A Floating-Gate MOS Learning Array with Locally Computed Weight Updates

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

Abstract—We have demonstrated on-chip learning in an array of floating-gate MOS synapse transistors. The array comprises one synapse transistor at each node, and normalization circuitry at the row boundaries. The array computes the inner product of a column input vector and a stored weight matrix. The weights are stored as floating-gate charge; they are nonvolatile, but can increase when we apply a row-learn signal. The input and learn signals are digital pulses; column input pulses that are coincident with row-learn pulses cause weight increases at selected synapses. The normalization circuitry forces row synapses to compete for floating-gate charge, bounding the weight values. The array simultaneously exhibits fast computation and slow adaptation: The inner product computes in 10 μ s, whereas the weight normalization takes minutes to hours.

I. INTRODUCTION

OUR goal is to develop silicon learning systems. We believe that these systems must possess the following attributes: high device density; low power consumption; fast, parallel computation; and slow, local adaptation. We build our learning systems as integrated circuits, achieving high device density by using MOS IC technology, effecting low power consumption by using subthreshold channel currents, and performing the requisite computations and adaptation by using innate features of the silicon-MOS physics.

We began our investigations by building single-transistor *silicon synapses* [1]–[5] modeled loosely after biological synapses [6]. Our synapse transistors are floating-gate MOSFETs; they possess nonvolatile analog weight storage, compute locally the product of their stored weight and an applied control-gate input, permit simultaneous computation and weight modification, and determine locally their own weight updates. We select source current as the synapse output, store the weights as floating-gate charge, and achieve bidirectional learning by using a combination of electron tunneling and hot-electron injection to modify the floating-gate charge.

Because our synapse transistors comprise a single device, and employ subthreshold channel currents, we can use them to build dense, low-power, silicon learning systems. Although

Manuscript received February 21, 1997; revised June 19, 1997. The review of this paper was arranged by Editor C.-Y. Lu. This work was supported by the Office of Naval Research, the Advanced Research Projects Agency, the Beckman Hearing Institute, the Center for Neuromorphic Systems Engineering as a part of the National Science Foundation Engineering Research Center Program, and 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(97)08311-1.

Fig. 1. The learning-array block diagram. For clarity, we show only 2×2 of the 4×4 synapses in the array. Each synapse multiplies its column input with its stored analog weight, and outputs a current to the row-output wire; the row wire sums the synapse-output currents along the row. The stored weights are nonvolatile; column inputs that are coincident with row-learn signals cause weight increases at selected synapses. The error signal constrains the time-averaged sum of the row-synapse weights to be a constant, bounding the row weights by forcing the synapses to compete for weight value.

a single transistor cannot model the complex behavior of a neural synapse completely, our synapse transistors can learn from an input signal without interrupting the ongoing computation.

In this paper, we demonstrate on-chip learning in a 4×4 array of our four-terminal nFET synapse transistors. We show the array block diagram in Fig. 1. The input vector comprises 10- μ s pulses; the array computes the inner product of this input vector and the stored analog weight matrix. The weights are nonvolatile; column input pulses that are coincident with row-learn pulses cause weight increases at selected synapses. To prevent unbounded weight values, we enforce a constraint: The time-averaged sum of the synapse weights, in each row of the array, is held constant. This constraint forces row synapses to compete for floating-gate charge, stabilizing the learning.

The array computation and synapse-weight modification occur locally and in parallel. We describe both the computation and weight modification using rules that we derive from the MOS-transistor and MOS-oxide physics. The array achieves our goals of fast computation and slow adaptation: The inner product computes in 10 μ s, whereas the weight normalization takes minutes to hours.

II. THE nFET SYNAPSE TRANSISTOR

We begin by reviewing our four-terminal nFET synapse transistor. As we show in Fig. 2, this device is an n-type

floating-gate MOSFET, to which we add a fourth terminal for gate-oxide tunneling. We operate the synapse from a single-polarity supply, use Fowler–Nordheim (FN) tunneling [7] to remove electrons from the floating gate, and use channel hot-electron injection (CHEI) [8] to add electrons to the floating gate. We fabricate the synapse in a 2- μm n-well CMOS process (with NPN option) available from MOSIS.

A. The Synapse Stores a Weight

We select source current as the synapse output. We apply signal inputs to the poly2 control gate, which, in turn, couples capacitively to the poly1 floating gate. We operate the MOSFET in the subthreshold regime [9], for three reasons. First, subthreshold channel currents ensure low power consumption. Second, because a subthreshold MOSFET's source current increases exponentially with gate voltage, only small quantities of oxide charge are required for learning. Third, the synapse output is the product of the stored weight and the control-gate input, as we derive from the subthreshold MOSFET equation

$$I_s = I_o e^{\kappa V_{fg}/U_t} = I_o e^{\kappa(Q_{fg} + C_{in} V_{in})/C_T U_t}$$

$$= I_o e^{Q_{fg}/Q_T} e^{\kappa' V_{in}/U_t} \quad (1)$$

$$= W I_o e^{\kappa' V_{in}/U_t} \quad (2)$$

where I_s is the source current, I_o is the pre-exponential current, κ is the coupling coefficient from the floating gate to the channel, 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 control-gate voltage, $Q_T \equiv C_T U_t/\kappa$, $\kappa' \equiv \kappa C_{in}/C_T$, $W \equiv \exp(Q_{fg}/Q_T)$, and, for simplicity, we have assumed the source potential to be ground ($V_s = 0$). The synapse weight W is the learned quantity: Its value derives from the floating-gate charge, which can change with synapse use. The synapse output is the product of W and the source current of an idealized MOSFET that has a control-gate input V_{in} and a coupling coefficient κ' from the control gate to the channel.

B. Electron Tunneling Increases the Weight

We increase W by tunneling electrons off the floating gate. In Fig. 3, we show the tunneling gate current (the oxide current) versus the reciprocal of the voltage across the tunneling oxide. We fit these data with an FN fit [7], [10]

$$I_g = I_{to} e^{-(V_f/V_{ox})} \quad (3)$$

where I_g is the gate current; V_{ox} is the oxide voltage; $V_f = 984$ V is consistent with a recent survey [11] of SiO_2 tunneling, given the synapse transistor's 400 Å gate oxide; and I_{to} is a pre-exponential current.

The present synapse requires large tunneling voltages, because the gate-oxide thickness is 400 Å. Synapses fabricated in more modern processes with thinner oxides have much lower tunneling voltages. In addition, at lower voltages, the well implant that we use for tunneling can be replaced with a graded implant, reducing the synapse size.

Fig. 2. The nFET synapse, showing the electron tunneling and injection locations. The three diagrams are aligned vertically. Diagrams A and C are drawn to scale; for clarity, we have exaggerated the vertical scale in diagram B. In the 2- μm Orbit process, the synapse length is 48 μm , and the width is 17 μm . All voltages in the conduction-band diagram are referenced to the source potential, and we have assumed subthreshold source currents ($I_s < 100$ nA). Although the gate-oxide band diagram actually projects into the plane of the page, for clarity we have rotated it by 90° and have drawn it in the channel direction. When compared with a conventional nFET, the p-type substrate implant quadruples the MOS gate-to-channel capacitance. With a 50 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, thereby increasing the coupling to 0.8.

C. CHEI Decreases the Weight

We decrease W by injecting electrons onto the floating gate. To permit CHEI with subthreshold channel currents, we add a bulk p-type implant to the synapse transistor's channel region. This implant serves two functions. First, it increases the peak drain-to-channel electric field, thereby increasing the hot-electron population in the drain-to-channel depletion region.

Fig. 3. Tunneling (gate) current I_g versus $-1/V_{ox}$. We define V_{ox} to be the potential difference between the n^+ tunneling implant and the floating gate. We fit the data using a conventional Fowler–Nordheim expression. We normalized the data to the tunneling-junction gate-to- n^+ edge length, in lineal microns, because the floating gate induces a depletion region in the lightly doped n-well, reducing the effective oxide voltage and with it the tunneling current. Because the gate cannot deplete the n^+ well contact appreciably, the oxide field is higher where the self-aligned floating gate overlaps the n^+ . Because I_g increases exponentially with V_{ox} , gate-oxide tunneling in the synapse transistor is primarily an edge phenomenon.

Fig. 4. CHEI efficiency (gate current divided by source current) versus drain-to-channel voltage. We fixed the gate-to-channel voltage at $V_{gc} = 6.1$ V during the experiment. We plot the data as efficiency because the gate current increases linearly with the source current over the entire subthreshold range. The exponential fit matches the data for the range of drain voltages encountered in the learning array.

Second, it raises the transistor's threshold voltage V_t from 0.8 to 6 V; this increase ensures that, for typical floating-gate and drain voltages of about 5.5 and 3 V, respectively, the drain-to-gate oxide electric field transports injected electrons to the floating gate, rather than returning them to the drain.

In Fig. 4, we show the CHEI efficiency (gate current I_g divided by source current I_s) versus the drain-to-channel potential V_{dc} , for a typical value of gate-to-channel potential. We plot the data versus drain-to-channel potential because the hot-electron population derives from the drain-to-channel electric field. We can re-reference our results to the source potential by using the relationship between source and channel potential in a subthreshold MOSFET [12], [13].

When V_{dc} is less than 2 V, the CHEI gate current is exceedingly small, and the weight W remains nonvolatile.

When V_{dc} is greater than 2.5 V, the CHEI gate current causes measurable changes in the synapse weight W . For reasons that we discuss in Section V, V_{dc} , in this application, typically is less than 3 V, and always is less than 3.5 V. Consequently, we approximate the data of Fig. 4 with a simple exponential

$$I_g = \beta I_s e^{V_{dc}/V_{inj}} \quad (4)$$

where I_g is the gate current; I_s is the source current; V_{dc} is the drain-to-channel potential; and β , V_{inj} are fit constants.

As a consequence of the synapse transistor's 6 V threshold, the floating-gate voltage usually exceeds 5 V, and the drain-to-gate oxide electric field strongly favors the transport of injected electrons to the floating gate. The CHEI efficiency therefore is, to first order, independent of the gate-to-channel potential, and we model the CHEI process using only (4).

D. Synapse Weight Updates Follow a Power Law

A synapse's weight updates derive from the tunneling and CHEI oxide currents that alter the floating-gate charge. Because these oxide currents vary with the synapse's terminal voltages and source current, W varies with the terminal voltages, which are imposed on the device, and with the source current, which is the synapse output. Consequently, the synapse learns: Its future output depends on both the applied input and the present output.

In Fig. 5, we show the temporal derivative of the source current versus the source current, for a synapse transistor with (part A) a set of fixed tunneling voltages, and (part B) a set of fixed drain voltages. In both experiments, we held the control-gate input V_{in} fixed; consequently, these data show the synapse weight updates $\partial W / \partial t$, as can be seen by differentiating (2).

In Appendix A, we show that the tunneling-induced weight increments follow a power law

$$\frac{\partial W}{\partial t} = \frac{1}{\tau_{tun}} W^{(1-\sigma)} \quad (5)$$

where we define σ and τ_{tun} in (16) and (17), respectively. In Appendix B, we show that the CHEI-induced weight decrements also follow a power law

$$\frac{\partial W}{\partial t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)} \quad (6)$$

where we define ε and τ_{inj} in (26) and (27), respectively.

III. THE LEARNING ARRAY

In Fig. 6, we show one row of the learning array, comprising a synapse transistor at each array node and a normalization circuit at the row boundary. The column inputs X_i and the row-learn signals Y_j are 10 μ s digital pulses. Each synapse multiplies its binary-valued input X_i with its stored weight W_{ij} , and outputs a source current I_{sij} whose magnitude is given by (2). The total row current I_{out} is the sum of the source currents from all the synapses in the row. Synapses ordinarily are on; low-true gate inputs X_i turn off selected synapses, decreasing the current I_{out} transiently. This decrease in I_{out} , in response to an input vector $\mathbf{X}$, is the row computation.

Fig. 5. Synapse-transistor (A) tunneling and (B) CHEI weight updates. In both experiments, we measured the synapse's source current I_s versus time, and plotted $(\partial I_s / \partial t)$ versus I_s . We fixed the synapse's terminal voltages; consequently, the change in I_s is a result of changes in the synapse's weight W . In part A, we applied $V_{in} = 5$ V, $V_s = 0$ V, $V_{ds} = 2$ V, and stepped V_{tun} from 29 to 35 V in 1 V increments; in part B, we applied $V_{in} = 5$ V, $V_s = 0$ V, $V_{tun} = 20$ V, and stepped V_{ds} from 2.9 to 3.5 V in 0.1 V increments. We turned off the tunneling and CHEI at regular intervals, to measure I_s . Because, for a fixed V_{in} , the synapse's weight updates $\partial W / \partial t$ are proportional to $(\partial I_s / \partial t)$ [see (2)], these data show that the weight updates follow a power law. The mean values of (σ) and (ϵ) are 0.17 and 0.24, respectively.

Synapse-weight increases occur only when both the row and column inputs, Y_j and X_i , are true. To see why, we first consider the case when the row learn signal Y_j is false (V_{tun} is low). Because $V_{ox} \equiv V_{tun} - V_{fg}$, when V_{tun} is low, V_{ox} is small for every synapse in the row. When V_{ox} is small, the tunneling currents are small, and there is no weight increase at any row synapse.

Now we consider the case when Y_j is true (V_{tun} is high). V_{ox} increases as V_{fg} decreases, and V_{fg} follows X_i . If a low-true column input X_i is true, then V_{fg} is low; V_{ox} is large, and electron tunneling causes a weight increase at the selected synapse. If, on the other hand, the low-true column input X_i is false, then V_{fg} is high; V_{ox} is too small to cause appreciable tunneling, and there is little change in the synapse's weight.

Tunneling increases the weight value of a row-column selected synapse. Because this weight update is single quadrant,

tunneling allows unbounded weight increases. To constrain the array-weight values, we renormalize the weights in each row of the array. Our array affords unsupervised learning [14], with the following constraint: The sum of the row-synapse weights, averaged over time, is a constant. The array error metric is a weight normalization; we use CHEI feedback along each row of the array to enforce the constraint.

IV. WEIGHT NORMALIZATION

The weight-normalization circuit (see Fig. 6) compares I_{sum} , the sum of the synapse drain currents in a row, with I_b , the bias current in transistor M_1 ; if $I_{sum} > I_b$, then the circuit uses CHEI to renormalize the weights. To explain the renormalization, we begin by defining row equilibrium: A row is in equilibrium when $I_{sum} = I_b$. In equilibrium, the drain voltage V_d typically causes little or no CHEI in the row synapses.

The normalization circuit constrains I_{sum} as follows: Assume that the row initially is in equilibrium, and that tunneling then raises the weight values of selected synapses, increasing I_{sum} . The excess drain current ($I_{sum} - I_b$) is mirrored by M_2 and M_3 into capacitor C_{int} , causing V_c to rise; Q_1 forces V_d to follow V_c . When V_d rises, all the row synapses undergo CHEI, decreasing all the weights, causing I_{sum} to fall. As I_{sum} falls, V_d also falls, and the row returns to equilibrium. The drain-current constraint requires that, over time, $I_{sum} = I_b$. The normalization circuit creates a negative resistance at the synapses' common drain node, causing V_d to rise when I_{sum} increases.

We now show how the drain-current constraint renormalizes the synapse weights. We begin with the constraint

$$\sum_i I_{s_i} \approx \sum_i I_{d_i} \equiv I_{sum} = I_b. \quad (7)$$

In Section V, we show that the renormalization time constant τ_a exceeds 10 s; this value is 10^6 times longer than the 10- μ s input pulses X_i (where $V_{in} = X_i$). Consequently, for renormalization, we replace V_{in} in (2) with its temporal average $\bar{V}_{in}$, and we assume that $\bar{V}_{in}$ both is time invariant and has the same value for all the row synapses. Substituting (2) into (7), we have

$$\sum_i W_i I_o e^{\kappa' \bar{V}_{in} / U_t} = I_o e^{\kappa' \bar{V}_{in} / U_t} \sum_i W_i = I_b \quad (8)$$

$$\Rightarrow \sum_i W_i = \frac{I_b}{I_o} e^{(-\kappa' \bar{V}_{in}) / U_t} \equiv W_{sum} = \text{constant}. \quad (9)$$

The drain-current and weight-value constraints are equivalent; consequently, row feedback renormalizes the synapse weights.

Renormalization forces the row synapses to compete for floating-gate charge; when one synapse's weight value increases, the sum of the weight values of its row neighbors must decrease by the same amount. However, when a selected synapse tunnels, increasing its weight, renormalization forces *all* the row synapses to undergo CHEI, decreasing *all* the row-synapse weights. The selected synapse undergoes both tunneling and CHEI; because the exponent in the CHEI weight-update rule is larger than that in the tunneling rule [see

Fig. 6. One row of the learning array. The column input vector $\mathbf{X}$ comprises low-true, 5 V, 10- μ s digital pulses; the row input vector $\mathbf{Y}$ comprises high-true, 12 V, 10- μ s digital pulses. Because the 2- μ m CMOS process that we use has 400 Å gate oxides, the tunneling voltages are high; to cause measurable tunneling, we superimpose the row inputs onto a 25 Vdc bias. The voltage coupling between a synapse's control and floating gates is about 0.8. Consequently, a 5 V (low-true) input on column wire X_1 causes a 4 V decrease in syn1's floating-gate voltage, which, in turn, causes a 4 V increase in syn1's tunneling-oxide voltage. A column input X_1 that is coincident with a row-learn pulse Y_1 causes a 16 V increase in the tunneling-oxide voltage at syn1, but only a 12 V increase at the other synapses. Because electron tunneling increases exponentially with tunneling-oxide voltage (see Fig. 3), syn1's floating gate receives about 100 times more charge than do the other synapses' floating gates; because W increases exponentially with floating-gate charge [see (2)], syn1's weight increases much more than do the other synapses' weights. The weight increase causes I_{sum} to rise, which, in turn, causes the normalization circuit to raise V_d . Because the CHEI efficiency increases with V_{ds} (see Fig. 4), a higher V_d causes CHEI in all the synapses, decreasing all the weights. The array eventually settles back to equilibrium, with I_{sum} equal to I_b , but syn1 now takes a larger share of the total row current, and the other synapses each take a smaller share. The inverting amplifier in the weight-normalization circuit enhances loop stability, for reasons that we discuss in Section V.

(5) and (6)], renormalization constrains a synapse's weight-update rate, in addition to its weight value.

Tunneling and CHEI effectively redistribute a fixed quantity of floating-gate charge among the row synapse transistors. In Appendix C, we derive the array learning rule, for coincident (x, y) pulse inputs to synapse j

$$W_{i,i \neq j}(n+1) = W_{i,i \neq j}(n) - f_{\text{learn}} W_i(n)^{(2-\varepsilon)} \quad (10)$$

$$W_j(n+1) = W_j(n) + f_{\text{learn}} \sum_{i,i \neq j} W_i(n)^{(2-\varepsilon)} \quad (11)$$

where we define ε and f_{learn} in (26) and (36), respectively. In Figs. 7 and 8, we show unsupervised learning in one row of our 4 × 4 array; these data highlight both the synapse weight and the update-rate constraints. We fit the data by applying (10) and (11), recursively; the only inputs to the fit equations are the synapse weights at $n = 0$ and the fit constants τ_{tun} , t_{pw} , σ , and ε .

V. NORMALIZATION-CIRCUIT STABILITY

The normalization circuit creates a negative resistance at the synapses' common drain node: When I_{sum} increases, V_d rises. The loop output is V_d , and the loop feedback comprises CHEI oxide currents: When V_d rises, CHEI decreases the synapse weights, causing I_{sum} to fall. Because the CHEI oxide currents increase exponentially with V_d , the loop dynamics are highly nonlinear. We therefore describe qualitative, rather than quantitative, loop-stability criteria.

The normalization circuit employs positive feedback; to ensure stability, we must make the loop gain less than unity for all frequencies. This requirement implies that the small-

signal impedance z_d , looking into the synapse drain terminals, must be greater than the total impedance z_c , at capacitor C_{int} . To see why, we assume instead that $z_c > z_d$. A rising V_d induces a small-signal current $i_{\text{sum}} = v_d/z_d$; i_{sum} is mirrored by M_2 and M_3 into C_{int} , causing V_c to rise by an amount $v_c = i_{\text{sum}} z_c$. Because V_d follows V_c , if $z_c > z_d$, then $v_c > v_d$; i_{sum} will increase rapidly, causing V_c to rise toward V_{dd} .

The impedance z_d is limited by interconnect capacitances, and by synapse-transistor channel-length modulation, floating-gate-to-drain overlap capacitance, and drain-current impact ionization. We consider each of these limitations in turn.

A. Interconnect Capacitance

Interconnect capacitance at the synapses' common drain node causes z_d to decrease with frequency. We choose C_{int} to be much larger than this parasitic capacitance, so the reactive impedance ratio, z_c/z_d , favors loop stability for all frequencies.

B. Channel-Length Modulation

Channel-length modulation reduces a synapse's drain impedance, limiting z_d . Fortunately, the synapse transistor's Early voltage exceeds 100 V, as a result of both the 10 μ m channel length and the p-type channel implant; consequently, the channel-length modulation is small.

C. Floating-Gate-to-Drain Overlap Capacitance

V_d couples to a synapse transistor's floating gate, by means of the floating-gate-to-drain overlap capacitance C_{dg} . The coupling coefficient is C_{dg}/C_T , where C_T is the total floating-gate

Fig. 7. Array learning behavior, with fits. We initialized all synapses to the same source-current value prior to starting the experiment. We first applied a train of coincident (x, y) $10\text{-}\mu\text{s}$ pulses to synapse 1, causing its weight value and source current to increase. Renormalization caused the weight values and source currents of the other synapses to decrease. Once synapse 1 had acquired 90% of the total row current, we removed the pulse-train stimulus and instead applied it to synapse 2, and then, in turn, to synapses 3 and 4. We measured the synapse source currents after every 10^3 input pulses. In the lower half of the figure, we highlight the first 1600 data points, and fit these data by applying (10) and (11), recursively. The inputs to the fit equations are the initial synapse source-current values (at $n = 0$); the pulsewidth $t_{pw} = 10\text{ }\mu\text{s}$; and the empirical constants τ_{tun} , σ , and ε . These data show that we can address individual synapses with good selectivity, and can achieve wide separation in the weight values of selected versus deselected synapses.

capacitance. Because I_s increases exponentially with V_{fg} , C_{dg} causes I_{sum} to increase exponentially with V_d , limiting z_d . To minimize the effect, we use a large interpoly capacitor ($C_T = 1\text{ pF}$); we also apply inverting feedback from V_d to the floating gate, increasing z_d (see Fig. 6). We use an off-chip amplifier to generate this inverting feedback; in future arrays, we will use instead our on-chip adaptive floating-gate amplifier [15].

D. Drain-Current Impact Ionization

Channel electrons that possess sufficient energy for CHEI also possess sufficient energy for impact ionization [16], [17]. In the synapse transistor, a drain-to-channel electric field that causes CHEI also creates additional electron-hole pairs, causing I_d to increase exponentially with V_{dc} . As a result, I_{sum} increases exponentially with V_d , limiting z_d . If V_d

Fig. 8. Array learning behavior, with fits. We replot the lower half of Fig. 7, this time on a logarithmic, rather than on a linear, scale. This plot highlights both the synapse weight and update-rate constraints, and shows that the weight values of deselected synapses do not saturate, but instead follow a power-law decay as predicted by (6) and (10).

becomes greater than about 4 V, the rate of drain-current increase causes loop instability, and V_d rises rapidly. As V_d rises, CHEI decreases all the synapse-transistor weights; as V_d saturates near V_{dd} , CHEI causes I_{sum} to fall below I_b , causing V_d to fall, and the loop to return to a stable operating regime. Loop instability causes V_d to undergo a single brief ($\sim 10\text{ }\mu\text{s}$) voltage spike, and reduces all the synapse weights substantially. Fortunately, because the synapse CHEI efficiency is high, weight renormalization rarely causes V_d to exceed 3.5 V; consequently, the loop is stable.

In Fig. 9, we show the normalization-circuit impedance versus frequency; in Fig. 10, we show the circuit's impulse response. Although the low-frequency time constant τ_a (the adaptation time constant) decreases as V_d increases, τ_a typically exceeds 10 s. The loop impulse response shows that, for short timescales, the total drain current I_{sum} can exceed I_b , violating the normalization constraint; for long timescales, $I_{sum} = I_b$.

The parasitic coupling between a synapse's tunneling junction and its floating gate is about 5 fF. With $C_T = 1\text{ pF}$, a 12 V row-learn pulse Y_j increases the floating-gate voltage of every row synapse by about 60 mV. This coupling does not affect the row computation significantly, for two reasons. First, 5 V low-true column inputs X_i always turn off selected synapses, regardless of Y_j . Second, because row-learn pulses Y_j increase the floating-gate voltage of every deselected synapse by a fixed 60 mV, we can calculate the corresponding source-current increase using (1), and can adjust I_{out} accordingly.

VI. CONCLUSION

We have shown simultaneous computation and unsupervised learning in a 4×4 array of nFET synapse transistors. The array computes the inner product of an input vector and a stored analog weight matrix. The array weights are nonvolatile; coincident row and column input pulses cause weight increases at selected synapses. We constrain the time-averaged sum of the row-synapse weights to be constant, forcing row synapses to compete for weight value.

Fig. 9. Normalization-circuit impedance magnitude versus frequency. We applied a small-signal sinusoidal current i_{in} to the synapses' row-drain node (see Fig. 6), measured the resulting small-signal voltage v_d , and plotted $z_d = v_d/i_{in}$. Because the loop feedback comprises CHEI oxide currents, which increase exponentially with V_d , the low-frequency corner increases with V_d . To hold this corner at a single frequency, we applied a constant $V_{tun} = 37$ V to all the row-synapse transistors, causing continuous tunneling. The normalization loop re-established equilibrium by setting V_d 3.3 V, inducing continuous CHEI to compensate the continuous tunneling. For these (artificial) operating conditions, the low-frequency corner comprises a single pole at about 0.03 Hz. The high-frequency rolloff comprises two poles: The first is the normalization-loop response, set by C_{int} ; the second is a consequence of interconnect capacitance at the synapses' common drain node, attenuating our injected signal i_{in} .

The array computation and synapse-weight modification occur locally and in parallel. The array achieves our goals of fast, single-transistor analog computation and of slow, locally computed weight adaptation. We describe the array computation and learning behavior using rules derived directly from the silicon-MOS and silicon-oxide physics.

SiO₂ trapping is a well-known issue in floating-gate transistor reliability [18]; in the synapse, oxide trapping decreases the weight-update rates. Fortunately, because our synapses require only small quantities of charge for their weight updates, we can ignore oxide trapping in the learning array safely.

Finally, although our array affords unsupervised learning, it uses a feedback error signal to constrain the weight values. Feedback error signals typically are used in supervised neural networks, to adjust the array weights according to the network learning rule. In future floating-gate arrays, rather than using unsupervised learning, we intend to use CHEI to adjust the synapse weights in a supervised fashion, using either pulsed, or continuously valued analog [19], inputs and row-error signals.

APPENDIX A

The Tunneling Weight-Increment Rule

We begin by taking the temporal derivative of the synapse weight W , where $W \equiv \exp(Q_{fg}/Q_T)$:

$$\frac{\partial W}{\partial t} = \frac{W}{Q_T} \frac{\partial Q_{fg}}{\partial t} = \frac{W}{Q_T} I_g, \quad (12)$$

We substitute (3) for the gate current I_g

$$\frac{\partial W}{\partial t} = \frac{I_{to}}{Q_T} W e^{-(V_f/V_{ox})}. \quad (13)$$

(a)

(b)

Fig. 10. Normalization-circuit impulse response. At time $t = 10$ min, we applied 2×10^5 coincident (x, y) 10 μ s pulses, over a 10 s period, to synapse 1. We plotted (A) the synapse source currents, and (B) the drain voltage V_d , for a period of about 5 h following the stimulus. Because the low-frequency loop time constant τ_a changes with V_d , the loop settling does not follow an RC decay; rather, the decay is closer to $1/t$. After 2 weeks, V_d was about 2.4 V. At time $t = 0$, V_d initially was decaying, because we had just finished resetting the synapse source currents to identical values.

We substitute $V_{ox} = V_{tun} - V_{fg}$ (where V_{tun} and V_{fg} are the tunneling-implant and floating-gate voltages, respectively), assume that $V_{tun} \gg V_{fg}$, expand the exponent using $(1 - x)^{-1} \approx 1 + x$, and solve

$$\frac{\partial W}{\partial t} \approx \frac{I_{to}}{Q_T} W e^{-(V_f/V_{tun}) - (V_f V_{fg}/V_{tun}^2)}. \quad (14)$$

We substitute $V_{fg} = U_t Q_{fg}/\kappa Q_T$, and solve for the tunneling weight-increment rule

$$\frac{\partial W}{\partial t} \approx \frac{1}{\tau_{tun}} W^{(1-\sigma)} \quad (15)$$

where

$$\sigma \equiv \frac{V_f U_t}{\kappa V_{tun}^2} \quad (16)$$

and

$$\tau_{tun} \equiv \frac{Q_T}{I_{to}} e^{V_f/V_{tun}}. \quad (17)$$

The parameters σ and τ_{tun} vary with the tunneling voltage V_{tun} .

APPENDIX B

The CHEI Weight-Decrement Rule

We begin by defining a synapse transistor's 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 [9] by

$$I_s = I_o e^{(\kappa V_{fg} - V_s)/U_t} \quad (18)$$

and the channel-surface potential, Ψ , is related to the floating-gate voltage, V_{fg} [12], [13] by

$$\Psi \approx \kappa V_{fg} + \Psi_o \quad (19)$$

where κ is the coupling coefficient from the floating gate to the channel, and Ψ_o derives from the MOS process parameters.

Using (18) and (19), we solve for the surface potential Ψ in terms of I_s and V_s

$$\Psi = V_s + \Psi_o + U_t \ln \left(\frac{I_s}{I_o} \right). \quad (20)$$

We now solve for V_{dc}

$$V_{dc} = V_d - \Psi = V_{ds} - \Psi_o - U_t \ln \left(\frac{I_s}{I_o} \right). \quad (21)$$

The CHEI gate current I_g is given by (4). We add a minus sign to I_g , because CHEI decreases the floating-gate charge, and substitute for V_{dc} using (21)

$$\begin{aligned} I_g &= -\beta I_s e^{[V_{ds} - \Psi_o - U_t \ln(I_s/I_o)]/V_{inj}} \\ &= -\beta I_o^{U_t/V_{inj}} e^{(V_{ds} - \Psi_o)/V_{inj}} I_s^{[1 - (U_t/V_{inj})]}. \end{aligned} \quad (22)$$

We substitute for I_s using (2), and solve

$$I_g = -\beta I_o e^{\kappa' V_{in}/U_t + (V_{ds} - \kappa' V_{in} - \Psi_o)/V_{inj}} W^{(1 - U_t/V_{inj})}. \quad (23)$$

We substitute (23) into $\partial W/\partial t$, (12)

$$\frac{\partial W}{\partial t} = -\frac{\beta I_o}{Q_T} e^{\kappa' V_{in}/U_t + (V_{ds} - \kappa' V_{in} - \Psi_o)/V_{inj}} W^{(2 - U_t/V_{inj})} \quad (24)$$

to get the final weight-decrement rule

$$\frac{\partial W}{\partial t} = -\frac{1}{\tau_{inj}} W^{(2-\varepsilon)} \quad (25)$$

where

$$\varepsilon \equiv \frac{U_t}{V_{inj}} \quad (26)$$

and

$$\tau_{inj} \equiv \frac{Q_T}{\beta I_o} e^{-\kappa' V_{in}/U_t - (V_{ds} - \kappa' V_{in} - \Psi_o)/V_{inj}}. \quad (27)$$

The low-true column-input (X_i) pulse duty cycle typically is small, so V_{in} normally is high ($V_{in} = X_i$). We therefore assume that V_{in} is a constant ($V_{in} = 5$ V) in (27).

APPENDIX C

The Array Learning Rule

We consider the row-synapse weights at discrete time intervals $t \equiv nT$, where n is the step number and T is the timestep, and derive the row-learning rule for a single coincident (x, y) input to a single row synapse. We begin with the equilibrium

condition for the row-weight normalization

$$\sum_i W_i(n) = W_{\text{sum}}. \quad (28)$$

We assume that the normalization time constant τ_a is fixed, for the following reason: Coincident (x, y) input pulses cause a weight increase at a synapse; the normalization circuit responds by establishing a drain voltage V_d for which the total weight decay, summed over all the row synapses, balances the weight increase at the single synapse. If we assume that the mean density of the coincident input pulses is time-invariant, then V_d 's mean value, $\bar{V}_d$, is constant, and therefore the low-frequency loop time constant, τ_a , also is constant.

We assume that $\tau_a \ll T$. The synapse weight values can violate (28) for times $t \ll \tau_a \ll T$, but we require that they satisfy (28) at our measurement time intervals $t = nT$. We permit array inputs at times $(t + \delta t) \equiv (n + \delta)T$, immediately after we measure the synapse weight values at $t = nT$. The array inputs comprise a pulsed column vector $\mathbf{X}(n + \delta)$, where $X_i \in [0, 1] \equiv [5 \text{ V}, 0 \text{ V}]$, and a pulsed row vector $\mathbf{Y}(n + \delta)$, where $Y_j \in [0, 1] \equiv [0 \text{ V}, 12 \text{ V}]$. Without loss of generality, we assume that at time $t = nT$, the circuit is in equilibrium, and that at $(t + \delta t + t_{pw}) \equiv (n + \delta_{pw})T$, coincident row and column inputs, of duration t_{pw} , have caused synapse j 's weight to increase

$$W_j(n + \delta_{pw}) \approx W_j(n) + \frac{\partial W_j(n)}{\partial t} t_{pw} \quad (29)$$

$$\approx W_j(n) + \frac{t_{pw}}{\tau_{\text{tun}}} W_j(n)^{(1-\sigma)} \quad (30)$$

where in (29) we have made the first-order approximation that $\partial W/\partial t$ is constant over t_{pw} , and in (30) we have substituted for $\partial W/\partial t$ using (5). Because $t_{pw} \ll \tau_a$, at time $(n + \delta_{pw})$ the circuit no longer is in equilibrium

$$\sum_i W_i(n + \delta_{pw}) > W_{\text{sum}} \quad (31)$$

and the synapse weights inject down to reestablish equilibrium.

We wish to find the synapse weights at $(n + 1)$, when the row again satisfies (28). Using (25) and (30), we write weight-decrement expressions for the row synapses

$$\begin{aligned} \Delta W_{i, i \neq j}(n + 1) &= -\frac{T}{\tau_{inj}} W_{i, i \neq j}(n)^{(2-\varepsilon)} \end{aligned} \quad (32)$$

$$\begin{aligned} \Delta W_j(n + 1) &\approx -\frac{T}{\tau_{inj}} \left[W_j(n) + \frac{t_{pw}}{\tau_{\text{tun}}} W_j(n)^{(1-\sigma)} \right]^{(2-\varepsilon)} \end{aligned} \quad (33)$$

where, because the row drain voltage V_d settles during renormalization, τ_{inj} may vary over T (recall that $T \gg \tau_a \gg t_{pw}$). For reasonable values of V_{tun} and t_{pw} , the weight increment from a single coincident (x, y) input is small; consequently, we can simplify (33) using $(1 + x)^n \approx 1 + nx$

$$\begin{aligned} \Delta W_j(n + 1) &\approx -\frac{T}{\tau_{inj}} W_j(n)^{(2-\varepsilon)} \left[1 + (2 - \varepsilon) \frac{t_{pw}}{\tau_{\text{tun}}} W_j(n)^{-\sigma} \right]. \end{aligned} \quad (34)$$

Because τ_{inj} varies over T , we now re-express T/τ_{inj} in terms of quantities that we know at n . We equate the weight

increment at synapse j [see (30)] to the sum of the weight decrements at synapses i , $i \neq j$ (32) and j (34)

$$\begin{aligned} & \frac{t_{pw}}{\tau_{tun}} W_j(n)^{(1-\sigma)} \\ &= \frac{T}{\tau_{inj}} \sum_{i, i \neq j} W_i(n)^{(2-\varepsilon)} \\ &+ \frac{T}{\tau_{inj}} W_j(n)^{(2-\varepsilon)} \left[1 + (2-\varepsilon) \frac{t_{pw}}{\tau_{tun}} W_j(n)^{-\sigma} \right] \end{aligned} \quad (35)$$

and we solve for T/τ_{inj} :

$$\frac{T}{\tau_{inj}} = \frac{\frac{t_{pw}}{\tau_{tun}} W_j(n)^{(1-\sigma)}}{(2-\varepsilon) \frac{t_{pw}}{\tau_{tun}} W_j(n)^{(2-\varepsilon-\sigma)} + \sum_i W_i(n)^{(2-\varepsilon)}}. \quad (36)$$

We define $f_{learn} \equiv T/\tau_{inj}$, substitute f_{learn} into (32), and use (28) to solve for the row-learning rule

$$W_{i, i \neq j}(n+1) = W_{i, i \neq j}(n) - f_{learn} W_i(n)^{(2-\varepsilon)} \quad (10)$$

$$W_j(n+1) = W_j(n) + f_{learn} \sum_{i, i \neq j} W_i(n)^{(2-\varepsilon)}. \quad (11)$$

Equations (10) and (11) describe the row weight-update rule for a single coincident (x, y) pulse input to synapse j .

REFERENCES

- [1] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A semiconductor structure for long term learning," U.S. Patent 5 627, 392, May 6, 1997.
- [2] P. Hasler, C. Diorio, B. A. Minch, and C. Mead, "Single transistor learning synapses," in *Advances in Neural Information Processing Systems* 7, G. Tesauero, D. Touretzky, and T. Leen, Eds. Cambridge, MA: MIT Press, 1995, pp. 817–824.
- [3] —, "Single transistor learning synapses with long term storage," in *Proc. IEEE Int. Symp. Circuits and Systems*, 1995, vol. 3, pp. 1660–1663.
- [4] C. Diorio, P. Hasler, B. A. Minch, and C. Mead, "A single-transistor silicon synapse," *IEEE Trans. Electron Devices*, vol. 43, pp. 1972–1980, Nov. 1996.
- [5] —, "A complementary pair of four-terminal silicon synapses," *Analog Integrated Circuits and Signal Processing*, vol. 13, no. 1/2, pp. 153–166, 1997.
- [6] P. Churchland and T. Sejnowski, *The Computational Brain*. Cambridge, MA: MIT Press, 1993.
- [7] M. Lenzlinger and E. H. Snow, "Fowler–Nordheim tunneling into thermally grown SiO₂," *J. of Appl. Phys.*, vol. 40, no. 6, pp. 278–283, 1969.
- [8] J. J. Sanchez and T. A. DeMassa, "Review of carrier injection in the silicon/silicon-dioxide system," *IEEE Proc. G*, vol. 138, no. 3, pp. 377–389, 1991.
- [9] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA: Addison-Wesley, 1989.
- [10] S. M. Sze, *Physics of Semiconductor Devices*. New York: Wiley, 1981.
- [11] C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *J. VLSI Signal Processing*, vol. 8, no. 6, pp. 9–25, 1994.
- [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 Integrated Circuits and Signal Processing*, vol. 8, no. 1, 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] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1994.
- [15] P. Hasler, "Foundations of learning in analog VLSI," Ph.D. thesis, Dept. Computation and Neural Systems, CIT, Pasadena, CA, 1997.
- [16] W. Shockley, "Problems related to pn junctions in silicon," *Solid-State Electron.* vol. 2, no. 1, pp. 35–67, 1961.
- [17] S. Tam, P. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Trans. Electron Devices*, vol. 31, pp. 1116–1125, Sept. 1984.
- [18] S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masuoka, "Reliability issues of flash memory cells," *Proc. IEEE*, vol. 81, pp. 776–787, May 1993.
- [19] P. Hasler, "Foundations of learning in analog VLSI," Ph.D. thesis, Dept. Computation and Neural Systems, CIT, Pasadena, CA, 1997.

Chris Diorio (S'97) received the B.A. degree 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.

Currently, he is a Staff Engineer at TRW, Inc., Redondo Beach, CA, and has worked as a Senior Staff Scientist at 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'97) 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 SiO₂.

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

Bradley A. Minch (S'97) received the B.S. degree 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) is a Gordon and Betty Moore Professor of Engineering and Applied Science and has taught at the California Institute of Technology, Pasadena, 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 L. 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. He authored a book on this topic entitled *Analog VLSI and Neural Systems* (Reading, MA: Addison-Wesley, 1989).

Prof. 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.