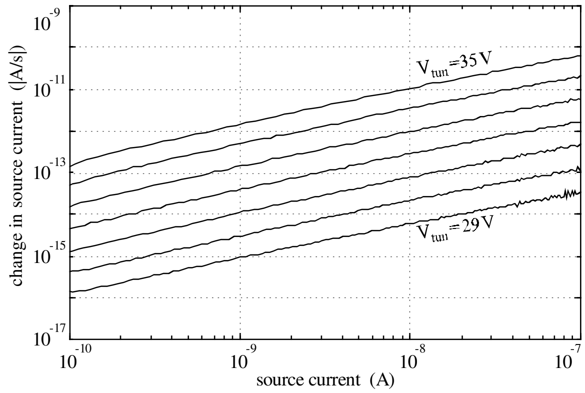

Fig. 12. Tunneling delta-weight versus weight for the *n*FET synapse, with source current chosen as the synapse weight. The {1,1} synapse was tunneled up as in Fig. 7, with the source at ground and the ground-referenced tunneling voltage stepped from 29 V to 35 V in 1 V increments. We here plot the magnitude of the temporal derivative of the weight value as a function of the weight value. The mean tunneling slope is +0.83.

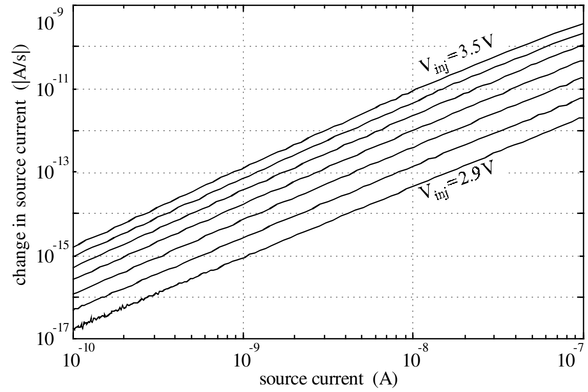

Fig. 13. Injection delta-weight versus weight for the *n*FET synapse, with source current chosen as the synapse weight. The {1,1} synapse was injected down as in Fig. 7, with the source at ground and the ground-referenced drain voltage stepped from 2.9 V to 3.5 V in 0.1 V increments. We plot the magnitude of the temporal derivative of the weight value as a function of the weight value. The mean injection slope is −1.76; the minus sign has been added because the synapse weight is injecting down.

solve for the injection weight-update rule:

$$\frac{\partial w}{\partial t} = -\frac{\eta' \kappa}{\mathrm{Q_T}} w^{(2-\alpha-\beta)} \tag{13}$$

where

$$\alpha \equiv 2\frac{\mathrm{U_t V_\beta^2}}{\left(\mathrm{V_{ds}} + \mathrm{V_\eta}\right)^3} \tag{14}$$

and

$$\beta \equiv 3\frac{\mathrm{U_t^2 V_\beta^2}}{\left(\mathrm{V_{ds}} + \mathrm{V_\eta}\right)^4} \ln\left(\frac{I_\mathrm{s}}{\mathrm{I_o'}}\right) \tag{15}$$

Because $\ln(I_s)$ changes slowly, we approximate β to be constant. Equation (13) fits accurately the injection weight-update data for both synapses. In the *n*FET, $0.14 < \alpha + \beta < 0.28$; in the *p*FET, $0.08 < \alpha + \beta < 0.14$.

5.3 The Learning Rule

We obtain the synapse learning rule by adding Eqns. (10) and (13), with a leading (±) added because the sign of the learning is different in the *n*FET and *p*FET synapses:

$$\frac{\partial w}{\partial t} = \pm\frac{\kappa}{\mathrm{Q_T}}\left[\xi' w^{(1-\sigma)} - \eta' w^{(2-\alpha-\beta)}\right] \tag{16}$$

Learning in the *n*FET synapse is described by selecting the (+) in Eqn. (16); in the *p*FET synapse, the (−) is chosen.

5.4. Learning-Rate Degradation

SiO_2 trapping is a well-known issue in floating-gate transistor reliability [23]. In digital EEPROM memories, it ultimately limits the transistor life. In the synapses, trapping decreases the learning rate. However, unlike the transistors in a digital memory, the synapses in a typical learning system will transport only a small quantity of total oxide charge over the system lifetime. We tunneled and injected 1 nC of gate charge in both synapses, and measured a ~20% drop in both the tunneling and injection weight-update learning rates. Because 1 nC of gate charge represents an enormous change in synapse gate voltage, we believe that oxide trapping can be ignored safely.

6. Conclusion

We have described complementary single-transistor silicon synapses with nonvolatile analog memory, simultaneous memory reading and writing, and bidirectional memory updates that are a function of both

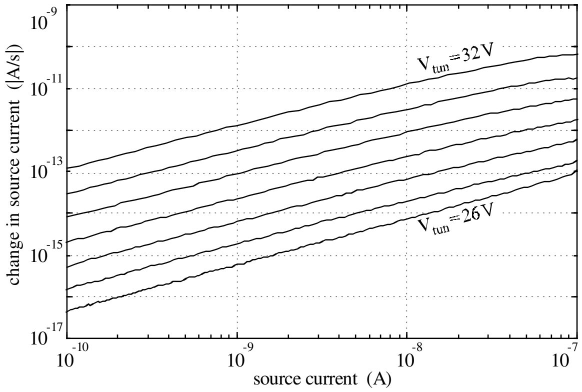

Fig. 14. Tunneling delta-weight versus weight for the *p*FET synapse, with source current chosen as the synapse weight. The {1,1} synapse was tunneled down as in Fig. 10, with the source and well at +12 V and the tunneling voltage, referenced to the well potential, stepped from 26 V to 32 V in 1 V increments. We plot the magnitude of the temporal derivative of the weight value as a function of the weight value. The mean tunneling slope is −0.99; the minus sign has been added because the synapse weight is tunneling down.

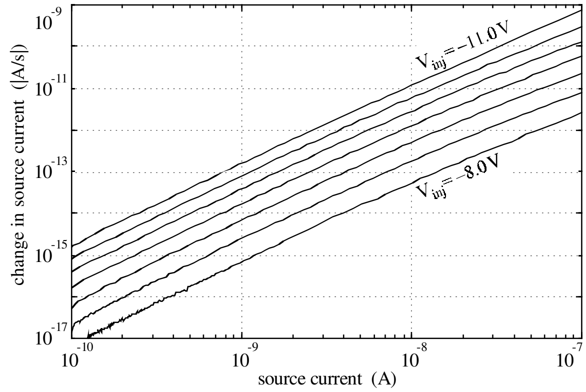

Fig. 15. Injection delta-weight versus weight for the *p*FET synapse, with source current chosen as the synapse weight. The {1,1} synapse was injected up as in Fig. 10, with the source and well at +12 V and the drain voltage, referenced to the source potential, stepped from −8.0 V to −11.0 V in −0.5 V increments. We plot the magnitude of the temporal derivative of the weight value as a function of the weight value. The mean injection slope is +1.89.

the applied terminal voltages and the present output. We have demonstrated that a learning system can be realized as a two-dimensional synaptic array, and have shown that we can address individual array nodes with good selectivity. We have derived a synapse learning rule, and believe that we can build autonomous learning systems, combining single-transistor analog computation with memory updates computed both locally and in parallel, with these devices. Finally, we anticipate that our single-transistor synapses will allow the development of dense, low-power, silicon learning systems.

Appendix A

A.1. The Tunneling Weight-Update Rule

We begin with Eqn. (9):

$$\frac{\partial I_s}{\partial t} = \frac{\kappa}{Q_T} I_o e^{\frac{\kappa' V_{in}}{U_t}} e^{\frac{\kappa Q_{fg}}{Q_T}} \frac{\partial Q_{fg}}{\partial t} = \frac{\kappa}{Q_T} I_s I_g \qquad (9)$$

We substitute Eqn. (3) for the gate current I_g:

$$\frac{\partial I_s}{\partial t} = \frac{\kappa\xi}{Q_T} I_s \left(V_{ox} + V_{bi}\right)^2 e^{-\frac{V_o}{V_{ox}+V_{bi}}} \qquad (17)$$

We substitute $V_{ox} = V_{tun} - V_{fg}$ (where V_{tun} is the tunneling-node voltage and V_{fg} is the floating–gate voltage), approximate $V_{tun} + V_{bi} >> V_{fg}$, expand the tunneling exponential by $(1-x)^{-1} \approx 1 + x$, and solve for the tunneling weight-update rule:

$$\frac{\partial I_s}{\partial t} \approx \frac{\kappa\xi}{Q_T} e^{\frac{-V_o}{V_{tun}+V_{bi}}} \left(V_{tun} + V_{bi} - V_{fg}\right)^2 I_o^{\sigma} I_s^{(1-\sigma)} \qquad (18)$$

where
$$\sigma \equiv \frac{V_o U_t}{\kappa\left(V_{tun} + V_{bi}\right)^2} \qquad (19)$$

Because, for subthreshold source currents, the floating-gate voltage changes slowly, we approximate $(V_{tun} + V_{bi} - V_{fg})^2$ to be constant. We define:

$$\xi' \equiv \xi\, e^{\frac{-V_o}{V_{tun}+V_{bi}}} \left(V_{tun} + V_{bi} - V_{fg}\right)^2 I_o^{\sigma} \qquad (20)$$

Finally, we substitute ξ' into Eqn. (18), redefining I_s as a weight w:

$$\frac{\partial w}{\partial t} \approx \frac{\kappa\xi'}{Q_T} w^{(1-\sigma)} \qquad (21)$$

A.2. The Injection Weight-Update Rule

We begin by rewriting the drain-to-channel potential, V_{dc}, in terms of V_{ds} and I_s. In a subthreshold floating-gate MOSFET, the source current is related to the floating-gate and source voltages [14] by:

$$I_s = I_o e^{\frac{\kappa V_{fg} - V_s}{U_t}} \tag{22}$$

Using Eqns. (5) and (22), we solve for the surface potential Ψ in terms of I_s and V_s:

$$\Psi = V_s + U_t \ln\left(\frac{I_s}{I_o'}\right) \tag{23}$$

where

$$I_o' \equiv I_o e^{-\frac{\Psi_o}{U_t}} \tag{24}$$

We now solve for V_{dc}:

$$V_{dc} = V_d - \Psi = V_{ds} - U_t \ln\left(\frac{I_s}{I_o'}\right) \tag{25}$$

The injection gate current I_g is given by Eqn. (6). We add a (–) sign to the gate current, because hot-electron injection decreases the floating-gate charge, and substitute for V_{dc} using Eqn. (25):

$$I_g = -\eta I_s e^{-\left(\frac{V_\beta}{V_{ds} + V_\eta - U_t \ln\left(\frac{I_s}{I_o'}\right)}\right)^2} \tag{26}$$

We expand the exponent by $(1-x)^{-2} \approx 1 + 2x + 3x^2$, and solve for I_g:

$$I_g \approx -\eta' I_s^{(1-\alpha-\beta)} \tag{27}$$

where:

$$\alpha \equiv 2\frac{U_t V_\beta^2}{\left(V_{ds} + V_\eta\right)^3} \tag{28}$$

$$\beta \equiv 3\frac{U_t^2 V_\beta^2}{\left(V_{ds} + V_\eta\right)^4} \ln\left(\frac{I_s}{I_o'}\right) \tag{29}$$

$$\eta' \equiv \eta I_o'^{(\alpha+\beta)} e^{-\frac{V_\beta^2}{(V_{ds}+V_\eta)^2}} \tag{30}$$

Because $\ln(I_s)$ changes slowly, we approximate β to be constant. Finally, we substitute Eqn. (27) into Eqn. (9), replacing I_s with w.

$$\frac{\partial w}{\partial t} \approx -\frac{\eta' \kappa}{Q_T} w^{(2-\alpha-\beta)} \tag{31}$$

Acknowledgments

This work was supported by the Office of Naval Research, by the Advanced Research Projects Agency, by the Beckman Hearing Institute, by the Center for Neuromorphic Systems Engineering as a part of the National Science Foundation Engineering Research Center Program, and by the California Trade and Commerce Agency, Office of Strategic Technology.

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, pp. 262-267, 1991.
2. P. Hollis and J. Paulos, "A neural network learning algorithm tailored for VLSI implementation," *IEEE Tran. Neural Networks*, vol. 5, no. 5, pp. 784-791, 1994.
3. F. Masuoka, R. Shirota, and K. Sakui, "Reviews and prospects of non-volatile semiconductor memories," *IEICE Trans.,* vol. E 74, no. 4, pp. 868–874, 1991.
4. J. Lazzaro, *et al.*, "Systems technologies for silicon auditory models," *IEEE Micro*, vol. 14, no. 3, pp. 7-15, 1994.
5. T. Allen, *et al.*, "Writeable analog reference voltage storage device," *U.S. Patent No. 5,166,562*, 1991.
6. P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses," *Advances in Neural Information Processing Systems 7*, MIT Press, pp. 817–824, 1995.
7. P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses with long term storage," *IEEE Intl. Symp. on Circuits and Systems*, vol. 3, pp. 1660–1663, 1995.
8. C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A single transistor silicon MOS device for long term learning," U.S. Patent Office serial no. 08/399966, Mar. 7, 1995.
9. C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A high-resolution non-volatile analog memory cell," *Proc. IEEE Intl. Symp. on Circuits and Systems*, vol. 3, pp. 2233–2236, 1995.

10. P. Hasler, B. A. Minch, C. Diorio, and C. Mead, "An autozeroing amplifier using *p*FET hot-electron injection," *Proc. IEEE Intl. Symp. on Circuits and Systems*, vol. 3, pp. 325 – 328, 1996.
11. C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A single-transistor silicon synapse," *IEEE Trans. Electron Devices,* vol. 43, no. 11, pp. 1972-1980, 1996.
12. E. Takeda, C. Yang, and A. Miura-Hamada, *Hot-Carrier Effects in MOS Devices,* San Diego, CA: Academic Press, Inc., 1995.
13. M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling into thermally grown SiO_2," *J. of Appl. Phys.*, vol. 40, no. 6, pp. 278–283, 1969.
14. C. Mead, *Analog VLSI and Neural Systems,* Reading, MA: Addison-Wesley, Inc., 1989.
15. A. S. Grove, *Physics and Technology of Semiconductor Devices*. New York: John Wiley & Sons, Inc., 1967.
16. S. M. Sze, *Physics of Semiconductor Devices,* New York: John Wiley & Sons, Inc., 1981.
17. C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *J. of VLSI Signal Processing,* vol. 8, pp. 9–25, 1994.
18. J. J. Sanchez and T. A. DeMassa, "Review of carrier injection in the silicon/silicon-dioxide system," *IEE Proceedings-G*, vol. 138, no. 3, pp. 377–389, 1991.
19. C. C. Enz, F. Krummenacher, and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and Signal Processing,* vol. 8, pp. 83–114, 1995.
20. A. G. Andreou and K. A. Boahen, "Neural information processing II," in M. Ismail and T. Fiez, eds., *Analog VLSI Signal and Information Processing,* New York: McGraw-Hill, Inc., pp. 358–413, 1994.
21. W. Shockley, "Problems related to *p-n* junctions in silicon," *Solid-State Electronics*, vol. 2, no. 1, pp. 35–67, Pergamon Press, 1961.
22. S. Tam, P. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Trans. Electron Devices*, vol. ED-31, no. 9, pp. 1116–1125, 1984.
23. S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of flash memory cells," *Proc. of the IEEE,* vol. 81, no. 5, pp. 776–787, 1993.

Chris Diorio received his B.A. in physics from Occidental College in 1983, and an M.S. in electrical engineering from the California Institute of Technology in 1984. Since September 1992, he has been a doctoral candidate in electrical engineering at the California Institute of Technology. His interests include analog integrated circuit design, ultra-high-speed digital circuit design, and semiconductor device physics. His current research involves using floating-gate MOS transistors to build adaptive systems in silicon. He is employed as a Staff Engineer at TRW, Inc., in Redondo Beach, CA, and has worked as a Senior Staff Scientist at American Systems Corporation in Chantilly, VA, and as a Technical Consultant at The Analytic Sciences Corporation in Reston, VA. Mr. Diorio is a member of Sigma Pi Sigma, and is a student member of the IEEE.

Paul Hasler received his B.S.E. and M.S. degrees in electrical engineering from Arizona State University in August 1991. Since September 1992, he has been a doctoral candidate in computation and neural systems at the California Institute of Technology. His research interests include using floating-gate MOS transistors to build adaptive systems in silicon, investigating the solid-state physics of floating-gate devices, and modeling high-field carrier transport in Si and SiO_2. Mr. Hasler is a member of Tau Beta Pi, a member of Eta Kappa Nu, and a student member of the IEEE.

Bradley A. Minch received his B.S. in electrical engineering, with distinction, from Cornell University in 1991. Since September 1991, he has been a doctoral candidate in computation and neural systems at the California Institute of Technology. His research interests include current-mode circuits and signal processing, the use of floating-gate MOS transistors to build adaptive systems in silicon, and silicon models of dendritic computation. Mr. Minch is a member of Tau Beta Pi, a member of Eta Kappa Nu, a member of Phi Kappa Phi, and a student member of the IEEE.

Carver A. Mead, Gordon and Betty Moore Professor of Engineering and Applied Science, has taught at the California Institute of Technology for more than 30 years. He has contributed in the fields of solid-state electronics and the management of complexity in the design of very large-scale integrated circuits, and has been active in the development of innovative design methodologies for VLSI. He wrote, with Lynn Conway, the standard text for VLSI design, *Introduction to VLSI Systems*. His more recent work is concerned with modeling neuronal structures, such as the retina and the cochlea, using analog VLSI systems. His book on this topic, *Analog VLSI and Neural Systems*, was published in 1989 by Addison-Wesley. Professor Mead is a member of the National Academy of Sciences, the National Academy of Engineering, the American Academy of Arts and Sciences, a foreign member of the Royal Swedish Academy of Engineering Sciences, a Fellow of the American Physical Society, a Fellow of the IEEE, and a Life Fellow of the Franklin Institute. He is also the recipient of numerous awards, including the Centennial Medal of the IEEE.