A Single-Transistor Silicon Synapse

Chris Diorio, *Member, IEEE,* Paul Hasler, *Student Member, IEEE,* Bradley A. Minch, *Student Member, IEEE,* and Carver A. Mead, *Fellow, IEEE*

***Abstract*—We have developed a new floating-gate silicon MOS transistor 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 synapse can implement a learning function. We have derived a memory-update rule from the physics of the tunneling and injection processes, and have investigated synapse learning in a prototype array. Unlike conventional EEPROM devices, the synapse allows simultaneous memory reading and writing. Synapse transistor arrays can therefore compute both the array output, and local memory updates, in parallel. The synapse is small, and typically is operated at subthreshold current levels; it will permit the development of dense, low-power silicon learning systems.**

I. Introduction

We have fabricated single-transistor learning devices that integrate nonvolatile analog memory storage with bidirectional memory modification; we call these devices *silicon synapses*. Like a neural synapse [1], our silicon synapses compute the product of the stored analog memory value and the applied input. Also like a neural synapse, they can learn from the input signal without interrupting the ongoing computation. Although we do not believe that a single transistor can model completely the complex behavior of a neural synapse, our transistors do implement a local learning function. With them, we intend to construct autonomous silicon learning systems.

We have described previously a four-terminal nFET synapse [2]–[4]; it is a high-threshold floating-gate MOSFET with an associated tunneling junction. It shows promise for a range of applications, including high-resolution analog memories [5] and analog learning arrays [3]. The synapse described here integrates the tunneling function within the transistor drain, yielding a three-terminal device. Like the four-terminal device, the three-terminal device possesses five attributes that we believe are essential in a silicon synapse. First, when the synapse is not learning, the analog memory is nonvolatile; when the synapse is learning, memory updates can be bidirectional. Second, the synapse output is the product of the input signal and the stored memory value. Third, memory reading and writing can occur simultaneously. Fourth, the memory updates vary with both the input signal and the stored memory value. Fifth, the synapse is compact, and operates off a single-polarity supply with low power consumption.

Our synapse differs from conventional EEPROM transistors both in its function and in its potential applications. Not only does it provide nonvolatile analog memory storage, and compute locally the product of its stored memory value and the applied input, but it also permits simultaneous memory reading and writing, and can even compute locally its own memory updates. We anticipate building synapse-based learning systems in which both the system outputs, and the memory updates, are computed both locally and in parallel. By contrast, because conventional EEPROM transistors are optimized for digital programming and binary-valued data storage [6], they typically possess few if any of these features, and therefore have seen only limited use in silicon learning systems.

II. The Silicon Synapse

The silicon synapse is an n-type MOSFET with a poly1 floating gate, a poly2 control gate, a moderately doped channel, and a lightly doped drain (LDD). It uses channel hot-electron injection (CHEI) to add electrons to its floating gate, and Fowler-Nordheim (FN) tunneling [7] to remove them. It has been fabricated in the 2 μm n-well Orbit BiCMOS process available from MOSIS. Top and side views of the device are shown in Figs. 1 and 2, respectively. Its principal features are

- Electrons tunnel from floating gate to drain through 350 Å gate oxide. High drain voltages provide the oxide E-field required for tunneling. The lightly doped ($\sim 5\times 10^{15}$ cm^{-3}) well–drain prevents pn-junction breakdown.
- Electron tunneling is enhanced where the poly1 floating gate overlaps the heavily doped ($\sim 1 \times 10^{19}$ cm^{-3}) well–drain 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 bipolar-transistor base implant ($\sim 1\times 10^{17}$ cm^{-3}) to the MOS channel region. As a result, the channel-to-drain depletion layer approximates a one-sided step junction, increasing the injection likelihood. The channel implant also raises the transistor's threshold voltage V_t, favoring the collection of the injected electrons by the floating gate.
- The channel-to-drain space-charge layer appears primarily on the drain side of the junction. We extend the MOS gate oxide 2 μm beyond the channel-drain edge,

Manuscript received December 5, 1995; revised April 1, 1996. The review of this paper was arranged by Editor C.-Y. Lu. 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.

The authors are with the Physics of Computation Laboratory, California Institute of Technology, Pasadena, CA 91125 USA.

Publisher Item Identifier S 0018-9383(96)07734-9.

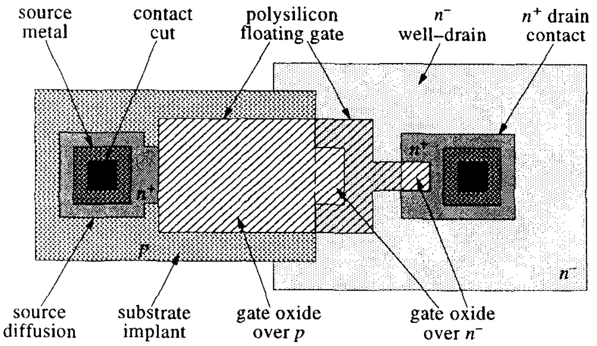


Fig. 1. Synapse transistor, top view. The poly2 control gate is not shown. In the Orbit 2 μm process, the channel width is 8 μm, and the channel length is 11 μm.

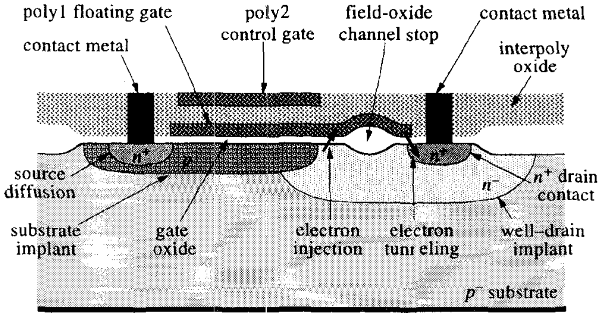


Fig. 2. Synapse transistor, side view, showing the tunneling and injection locations. As a consequence of the p-type substrate implant, the floating-gate to channel-surface coupling coefficient, κ, is 0.2.

over the space-charge layer. Because the injected channel electrons encounter gate oxide, rather than a field-oxide channel stop, CHEI is greatly facilitated by this gate-oxide extension.

- 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.

The stored memory value is represented by the floating-gate charge. Either channel current or channel conductance can be selected as the synapse output. Inputs are typically 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. CHEI adds electrons to the floating gate, reducing the charge and raising the threshold; tunneling removes electrons, increasing the charge and lowering the threshold.

We typically operate the synapse in its subthreshold regime [8], to limit the power consumption, and typically select either drain current or source current to be the synapse output. When operated in this fashion, the synapse output is the product of a stored memory value and the applied input as follows:

$$I_s = I_0 e^{\frac{\kappa Q_{\mathrm{fg}}}{C_T U_t}} e^{\frac{\kappa C_{\mathrm{in}} V_{\mathrm{in}}}{C_T U_t}} = I_0 e^{\frac{\kappa Q_{\mathrm{fg}}}{Q_T}} e^{\frac{\kappa' V_{\mathrm{in}}}{U_t}} \tag{1}$$

$$= I_m e^{\frac{\kappa' V_{\mathrm{in}}}{U_t}} \tag{2}$$

where I_s is the source current, I_0 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 $kT/q, C_{\mathrm{in}}$ is the input (poly1 to poly2) coupling capacitance, V_{in} is the signal voltage applied to the control gate, $Q_T \equiv C_T U_t$, and $\kappa' \equiv \kappa C_{\mathrm{in}}/C_T$.

The current I_m is a learned quantity; its value changes with synapse use. The synapse output is the product of I_m and the exponentiated gate input. Because the CHEI and tunneling 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 synapse exhibits a type of learning by which its future output depends on both the applied input and the present output.

III. The Channel Energy Profile

To be injected onto the floating gate, channel electrons must (1) acquire the 3.2 eV required to surmount the Si-SiO_2 work-function barrier, (2) scatter upward into the gate oxide, and (3) be transported across the oxide to the floating gate. CHEI in conventional MOSFET's is well known [9]. It occurs in short-channel devices with continuous channel currents, when a high gate voltage is combined with a 3.2 V drop across the short channel. It also occurs in switching transistors, when both the drain and gate voltages are transiently high. In neither case is the CHEI suitable for use in a learning system. The short-channel CHEI requires large channel currents, consuming too much power; the switching-induced CHEI is a poorly controlled transient phenomenon.

We impart 3.2 eV to the channel electrons by accelerating them in the synapse transistor's channel-to-drain E-field. However, merely generating a 3.2 eV electron population is not, by itself, sufficient for CHEI. As shown in Fig. 3, a conventional well–drain MOSFET can experience a channel-to-drain E-field exceeding 10 V/μm, thereby inducing a large 3.2 eV carrier population. Still, when operating in the subthreshold regime, this device experiences little or no CHEI. Under similar conditions, the synapse transistor's injection efficiency (gate current divided by source current) can exceed 1×10^{-8}. This efficiency improvement is a consequence of the synapse transistor's higher p-type substrate doping, for two reasons.

First, the synapse transistor's channel-to-drain depletion region is one-sided, with 95% of the space-charge layer appearing on the drain side of the junction. When $V_{\mathrm{dc}} = 30$ V, peak field occurs a mere 0.14 μm into this space-charge layer. At peak field, the conduction-band potential rises 3.2 V in 25λ (where $\lambda \sim 7$ nm is the electron mean-free-path length [10]). A hot-electron population is therefore available near the channel edge of the space-charge layer. By contrast, in the conventionally doped well–drain transistor, the channel-to-drain depletion region is symmetric rather than one-sided; peak field is not reached until 2 μm into the space-charge layer.

Second, the higher surface-acceptor concentration raises the synapse transistor's threshold voltage V_t from 0.8 V to 6.2 V. It is evident from Fig. 3 that electron transport within the SiO_2 depends on the direction of the oxide E-field. Where the gate voltage exceeds the surface potential, the oxide field

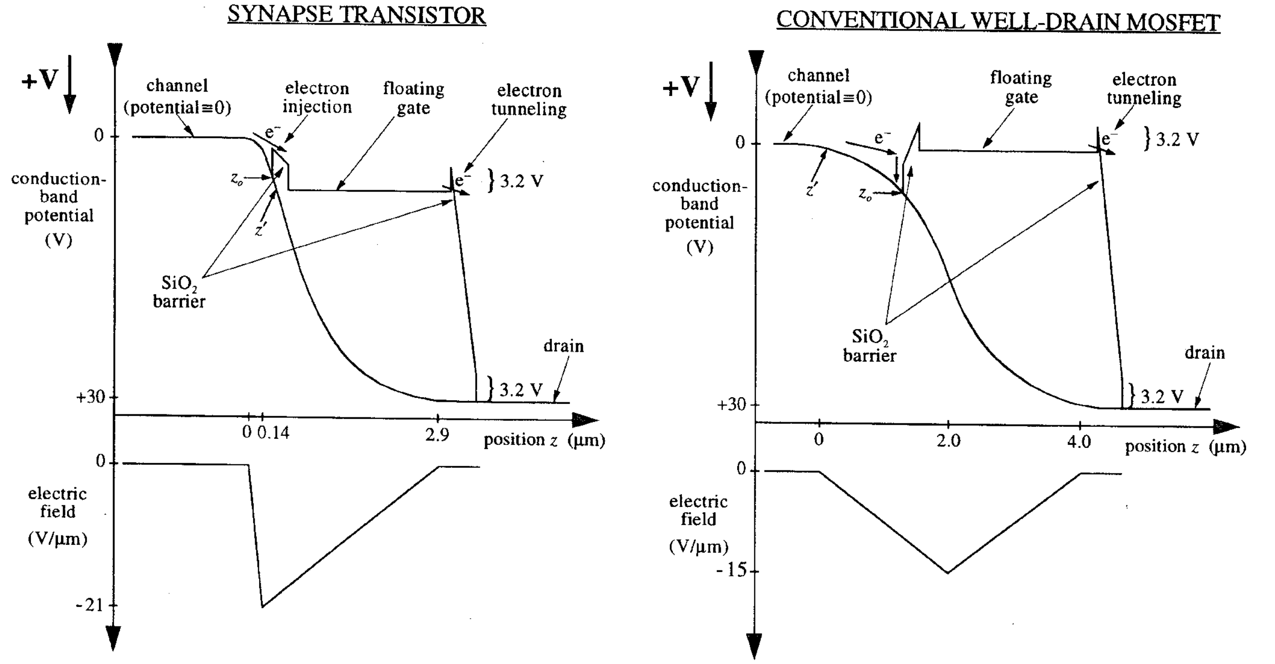


Fig. 3. Drain-to-channel space-charge conduction-band potential and electric field, calculated from implant impurity concentrations [11]. The transistors are identical except for their channel impurity doping, which in the synapse is 1×10^{17} cm^{-3} and in the conventional well–drain MOSFET is 5×10^{15} cm^{-3}. Both a step doping profile and subthreshold operation ($I_s < 100$ nA) are assumed. The synapse impact-ionization data of Fig. 8, and an observed 70 V drain-avalanche onset, are both consistent with the step-junction approximation. All voltages are referenced to the channel potential; all positions are measured from the channel edge of the channel-to-drain space-charge layer. Although the gate-oxide band diagrams actually project into the plane of the page, for convenience they have been rotated 90° and drawn in the channel direction. Because, for both devices, the conduction-band edge provides the reference potential for the oxide barrier's leading edge, the barrier shape varies with position z along the channel. For clarity, oxide barriers are drawn for only a single channel position, z_0. At $z = z'$, the oxide voltage is zero; for $z > z'$, the oxide field opposes the transport of injected electrons to the floating gate.

sweeps injected electrons across the SiO$_2$ to the floating gate. Where the surface potential exceeds the gate voltage, injected electrons tend to return to the silicon surface. When $V_{\mathrm{dc}} = 30$ V, the synapse's conduction-band potential is 3.2 V at $z = 0.22$ μm, whereas the surface potential does not exceed the gate voltage until $z = 0.37$ μm. The gate current arises primarily in the intervening region ($0.22 < z < 0.37$ μm). By contrast, in the conventional well–drain transistor with $V_{\mathrm{dc}} =$ 30 V, the conduction-band potential does not reach 3.2 V until 0.9 μm into the space-charge layer. Here the surface potential exceeds the gate voltage by 6.5 V, preventing a gate current.

IV. The Gate-Current Equation

We would like to use our synapse to build a silicon learning system. Because the learning behavior of any such system is determined in part by the CHEI and tunneling processes that alter the stored memory, we have investigated these processes over the anticipated synapse operating range. Based on a preliminary analysis, and on data taken from four-terminal synapses fabricated in a 1.2 μm process, we believe that the gate-current equation derived here, with modified fit constants, will describe generally the learning behavior of three-terminal well–drain synapses fabricated in more modern processes.

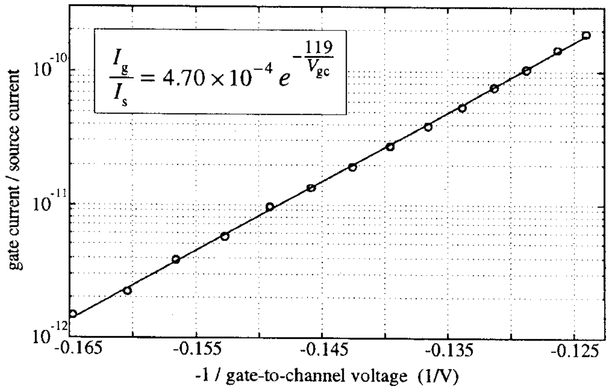


Fig. 4. Synapse-transistor CHEI efficiency versus gate-to-channel voltage, for $V_{\mathrm{dc}} = 20$ V and $I_s = 2\mu$A.

A. Hot-Electron Injection

To measure CHEI, we fabricated the device of Fig. 1 without a tunneling junction. We fit the measured injection data empirically; we are currently analyzing the relevant transport physics to derive equivalent analytic results. Because the CHEI 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 an MOS transistor [12], [13].

In Fig. 3, we define z' to be that location where the oxide E-field is zero. Because z' increases with gate voltage, the gate current also increases with gate voltage. Fig. 4 shows CHEI efficiency versus gate-to-channel potential. The data are fit by

$$I_g = \alpha I_s e^{-\frac{V_\alpha}{V_{gc}}} \tag{3}$$

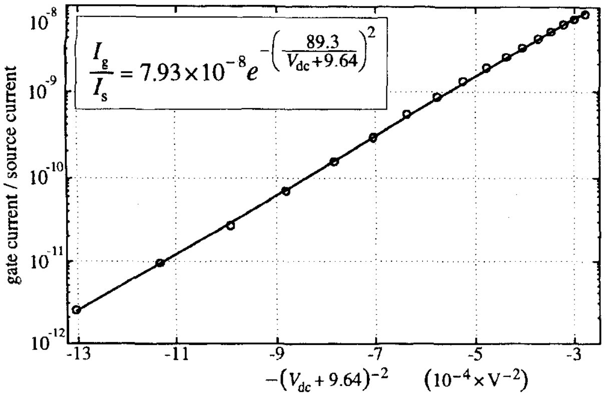


Fig. 5. Synapse-transistor CHEI efficiency versus drain-to-channel voltage, for $V_{gc} = 6.7$ V and $I_s = 2\mu$A.

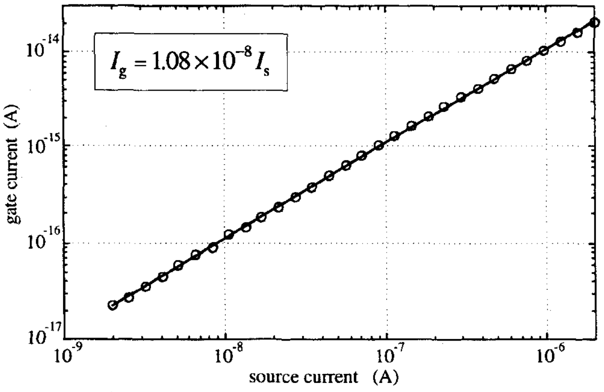


Fig. 6. Synapse-transistor gate current versus source current. The drain-to-bulk and gate-to-bulk voltages were held fixed at $V_{db} = 35$ V and $V_{gb} = 7$ V during the experiment.

where I_g is the gate current, I_s is the source current, V_{gc} is the gate-to-channel potential, and α and V_α are constants.

Because the channel-to-drain E-field increases with drain voltage, the gate current also increases with drain voltage. Fig. 5 shows the injection efficiency versus drain-to-channel potential. These data are fit by

$$I_g = \beta I_s e^{-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{4}$$

where V_{dc} is the drain-to-channel potential and β, V_β, and V_η are constants.

Fig. 6 shows the gate current to be directly proportional to the source current. Fig. 7 shows the CHEI efficiency versus both drain-to-channel and gate-to-channel potential. We fit these data by combining (3) and (4)

$$I_g = \eta I_s e^{-\frac{V_\alpha}{V_{gc}}-\left(\frac{V_\beta}{V_{dc}+V_\eta}\right)^2} \tag{5}$$

where $\eta = 3.63$ is a fit constant, and V_α, V_β, and V_η remain unchanged from (3) and (4).

We equate channel current with source current. Because the activation energy for impact ionization in silicon is less than 3.2 eV, a channel-to-drain E-field that generates 3.2 eV carriers must create additional electron-hole pairs [14] at the drain. We show synapse impact-ionization data in Fig. 8. The drain current is determined from the source current and drain-to-channel potential, [15] by

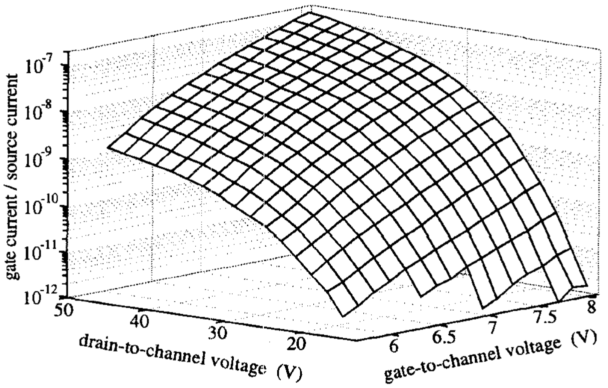


Fig. 7. Synapse-transistor injection efficiency versus drain-to-channel and gate-to-channel voltages. The RMS deviation between these data and (5) is 1.2 × 10^{-9}.

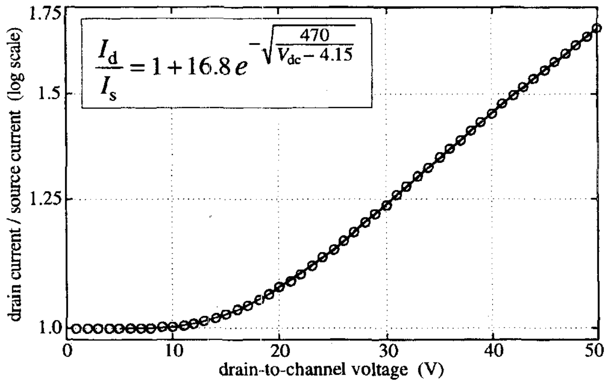


Fig. 8. Synapse-transistor impact ionization versus drain-to-channel voltage, for $V_{gc} = 6.15$ V and $I_s = 100$ nA. The fit function is independent of source current for 2 nA $< I_s <$ 2μA.

$$I_d = I_s\left(1 + \gamma e^{-\sqrt{\frac{V_m}{V_{dc}-V_\gamma}}}\right) \tag{6}$$

where I_d is the drain current and γ, V_m, and V_γ are constants.

B. Tunneling

The FN-tunneling process is illustrated in Fig. 3. The drain-to-gate potential reduces the effective oxide thickness, facilitating electron tunneling from the floating gate, through the SiO_2 barrier, into the oxide conduction band. The electrons are then swept by the oxide E-field over to the synapse drain. To measure the tunneling current, we fabricated a separate tunneling junction. Fig. 9 shows gate tunneling current versus oxide voltage. We fit these data with a modified FN fit, which employs a built-in potential, V_{bi}, to account for oxide traps

$$I_g = \xi(V_{dg}+V_{bi})^2 e^{-\frac{V_0}{V_{dg}+V_{bi}}} \tag{7}$$

where V_{dg} is the drain-to-gate potential and ξ, V_{bi}, and V_0 are constants. For comparison, we also show the conventional FN

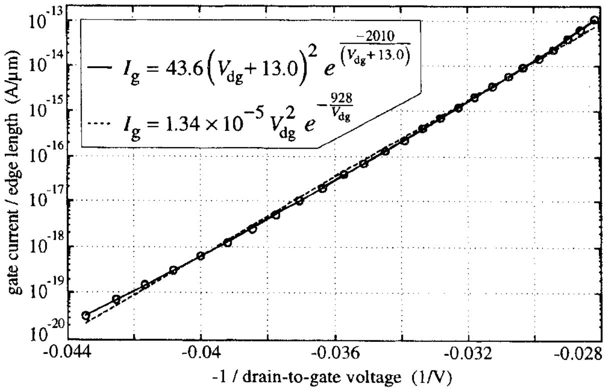


Fig. 9. Synapse-transistor tunneling gate current versus drain-to-gate voltage, normalized to the tunneling-junction edge length in lineal microns. The modified Fowler-Nordheim fit (solid line) employs a built-in voltage to account for oxide traps; the conventional Fowler-Nordheim fit (dashed line) is shown for comparison.

fit [7], [16]

$$I_g = \varphi V_{\mathrm{dg}}^2 e^{-\frac{V_f}{V_{\mathrm{dg}}}} \tag{8}$$

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

The data of Fig. 9 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$^-$ well–drain, 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 self-aligned floating gate overlaps the n$^+$. Because the tunneling current is exponential in the oxide voltage, gate-oxide tunneling in the synapse transistor is primarily an edge phenomenon.

C. The Gate-Current Equation

Because the tunneling and injection gate currents are in opposite directions, we obtain the complete synapse gate-current equation by subtracting (5) from (7)

$$I_g = \xi(V_{\mathrm{dg}} + V_{\mathrm{bi}})^2 e^{-\frac{V_0}{V_{\mathrm{dg}}+V_{\mathrm{bi}}}} - \eta I_s e^{-\frac{V_\alpha}{V_{\mathrm{gc}}} - \left(\frac{V_\beta}{V_{\mathrm{dc}}+V_\eta}\right)^2}. \tag{9}$$

The synapse exhibits four operating regimes:

1) $V_{\mathrm{dc}} < 10$ V: The tunneling and injection gate currents are both exceedingly small; the floating-gate charge is retained in a nonvolatile state.
2) 10 V $< V_{\mathrm{dc}} <$ 30 V: The tunneling current is small, but the injection current is not small; electrons are added to the floating gate, increasing the threshold voltage.
3) 30 V $< V_{\mathrm{dc}} <$ 40 V: Neither the tunneling nor the injection current is small; the floating-gate asymptotes to a voltage where the gate current of (9) is zero.
4) $V_{\mathrm{dc}} > 40$ V: The tunneling current is larger than the injection current; electrons are removed from the floating gate, decreasing the threshold voltage. Although drain voltages that transiently exceed 40 V are useful for learning, drain voltages that continuously exceed 40 V can lead to excessive power dissipation, damaging the synapse.

V. Future Development

Although the synapse already possesses those attributes that we believe are essential for building a silicon learning system, further development will improve the device substantially. We identify three areas for improvement: (A) drain voltage, (B) drain-to-gate capacitance, and (C) drain leakage current. More modern processing will readily allow these improvements.

A. Drain Voltage

The present synapse requires drain voltages up to 45 V. Although such high voltages limit potential applications, the 45 V requirement is a consequence of the 350 Å gate oxides found in the Orbit 2 μm process, rather than an inherent limitation in the synapse itself. If the synapse were fabricated in a modern EEPROM process with 80 Å oxides, it would operate from a 12 V supply. In addition, at lower voltages, the well–drain structure that we use to prevent drain breakdown could be replaced with a graded junction, reducing the synapse size.

B. Drain-to-Gate Capacitance

The synapse transistor's parasitic drain-to-gate capacitance C_{dg} is approximately 5 fF. Because the gate is floating, and the drain-voltage range is $0 < V_d < 45$ V, drain-to-gate coupling significantly affects the channel current. We reduce this effect by using a large (1 pF) gate capacitor. In a more modern process, however, two improvements are possible. First, the drain-voltage range can be smaller. Second, replacing the well–drain with a graded drain can reduce C_{dg}. These changes will permit us to use a substantially smaller gate capacitor.

C. Drain Leakage Current

As shown in Fig. 2, the poly1 gate extension that forms the tunneling junction crosses a region of field oxide (FOX). This FOX was intended to form a channel stop, preventing the channel-surface depletion layer from reaching the n$^+$ well–drain contact. Unfortunately, in the Orbit process, the FOX transistor threshold voltage $V_t \approx 20$ V. For $V_{\mathrm{dg}} > 20$ V, a parasitic p-type MOS channel forms in the n^- well, beneath the channel stop. For $V_{\mathrm{dg}} > 35$ V, pn-junction breakdown occurs where the FOX-induced, p-type channel meets the n$^+$ well contact [11].

The drain leakage current is shown in Fig. 10. Because the FOX-transistor channel conductance restricts the leakage current, the breakdown process is self-limiting. Unfortunately, junction breakdown induces a hot-electron gate current not included in (9). Although we could model this effect, we prefer to eliminate it in future synapses by using lower drain voltages or an improved channel stop.

VI. A Synaptic Array

A synaptic array, with a synapse transistor at each node, can form the basis of a silicon learning system. We fabricated the

TABLE I
THE VOLTAGES APPLIED TO THE ARRAY OF FIG. 11, TO OBTAIN THE DATA OF FIGS. 12 AND 13

	col 1 gate	col 2 gate	row 1 drain	row 2 drain	row 1 source	row 2 source
read	+5	0	+5	0	0	0
tunnel	0	+4.5	+35	0	+2	0
inject	+5	0	+25	0	0	0

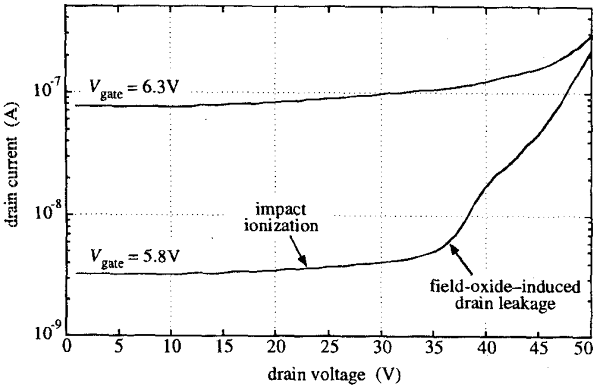


Fig. 10. Synapse-transistor drain current versus drain voltage, with the source and bulk terminals grounded.

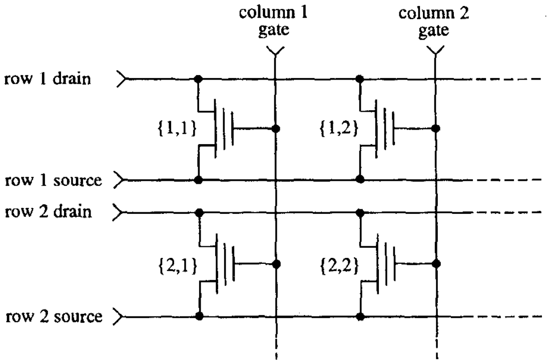


Fig. 11. A 2 × 2 synaptic array. The row synapses share a common drain wire, so tunneling at one synapse can cause undesired tunneling and injection at other row synapses.

simplified 2×2 array of Fig. 11 to investigate synapse isolation during tunneling and injection. Because a 2 × 2 array uses the same row-column addressing employed by larger arrays, it allows us to characterize completely the synapse isolation.

We chose, from among the many possible ways of using the array, to select the source current as the synapse output, and to turn off the synapses while tunneling. We applied the voltages in Table I to read, tunnel, or inject synapse $\{1, 1\}$ selectively, while ideally leaving the other synapses unchanged.

Because the synapse drain terminals are connected within a row, but not within a column, crosstalk between column synapses is small. Crosstalk between row synapses depends on the operation being performed. When a row synapse is read or injected, crosstalk to the other row synapses is small. When a row synapse is tunneled, the high drain voltage can cause both parasitic tunneling and FOX injection at other row synapses.

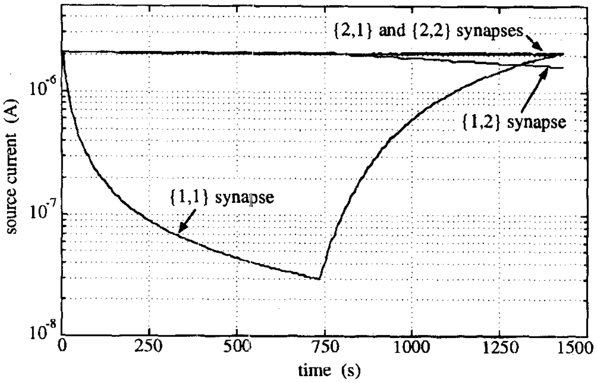


Fig. 12. Isolation in a 2 × 2 synaptic array. Source current is the synapse output. The $\{1, 1\}$ synapse first is injected down to 30 nA, then is tunneled back up to 2 μA. Crosstalk to the $\{1, 2\}$ synapse, defined as the fractional change in the $\{1, 2\}$ synapse divided by the fractional change in the $\{1, 1\}$ synapse, is 0.43%.

To obtain the data in Fig. 12, we initially set all synapses to $I_s = 2$ μA. We injected the $\{1, 1\}$ synapse down to 30 nA, and then tunneled it back up to 2 μA, 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. The $\{1, 2\}$ synapse was similarly unaffected by the injection, but during tunneling experienced both FOX injection and parasitic tunneling. A 4.7 V signal on the column 2 gate input exactly balanced these parasitic effects; unfortunately, this optimum gate voltage varied with the $\{1, 2\}$ synapse memory value. We chose a 4.5 V gate signal, so FOX injection slightly exceeded parasitic tunneling at the $\{1, 2\}$ synapse.

To obtain the data in Fig. 13, we first set all four synapses to $I_s = 30$ nA. We tunneled the $\{1, 1\}$ synapse up to 2 μA, and then injected it back down to 30 nA. Like the experiment of Fig. 12, when the $\{1, 1\}$ synapse tunneled, the $\{1, 2\}$ synapse experienced both FOX injection and parasitic tunneling. A 4.3 V gate input exactly balanced these parasitic effects. With the chosen 4.5 V gate signal, parasitic tunneling slightly exceeded FOX injection at the $\{1, 2\}$ synapse.

The measured crosstalk between row synapses was ~0.5% during tunneling, and $\ll$0.1% for all other operations. We anticipate that, with an improved channel stop and thinner gate oxide, we can achieve <0.1% crosstalk for all operations.

In the experiments of Figs. 12 and 13, the synapse tunneling and injection rates were small, for two reasons. First, the 1 pF gate capacitors that we employed to reduce the drain-to-

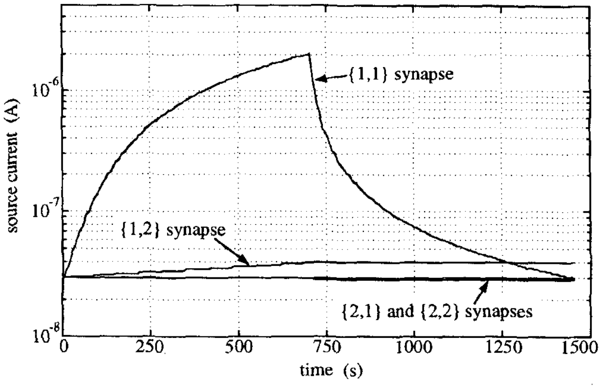


Fig. 13. The same experiment as in Fig. 12, but here the $\{1,1\}$ synapse first is tunneled up to 2 μA, then is injected back down to 30 nA. Crosstalk to the $\{1,2\}$ synapse is 0.52%.

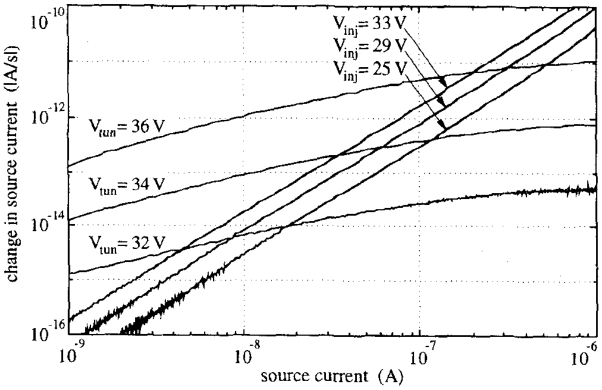


Fig. 14. Synapse-transistor delta-weight versus weight, with source current chosen as the synapse weight. We tunneled and injected synapse $\{1,1\}$, as in Fig. 13, with the source at ground and the drain at the indicated tunneling and injection voltages. We here plot the magnitude of the temporal derivative of the weight value as a function of the weight value. The mean injection slope is 2.01.

gate coupling slowed the learning rate. Second, we limited the synapse drain voltage to 35 V, to avoid FOX injection. We anticipate much faster learning rates in an improved synapse.

VII. A Synapse Learning Rule

We repeated the experiment of Fig. 13, for several tunneling and injection voltages; in Fig. 14, we plot 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 synapse weight-update rate. Starting from the gate-current equation, (9), we now derive a synapse learning rule that fits these data.

A. Injection

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

$$\frac{\partial I_s}{\partial t} = \frac{\kappa}{Q_T} I_0 e^{\frac{\kappa' V_{\rm in}}{U_t}} e^{\frac{\kappa Q_{\rm fg}}{Q_T}} \frac{\partial Q_{\rm fg}}{\partial t} = \frac{\kappa}{Q_T} I_s I_g. \tag{10}$$

As shown in Fig. 6, the gate current I_g is proportional to the source current I_s; the proportionality factor is the CHEI efficiency of (5). We add a $(-)$ sign to the gate current, because CHEI decreases the synapse weight

$$I_g = -\eta I_s e^{-\frac{V_\alpha}{V_{\rm gc}} - \left(\frac{V_\beta}{V_{\rm dc}+V_\eta}\right)^2} = -f(V_{\rm dc}, V_{\rm gc}) I_s. \tag{11}$$

We substitute (11) into (10), replacing I_s with w

$$\frac{\partial w}{\partial t} = -\frac{\kappa f(V_{\rm dc}, V_{\rm gc})}{Q_T} w^2. \tag{12}$$

For fixed drain and source voltages, $V_{\rm gc}$ increases with w, whereas $V_{\rm dc}$ decreases with w; f, which depends on both, typically increases with w. As a result, the subthreshold weight-decrement rate varies as $w^{(2+x)}$, where x represents a positive-valued correction term. However, we often operate the synapse near threshold, to increase the learning rate. For source currents near threshold, the $\partial I_s / \partial Q_{\rm fg}$ slope declines relative to its subthreshold value. For 1 nA $< w <$ 1 μA, the decreasing $\partial I_s / \partial Q_{\rm fg}$ slope counteracts the effects of the increasing f. If we assume a perfect cancellation, the weight-decrement rule, with $f(V_{\rm gc}, V_{\rm dc}) = \rho$, models accurately the data of Fig. 14.

$$\frac{\partial w}{\partial t} \approx -\frac{\kappa \rho}{Q_T} w^2. \tag{13}$$

B. Tunneling

We begin by taking the temporal derivative of (1), substituting for the gate current using (7)

$$\frac{\partial I_s}{\partial t} = \frac{\kappa \xi}{Q_T} I_0 (V_{\rm dg} + V_{\rm bi})^2 e^{\frac{\kappa' V_{\rm in}}{U_t}} e^{\frac{\kappa Q_{\rm fg}}{Q_T}} e^{-\frac{V_0}{V_{\rm dg}+V_{\rm bi}}}. \tag{14}$$

We approximate $V_{\rm db} + V_{\rm bi} \gg V_{\rm gb}$ (where $V_{\rm db}$ is the drain-to-bulk voltage, $V_{\rm gb}$ is the gate-to-bulk voltage, and $V_{\rm dg} = V_{\rm db} - V_{\rm gb}$), expand the tunneling exponential by $(1+{\rm x})^{-1} \approx 1-{\rm x}$, and solve for the weight-increment rule

$$\frac{\partial w}{\partial t} \approx \frac{\kappa \xi}{Q_T} e^{-\frac{V_0}{V_{\rm db}+V_{\rm bi}}} (V_{\rm dg} + V_{\rm bi})^2 I_0^\sigma I_s^{(1-\sigma)} \tag{15}$$

where $\sigma \equiv V_0 U_t / \kappa (V_{\rm db} + V_{\rm bi})^2$. Because, for subthreshold source currents, the floating-gate voltage changes slowly, we approximate $(V_{\rm dg} + V_{\rm bi})^2$ to be constant. We combine the constant terms into a single parameter ε, and replace I_s with w

$$\frac{\partial w}{\partial t} \approx \varepsilon w^{(1-\sigma)}. \tag{16}$$

Equation (16) models accurately the weight-increment data for subthreshold source currents. For source currents near threshold, however, the fit is poor. As the weight w increases, the floating-gate voltage increases, causing (1) the tunneling current to decrease, and (2) the $\partial I_s / \partial Q_{\rm fg}$ slope to decrease. Whereas this first effect is included in (16), the second is not. In addition, our approximation that $(V_{\rm dg} + V_{\rm bi})^2$ is a constant becomes less valid for above-threshold source currents. We therefore extend (16) with the following approximation, which models accurately the weight-increment data for channel currents up to 1 μA

$$\frac{\partial w}{\partial t} \approx (\Delta w)_{\rm max} \frac{w^{(1-\sigma)}}{w_{\rm corner} + w^{(1-\sigma)}}. \tag{17}$$

We find the maximum weight change $(\Delta w)_{\max}$, and the saturation weight value w_{corner}, by empirical measurement; the values vary with the tunneling voltage.

C. The Learning Rule

We obtain the synapse learning rule by adding (13) and (17)

$$\frac{\partial w}{\partial t} \approx (\Delta w)_{\max} \frac{w^{(1-\sigma)}}{w_{\text{corner}} + w^{(1-\sigma)}} - \frac{\kappa \rho}{Q_T} w^2. \quad (18)$$

Whereas the data of Figs. 12–14 were taken using terminal voltages chosen to prevent simultaneous injection and tunneling, we have also investigated the synapse learning for terminal voltages that permit simultaneous injection and tunneling. Equation (18) describes adequately the synapse learning for both modes of operation.

D. Learning-Rate Degradation

SiO_2 trapping is a well-known issue in floating-gate transistor reliability [18]. In digital EEPROM memories, it ultimately limits the transistor life. In the synapse, 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, and measured a $\sim$20% drop in both the weight-increment and weight-decrement learning rates. Because 1 nC of gate charge represents an enormous change in gate voltage, we believe that oxide trapping in the synapse can be ignored safely.

VIII. Conclusion

We have described a single-transistor silicon synapse with nonvolatile analog memory, simultaneous memory reading and writing, and bidirectional memory updates that are a function of both 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 characterized a synapse learning rule, and believe that we can build an autonomous learning system, combining single-transistor analog computation with memory updates computed both locally and in parallel, with this synapse.

We have discussed the limitations of the present device, and anticipate that these limitations can be reduced or eliminated with more modern processing. We claim that we can halve the present device size in the current 2 μm process; further size reductions are possible in a more modern process. Finally, we anticipate that our single-transistor synapse will allow the development of dense, low-power, silicon learning systems.

References

[1] P. Churchland and T. Sejnowski, *The Computational Brain.* Cambridge, MA: MIT Press, 1993.

[2] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses," in *Advances in Neural Information Processing Systems 7.* Cambridge, MA: MIT Press, 1995, pp. 817–824.

[3] ——, "Single transistor learning synapses with long term storage," *IEEE Int. Symp. on Circuits and Systems*, 1995, vol. 3, pp. 1660–1663.

[4] 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.

[5] ——, "A high-resolution nonvolatile analog memory cell," *IEEE Int. Symp. on Circuits and Systems*, 1995, vol. 3, pp. 2233–2236.

[6] F. Masuoka, R. Shirota, and K. Sakui, "Reviews and prospects of nonvolatile semiconductor memories," *IEICE Trans.*, vol. E74, pp. 868–874, Apr. 1991.

[7] M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown SiO_2," *J. Appl. Phys.*, vol. 40, no. 6, pp. 278–283, Jan. 1969.

[8] C. Mead, *Analog VLSI and Neural Systems.* Reading, MA: Addison-Wesley, 1989.

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

[10] C. R. Crowell and S. M. Sze, "Temperature dependence of avalanche multiplication in semiconductors," *Appl. Phys. Lett.*, vol. 9, no. 6, pp. 242–244, Sept. 1966.

[11] A. S. Grove, *Physics and Technology of Semiconductor Devices.* New York: Wiley, 1967.

[12] 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 Integ. Circ. Sig. Proc.*, vol. 8, pp. 83–114, 1995.

[13] A. G. Andreou and K. A. Boahen, "Neural information processing II," in *Analog VLSI Signal and Information Processing*, M. Ismail and T. Fiez, Eds. New York: McGraw-Hill, 1994, pp. 358–413.

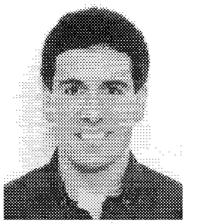
[14] W. Shockley, "Problems related to p-n junctions in silicon," *Solid-State Electron.*, vol. 2, no. 1, pp. 35–67, Pergamon Press, 1961.

[15] 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, Sept. 1984.

[16] S. M. Sze, *Physics of Semiconductor Devices.* New York: Wiley, 1981.

[17] C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *J. VLSI Sig. Proc.*, vol. 8, pp. 9–25, 1994.

[18] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of flash memory cells," in *Proc. IEEE*, vol. 81, no. 5, pp. 776–787, May 1993.



Chris Diorio (M'88) received the B.A. in physics from Occidental College, Los Angeles, CA, in 1983, and the M.S. degree in electrical engineering from the California Institute of Technology, Pasadena, in 1984. Since September 1992, he has been a doctoral candidate in electrical engineering at the California Institute of Technology.

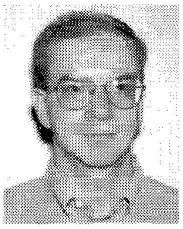
He is employed as a Staff Engineer at TRW, Inc., Redondo Beach, CA, and has worked as a Senior Staff Scientist, American Systems Corporation, Chantilly, VA, and as a Technical Consultant at The Analytic Sciences Corporation, Reston, VA. His interests include analog integrated circuit design, ultra-high-speed digital circuit design, and semiconductor device physics. His current research involves using floating-gate MOS transistors to build adaptive systems in silicon.

Mr. Diorio is a member of Sigma Pi Sigma.

Paul Hasler (S'87) received the B.S.E. and M.S. degrees in electrical engineering from Arizona State University, Tempe, in August 1991. Since September 1992, he has been a doctoral candidate in computation and neural systems at the California Institute of Technology, Pasadena.

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.

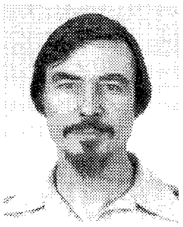
Mr. Hasler is a member of Tau Beta Pi and Eta Kappa Nu.



Bradley A. Minch (S'90) received the B.S. in electrical engineering, with distinction, from Cornell University, Ithaca, NY, in 1991. Since September 1991, he has been a doctoral candidate in computation and neural systems at the California Institute of Technology, Pasadena.

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, Eta Kappa Nu, and Phi Kappa Phi.



Carver A. Mead (S'53–M'60–SM'92–F'95) has taught at the California Institute of Technology, Pasadena, for more than 30 years, and is the Gordon and Betty Moore Professor of Engineering and Applied Science. 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 *Analog VLSI and Neural Systems* (Addison-Wesley, 1989) is concerned with modeling neuronal structures, such as the retina and the cochlea, using analog VLSI systems.

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, and a Life Fellow of the Franklin Institute. He is also the recipient of numerous awards, including the Centennial Medal of the IEEE.