

An Analog Memory Circuit for Spiking Silicon Neurons

John G. Elias, David P. M. Northmore, and Wayne Westerman

Departments of Electrical Engineering and Psychology
University of Delaware
Newark, DE 19716

A simple circuit is described that functions as an analog memory whose state and dynamics are directly controlled by pulsatile inputs. The circuit has been incorporated into a silicon neuron with a spatially extensive dendritic tree as a means of controlling the spike firing threshold of an integrate-and-fire soma. Spiking activity generated by the neuron itself and by other units in a network can thereby regulate the neuron's excitability over time periods ranging from milliseconds to many minutes. Experimental results are presented showing applications to temporal edge sharpening, bi-stable behavior, and a network that learns in the manner of classical conditioning.

Introduction

Neuromorphic engineers and other researchers interested in fabricating artificial neurons have long sought a simple and area-efficient on-chip analog memory device. Various types of analog memory devices have been used to store synaptic weights (e.g., Shoemaker et al., 1988; Holler et al., 1989; Lee et al., 1991; Murray and Tarassenko, 1994; Hasler et al., 1995), to provide offset correction in silicon retinas (Mead, 1989), to represent calcium concentration in silicon neurons (Mahowald and Douglas, 1991), to hold state in artificial dendrites (Elias and Northmore, 1995), and to set conductance parameters for voltage-sensitive channels (Douglas and Mahowald, 1995). Other applications of analog memory include setting the threshold and integration time constant in integrate-and-fire somata and the setting of dynamics and conductance parameters in silicon dendrites. Many of these applications may be best served with memory devices that respond rapidly to network activity and thus only need to retain state for short periods of time. As neuromorphic engineers draw closer to realizing accurate hardware models of neurons, the need for a simple analog memory device grows correspondingly.

Before discussing the state-of-the-art in integrated analog memory devices, it is best to provide some points of reference to which existing designs can be compared. We, therefore, define an ideal analog memory device as one that has the following attributes: fabricated with standard integrated circuit (IC) processing methods; non-volatile storage; high precision; wide dynamic range; rapid and easy programmability; small footprint; and operatable from a single power supply. There may be other desirable characteristics but these few represent the bulk of what is needed and

what is, typically, difficult to achieve.

The capacitor is the basis for most integrated circuit analog memory, as well as for several forms of what is generally regarded as digital memory. An ideal capacitor, once charged, retains its state indefinitely. It is free of voltage, temperature, and frequency effects, and its state is easily changed over a wide range. Of all integrated devices, the simple capacitor comes closest towards achieving ideal behavior. Various types of integrated planar capacitors are easily fabricated using the metal-oxide-semiconductor (MOS) standard fabrication process available through the MOS integration service (MOSIS). Foremost among the various types of integrated planar capacitors are the poly-oxide-poly and poly-oxide-substrate forms, with the former having more ideal characteristics. Because of the extreme purity of the oxide layer used to form the dielectric in these types of capacitors they can hold their state for long periods of time. For poly-oxide-substrate capacitors, the charge leakage is so low that significant amounts of information can be retained for decades (AMD, 1995).

When integrated capacitors are connected to other integrated circuit elements such as the drain or source terminals of a MOS field effect transistor (MOSFET) their charge retention times are greatly reduced. This occurs because of unavoidable charge leakage pathways that exist in the additional circuit elements. With MOSFETs, the charge leakage is primarily through reversed-biased PN junctions. Although small in magnitude, the leakage can still be large enough to drain a small capacitor of its charge in a matter of seconds. Capacitive-based digital memory such as DRAMs, therefore, need to be refreshed (recharged) periodically to maintain state. Analog memory may have to be refreshed as well. And, since a precise voltage level might be required, the refresh process may present implementation difficulties.

CAPS-and-DAC Of the three analog memories discussed in this paper, the simplest to understand is the CAPS-and-DAC, whose basic architecture is shown in Fig. 1a. The voltage on any one capacitor is precisely set by a digital-to-analog converter (DAC) and de-multiplexer. Address lines select the memory location to be set, and data lines to the DAC encode the desired voltage. Since there are unavoidable leakage paths, the voltage of each memory must be refreshed periodically by using digitally encoded states, which are stored, typically, off-chip in conventional digital memory (SRAM or DRAM). The de-multiplexer circuitry would normally be integrated along with the capacitor array but the DAC could be located off chip to save space and reduce pin count.

The rate of voltage decay due to charge leakage in CAPS-and-DAC memory depends on temperature, leakage pathway characteristics, and the capacity of each capacitor. The decay rate measured as a percentage of initial state is, to first order, independent of initial voltage. Therefore, to maintain state to a certain precision requires that refreshing occurs frequently enough to keep the associated voltage ripple, expressed as a binary fraction of the full scale swing, less than 0.5 of a least significant bit (LSB). If we assume the decay is a single exponential then the refresh frequency

needed to maintain state with a ripple less than 0.5 LSB is given by

$$f = \frac{-1.0}{\tau \cdot \log \left[1 - \frac{V_{fs}}{2V_0 \cdot (2^N - 1)} \right]} \quad (1)$$

Where τ is the time constant for voltage decay, V_{fs} is the full scale voltage, V_0 is the desired state, and N is the number of bits of precision. The worst case (i.e., the highest required refresh frequency) occurs when the desired state is equal to V_{fs} . Figure 1b is a plot of the worst case refresh frequency needed for various values of N , when the time constant for decay is 10 seconds. With this time constant, a 200 Hz refresh rate is needed to maintain a precision of 10 bits, while only a 12 Hz rate is needed to maintain 6 bits of precision.

In addition to the voltage ripple caused by the continuous decay and recharge of the capacitor there are other noise effects having to do with connecting the target capacitor to the voltage source during setting and refreshing operations. These effects, generally called clock feedthrough (e.g., Sheu and Hu, 1983; Wilson et al., 1985), can be significant depending on the design of the switching circuitry, the input voltage, and the size of the memory capacitance. The noise voltage generated is due to two effects: 1) capacitive coupling between the gate and drain electrodes that allows a portion of the gate signal to couple into the holding capacitor and 2), mobile charge left in the channel between drain and source electrodes when the transistor is turned off. A portion of this charge flows to the holding capacitor, thus changing its state.

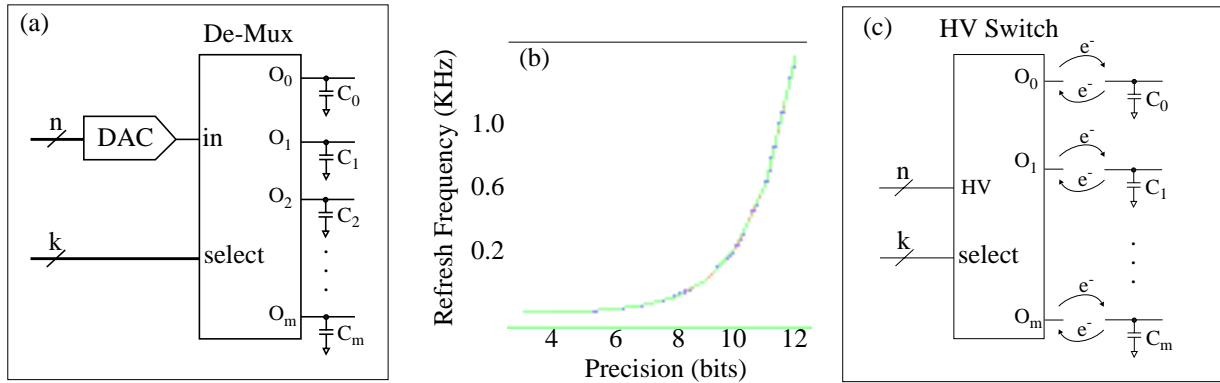


FIGURE 1. (a) Basic architecture for CAPS-and-DAC analog memory. An n -bit DAC is used to refresh capacitor voltages through a switching network. A k -bit address applied to the de-multiplexer connects a particular capacitor in the array to the DAC output which sets its state to the desired value. (b) Plot of the minimum refresh frequency needed to maintain a desired level of precision with CAPS-and-DAC analog memory when the decay time constant is 10 seconds. (c) Basic architecture for an array of floating gates. A k -bit address selects a particular floating gate for electron removal or electron addition. The direction and rate of charge flow during programming depends on the sign and magnitude of the high voltage (HV) inputs.

Floating Gates This type of analog memory (Frohman-Bentchkowsky, 1971) has several desirable characteristics: its charge retention time is extremely long, a memory cell requires only a single minimum-size transistor, and it is possible to set its state precisely. In these devices, there are essentially no pathways through which charge can flow because the capacitor serving as the memory is completely isolated from all other circuits. The capacitor is also the gate of a MOSFET, and therefore, directly affects the transistor's operation. The state of the capacitor can be set using Fowler-Nordheim tunneling (Lenzlinger and Snow, 1969), or, in some implementations, with both tunneling and hot electron injection (Diori et al., 1995). Both methods require elevated voltages and a relatively long time to set the desired state compared to CAPS-and-DAC memory. In general, the charging rate of the capacitor (floating gate) is not well known. Therefore, the accuracy and precision to which the memory can be set is limited unless a feedback technique is used (e.g., Diori et al., 1995). Floating gate memory devices were proposed for use in artificial neural networks by Alspector et al., (1987) and were integrated as a large array of over 10,000 memory elements by Holler et al., (1989). Figure 1c shows the basic architecture of a floating gate memory array in which each memory element only needs to be accessed once to set its state, which can be stable for years.

Despite the many positive attributes of CAPS-and-DAC and Floating-Gate analog memories, neither are very suitable as a network-dependent short-term memory element in artificial nervous systems. By short-term, we mean time scales of tens of microseconds to hundreds of seconds. To make CAPS-and-DAC memory devices network dependent would require constant monitoring of network activity and computing a new state for each memory cell. This could potentially be done using a dedicated processor, but it would add considerably to system complexity.

As a long-term adaptive memory element, devices based on floating gates are unequaled (e.g., Hasler et al. 1995), but they require high voltages for programming and their dynamics depend significantly on these voltage levels. By dynamics, we mean the rate of change in the memory state. In practice, it may be difficult to change floating gate memory dynamics on an individual basis, since each would require a different high-voltage level. In our work, we sought a memory device that could be fabricated in arrays, that required no high voltage supplies, and whose dynamics and state were easily set by pulsatile network activity. To this end, we developed a simple four-transistor analog memory circuit which we call a flux capacitor.

Flux Capacitor

The flux capacitor is based on techniques used with CAPS-and-DAC analog memory and switched-capacitors (e.g., Allen and Sanchez-Sinencio, 1984). It uses a conventional MOS capacitor to hold charge, but its state is set via two opposing local synapses: an upper that increases the capacitor voltage and a lower that decreases it. A diagram of the circuit is shown in Figure 2a. On every upper (S_U) synaptic activation, the holding capacitor, C_H , receives a small

packet of charge from the upper capacitor, C_u . And, with every lower (S_L) synaptic activation, C_h gives up a similarly small packet of charge to the lower capacitor, C_L . Therefore, to set the flux capacitor's state to any desired level requires a sequence of upper or lower synaptic activations. The effect that each activation has on the state of the flux capacitor is determined by the synaptic weight, which is set by the size of the synapse capacitors, C_u or C_L , relative to the holding capacitor, C_h .

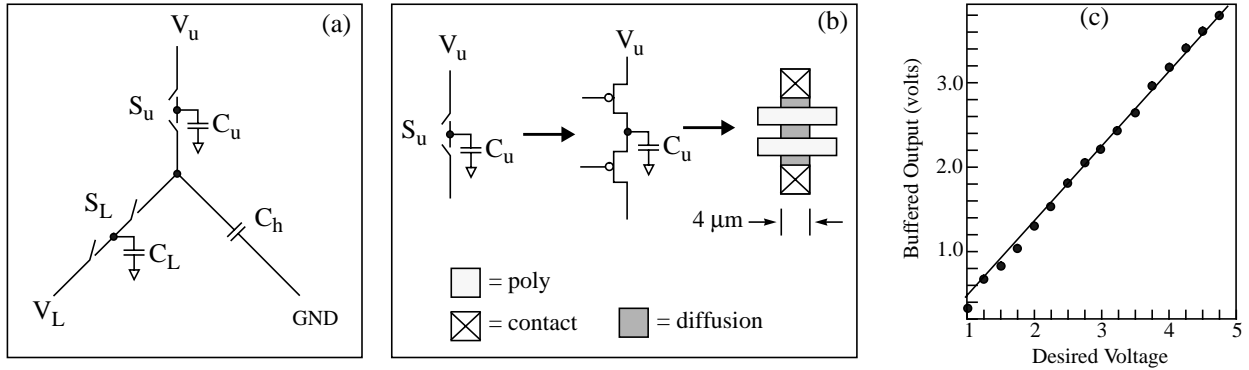


FIGURE 2. (a) Basic circuit for flux capacitor analog memory. S_u and S_L are the upper and lower synapses, respectively, V_u and V_L are the corresponding synaptic supply potentials, state is held on C_h . C_u and C_L relative to C_h are the effective synaptic weights. Each synapse is made up of two MOS transistors in series. (b) Synapse circuit schematic and VLSI layout. Only two MOS transistors make up each synapse. The physical layout is minimum size and makes use of the parasitic diffusion capacitance between the transistors to effect C_u or C_L . (c) Measured flux capacitor output voltage as a function of desired voltage computed from Eq. 5.

Each synapse (Fig. 2b) has three components: two MOS transistors which are series-connected, drain to source, and a small capacitor that connects between their common node and ground. Lower synapses use n-channel transistors; upper use p-channel. The two transistors of a particular synapse never turn on at the same time. Synaptic activation proceeds by first briefly (~ 50 nsec) turning on the transistor closest to either of the supply voltage sources (V_u or V_L) to charge the synapse capacitor to that potential. This is followed by briefly turning on the transistor nearest the holding capacitor, which either adds or removes a certain amount of charge from C_h .

The state and dynamics of the memory are set by the average charge flux entering and leaving the circuit through its synapses. The synapses are directly activated by afferent pulsatile signals coming from multiple sources (neuromorphs) anywhere in a network. A flux capacitor analog memory can represent transient or sustained information concerning the system, and, with the appropriate connections, respond to a wide range of network activities. Several examples of its utility as a dynamic memory for spike firing threshold will be presented in a latter section.

Statics

The steady state (average) flux capacitor voltage is given by

$$V = \frac{f_u \cdot W_u \cdot V_u + f_L \cdot W_L \cdot V_L + g \cdot V_g}{f_u \cdot W_u + f_L \cdot W_L + g} \quad (2)$$

where f_u and f_L are, respectively, the average frequencies at which the upper and lower synapses are activated, W_u and W_L are the respective synaptic weights, and g accounts for the state decay due to charge leaking away to voltage source V_g . Here, we make the assumption that the charge leakage between synaptic activations can be represented by a single-exponential function with time constant RC_h , where R represents the resistance of the various unwanted charge leakage pathways. The average period between synaptic activations is T . Equation 2 can also be solved for the ratio of upper to lower rates of synaptic activation given the desired voltage, V :

$$\frac{f_u}{f_L} = \frac{W_L \cdot (V - V_L) + g \cdot T \cdot (V - V_g)}{W_u \cdot (V - V_u) + g \cdot T \cdot (V - V_g)} \quad (3)$$

Definitions for the various parameters are

Upper Weight	Lower Weight	Decay Factor	Average Synapse Activation Period	
$W_u = \frac{C_u}{C_u + C_h}$	$W_L = \frac{C_L}{C_L + C_h}$	$g = \frac{1}{T} \cdot \left(1 - e^{-\frac{T}{RC_h}}\right)$	$T = \frac{1}{f_L + f_u}$	(4)

Experimentally, RC_h time constants have been found to be fairly long (~ 700 seconds with a $C_h = 0.5$ pF) compared to synaptic activation frequencies that, in our system, typically range between 0.1-500 activations/sec.

Equation 2 simplifies considerably under easily arranged conditions: equal synaptic weights (i.e., $C_L = C_u$), lower supply voltage, V_L , and leakage voltage, V_g , are zero, and the decay time constant, RC_h , is much larger than the period between synaptic activations, T ,

$$V = \frac{V_u}{1 + \frac{f_L}{f_u}} \quad (5)$$

Figure 2c shows the static output of a flux capacitor as a function of the desired voltage computed from Eq. 5. The desired voltage determined the activation frequency ratio of lower to upper synapses, which was then applied to the flux capacitor. To enable measurement of flux capacitor state we used an on-chip source-follower to buffer its output.

Therefore, all reported measurements are approximately one volt lower than the actual flux capacitor voltage. In addition, memory voltages less than 1 volt could not be measured because of the limited lower range of the source-follower buffer.

Noise The voltage on the holding capacitor, C_h , is always in a state of flux due to discrete synaptic events. Therefore, the steady state capacitor voltage will have a corresponding noise component. The magnitude of this noise will determine the precision to which the state of the flux capacitor can be set. The change in flux capacitor state due to each synaptic activation is given by

$$\Delta V_u = (W_u + SN_u) \cdot (V_u - V_h) \quad \Delta V_L = (W_L + SN_L) \cdot (V_L - V_h) \quad (6)$$

where V_h is the state before the synapse is activated, ΔV_u is the change in voltage due to activating an upper synapse, ΔV_L is the change due to activating a lower synapse, and SN_x is a constant representing the charge injection due to switching (i.e., the clock feedthrough effect). The voltage step depends on the current state, the supply potential, the synapse weight, and the charge injection factor. Henceforth, references to flux capacitor synaptic weights will include the contribution due to charge injection. As with real synapses (e.g., Shepherd and Koch, 1990) the change in membrane state due to synaptic activation is dependent on the instantaneous voltage of the local membrane and the synaptic conductance (weight).

If the synaptic weights are equal, the maximum peak-to-peak noise due to individual synaptic activations is

$$V_{pp} = \Delta V_u - \Delta V_L = W \cdot (V_u - V_L), \quad (7)$$

and the precision to which a particular state can be set in terms of bits is given by

$$N = -\log_2 \frac{V_{pp}}{V_u - V_L} = \log_2 \frac{1}{W} \quad \text{base 2} \quad (8)$$

Therefore, for 8 bit precision, the holding capacitor must be at least 256 times larger than either synapse capacitor. Of course, the relative size of the capacitors will depend on the application. In our work, we have fabricated and tested flux capacitors that have C_u/C_h ratios between 100 and 2250. In all of our flux capacitor memories, the lower and upper synapse transistors are minimum size, and, in most cases, the synapse capacitances are equal-value and derived solely from parasitic sources such as drain diffusions and short segments of wiring, which amounts to an effective capacitance on the order of 1-5 fF. The footprint required for the holding capacitor is not very large for low precision applications. For example, a 6-bit flux capacitor requires a C_h footprint of about 25 x 25 square microns for a standard

IC process available through MOSIS. The data shown in Figures 2, 3 and 4 were taken from a flux capacitor that had a C_h of approximately 0.5 pF, so its maximum effective precision was nearly 7 bits.

Since there is the inevitable change in state due to charge leakage, the precision to which a particular state can be set will also depend on the average synaptic activation rate, T . In general, the activation rate must be significantly faster than the leakage time constant to minimize the reduction in bit precision - if bit precision is important for the application. Figure 3a shows experimental data plotting flux capacitor output voltage as a function of average activation rate when both upper and lower synapses were activated at the same rate. The experimental data is plotted along with the expected output computed using Eq. 2 with parameters set to their measured or estimated values. A constant voltage of 0.9 was added to the experimental data to compensate for the source-follower transfer function. As can be seen, the output from this flux capacitor is fairly constant for frequencies above approximately 10 Hz.

An upper bound on the reduction in bit precision can be determined by equating the peak-to-peak noise voltage (Eq. 7) to an expression that describes the decay in voltage due to charge leakage alone, as in (9) below, left side. Thus, for the case when the synaptic weights are equal ($W = W_u = W_L$), the degradation in precision from theoretical, given by Eq. 8, can be limited to one bit by ensuring that the average activation period, T , is no greater than that given in (9), right side.

$$W \cdot (V_u - V_L) = (V_h - V_L) \cdot \left(1 - e^{-\frac{T}{RC_h}}\right) \quad T \leq -R \cdot C_h \cdot \log \left(1 - \frac{(V_u - V_L) \cdot W}{(V_h - V_L)}\right) \quad (9)$$

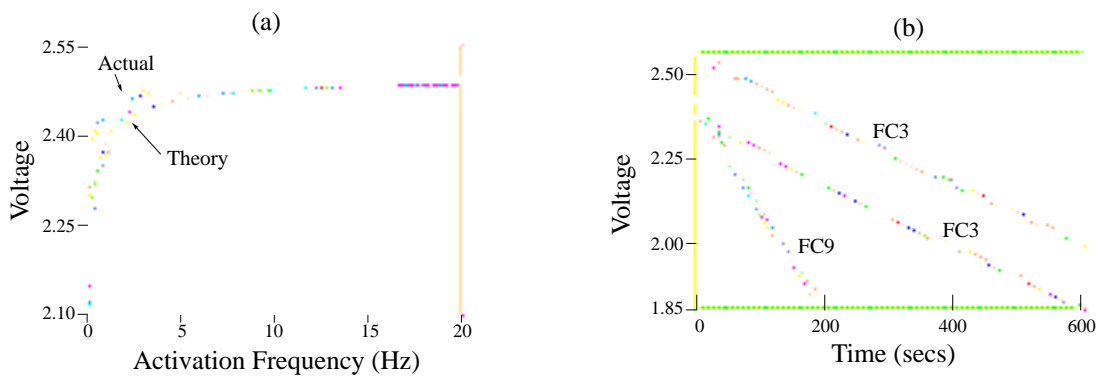


FIGURE 3. a) Flux capacitor output as a function of average activation frequency ($1/T$) when both upper and lower synapses are activated at the same rate. Theoretical curve computed from Eq. 2 using the following parameters: $W_L = W_u = 0.01$, $RC_h = 700$ sec, $V_u = 5.0$, $f_u = f_L$, $V_L = 0$, $V_I = 0$. (b) Measured flux capacitor voltage decay from a set state after synaptic activation ceases. Two different devices are shown: flux capacitor 9 (FC9) had a 25 x 40 square micron MOS holding capacitor; flux capacitor 3 had a 115 x 115 square micron poly-poly capacitor. Calculated time constants are 700 sec for FC9 and 2700 sec for FC3.

Dynamics While the flux capacitor could be used as a type of static analog memory device whose bit precision is determined by decay and synaptic activation rates, its principal use is intended to be as a network dependent dynamic memory. In nervous systems, there may exist many examples of adaptive, network-dependent effects such as membrane potentiation, modification of dendrite cable properties, or even structural changes that may have a short-term memory function. Such state changes could play an important role in modifying and regulating behavior. In artificial nervous systems, network-dependent, short-term memories may be essential for similar purposes. Although floating gate devices have been proposed as the basis for an adaptive memory (e.g., Hasler et al., 1995) we believe that the flux capacitor offers more advantages for use in short-term, network-dependent memory applications. In particular, the flux capacitor requires no elevated voltages, its memory state and dynamics are determined by direct synaptic activation, and its state can be precisely set without the need for additional feedback or monitoring circuitry.

Equation 6 shows that each synaptic activation produces either an upward or downward movement of flux capacitor voltage. Dynamics, therefore, is entirely determined by the synaptic activation rate: the more slowly flux capacitor synapses are activated, the slower its state changes (this assumes that the leakage time constant is long compared to the average rate of synaptic activation). The fastest rate at which synapses can be activated is dependent on the connection system employed (e.g., Mahowald, 1991; Elias, 1993). With our present virtual wire connection system, flux capacitor synapses could be activated as fast as 10^7 activations/sec.

The state trajectory that the flux capacitor follows depends on the ordering of synaptic activations. The final state, however, is independent of the order of activating a fixed number of upper and a fixed number of lower synapses, provided that the supply potentials (V_U or V_L) were not encountered at some point along the state trajectory. Figure 4 shows three examples of the flux capacitor state following a prescribed path. Each trajectory was generated by a particular temporal pattern of upper and lower synaptic activations applied to a flux capacitor through our virtual wire network connection circuit. The trajectories can easily be made to fit in a wide range of different time scales. As an example, the state trajectory shown in Figure 4b was generated with equal fidelity over several time scales ranging from 800 μ sec to 80 seconds. The limit to how far the time scale can be stretched or compressed depends, respectively, on the leakage decay rate and the maximum rate at which synapses can be activated.

Since the flux capacitor state can be rapidly changed and set via synaptic activations a question arises on how to embed the circuit in a network in order to produce desirable (i.e., appropriately adapting) behavior. In the next section, we present several experiments that show interesting adaptation responses to specific network activity when the flux capacitor state is used as the spike firing threshold for silicon neurons.

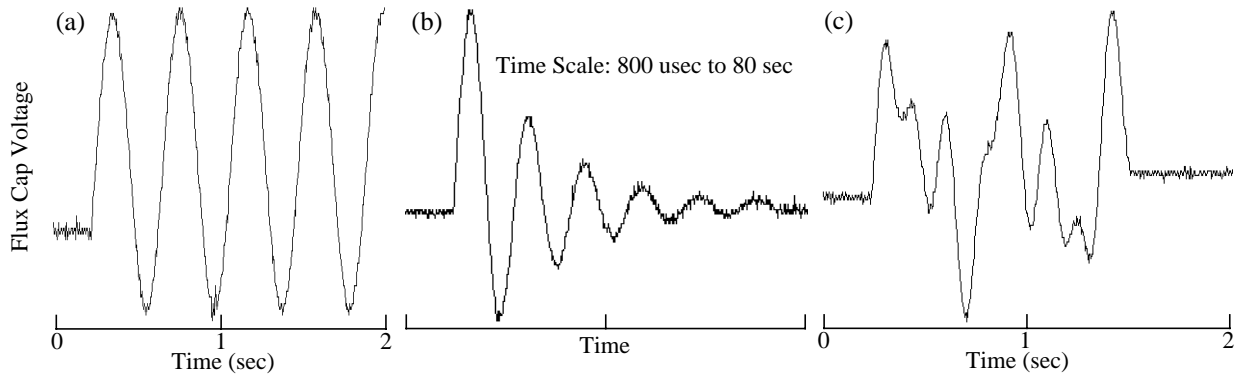


FIGURE 4. State trajectories measured from a flux capacitor using an on-chip source follower. The temporal patterns of synaptic activation that produced the trajectories were delivered to flux capacitor synapses by a network connection circuit called virtual wires (Elias, 1993). a) sine. b) damped sine. c) sum of three sine waves. The same or arbitrary trajectories can be generated over widely different time intervals. For example, data shown in (b) were collected using several different sampling periods spanning 800 usec to 80 seconds.

Network Dependent Spike Threshold

Hardware System Network dependency experiments were conducted using our silicon neuron or neuromorph (Elias and Northmore, 1995) having an integrated flux capacitor as a means of setting the spike firing threshold of its integrate-and-fire "soma" (Fig. 5). The dendritic branches of the neuromorph are composed of a series of compartments each with a capacitor representing a membrane capacitance, and two programmable resistors representing the axial cytoplasmic and membrane resistances. Synapses at each compartment (small vertical lines in Fig. 5) are emulated by a pair of MOSFETs, one of which, the excitatory synapse, enables inward current, moving the potential of the membrane capacitance in a depolarizing direction. The other transistor, emulating an inhibitory synapse, enables outward current, hyperpolarizing the membrane capacitance. There are also shunting or silent inhibitory synapses that pull the membrane potential towards a potential close to the resting value. The potential appearing at the soma (V_s , see Fig. 5) determines the output spike firing rate in conjunction with the integration time constant, RC , and the threshold, V_{th} . The spike firing threshold is derived directly from a flux capacitor whose state is set by network spiking: activating the upper threshold synapse (filled circle) raises the spike firing threshold voltage; activating the lower threshold synapse (open circle) reduces it. All synapses (on dendrites and on threshold-setting flux capacitor) are activated by a 50 nsec impulse signal.

Experiments were performed using multiple neuromorphs having 4- and 8-branched dendritic trees. The dendrite dynamics was set to give membrane time constants in the range of 10 msecs. The neuromorphs were embedded in the "virtual wire system" previously described (Elias, 1993). This system allows spikes generated by thousands of neuromorphs to be distributed with programmable delays to arbitrary synapses throughout a network. A host computer generated the spatiotemporal patterns of spikes used as test stimuli for the network. It also recorded the

soma membrane potential, V_s , its spiking threshold, V_{th} , and read the "spike history", which is part of the virtual wire system, to obtain the times of occurrence of the output spikes generated by all of the neuromorphs.

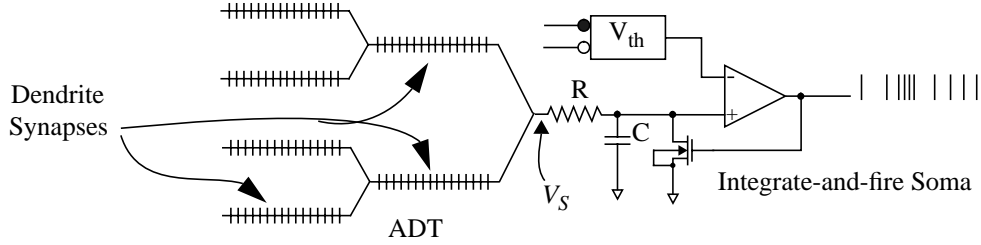


FIGURE 5. A silicon neuron with an artificial dendritic tree (ADT) and integrate-and-fire "soma". V_s is "membrane voltage" at the soma; R and C determine the integration time constant, R is programmable. The spike firing threshold, V_{th} , is the state of a flux capacitor which is set by network activity: upper threshold synapse (filled circle), when activated, raises spike firing threshold voltage; activating lower synapse reduces it.

Sustained and transient responses

Figure 6a depicts connections to a neuromorph showing how its excitability can be controlled by network activity. In the absence of other inputs to the threshold-setting synapses, a fixed threshold for the integrate-and-fire soma is achieved by delivering tonic spike trains with frequencies f_u and f_L to the upper and lower threshold-setting synapses of the soma (see Eq. 2). Using the virtual wire system, these spike trains could be derived from the activity of any neuromorph in the network, but for experimental purposes they are conveniently generated by pacemaker neuromorphs whose uniform spiking rates can be easily programmed. Because of the long time constant of the flux capacitor (~ 700 seconds), the pacemaker frequencies need to be only a few spikes per second.

An example of self-regulation of neuromorph excitability occurs when tonic spike trains are supplied to the threshold-setting synapses, and the neuromorph's own output spikes are fed back with zero delay to its threshold-setting synapses. The tonic spike input sets the spike firing threshold according to Eq. 2, which, if below the "membrane resting potential" produces output spiking in the absence of any dendritic input. The feedback connections adjust the spike firing threshold (Eq. 6) every time the neuromorph fires. In the absence of any synaptic input to the dendritic tree, the neuromorph achieves an average firing rate of f_0 spikes/sec given by

$$f_0 = \frac{-1}{RC \cdot \log \left(1 - \frac{V_u}{V_{soma} \cdot (1 + f_L/f_u)} \right)} \quad (10)$$

where V_{soma} is the "membrane potential", RC is the integrate-and-fire soma time constant (Fig. 5), f_L and f_u are the activation frequencies of the lower and upper threshold synapses. The threshold synapse activation frequencies are derived from the feedback of the neuromorph's own output and the outputs of any other neuromorphs in the network: $f_u = a \cdot f_0 + \Sigma f_{upper}$; $f_L = b \cdot f_0 + \Sigma f_{lower}$, where Σf_{upper} and Σf_{lower} represent the summed activation frequency due to all other neuromorphs whose outputs connect to the threshold setting synapses; a and b are the number of feedback connections to the upper and lower threshold synapses, respectively.

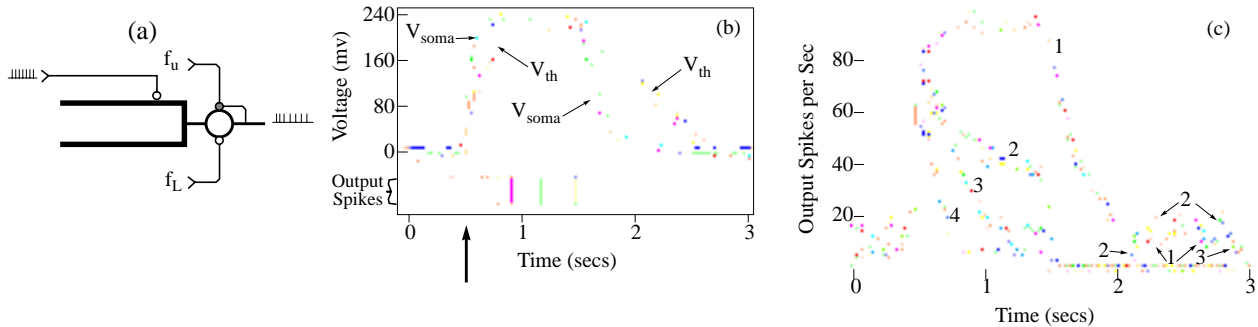


FIGURE 6. (a) Two-branch dendritic tree neuromorph. A spike train of 1 second duration is input to a proximal excitatory synapse on one branch. Tonic spike trains of frequency f_u and f_L are delivered to the upper and lower threshold-setting synapses along with various amounts of feedback to upper synapse from output. (b) Soma voltage (V_{soma}), and the threshold voltage (V_{th}) as function of time, together with output spikes. At 0.5 second (arrow), a 100 spikes/sec input train was applied to the excitatory synapse for 1 second while $f_u = 0$ and $f_L = 10$. When output stops firing (~ 1.5 sec) threshold slowly returns to previous level due to tonic activation of lower threshold synapse. (c) Spike output frequency for 4 different combinations of f_u and f_L and feedback. Curve 1: $f_u = 12.5$, $f_L = 20$, no feedback; Curve 2: $f_u = 42$, $f_L = 111$, feedback = 1 upper activation/output spike; Curve 3: $f_u = 0$, $f_L = 12.5$, feedback = 1 upper activation/output spike; Curve 4: $f_u = 0$, $f_L = 12.5$, feedback = 2 upper activations/output spike.

Figure 6b shows the response evoked by a 100 Hz input spike train of 1 second duration, delivered to an excitatory synapse on the dendritic tree. The soma potential, V_{soma} , and the threshold potential, V_{th} , are plotted together with a sample train of output spikes. A tonic spike train of 12.5 spikes/sec was applied to the lower threshold setting synapse and the neuromorph's output was fed back to its upper threshold synapse, which resulted in an average output firing rate of around 10 spikes/sec. During the 1 second input train (beginning at the arrow), V_{soma} is raised 240 mV, evoking a transient burst of output spikes followed by sustained firing at a slower rate. The output response accommodates because each fed back output spike raises the soma threshold according to Eq. 6. At the end of the input train, and for more than a second thereafter, threshold exceeds V_{soma} and spike firing ceases. With each activation of the lower threshold synapse, V_{th} drops according to Eq. 6. Eventually, V_{th} falls below V_{soma} causing the neuromorph to fire at a regular rate again. Figure 6c (curve 3) shows the instantaneous output firing frequency obtained with these connections.

The sustained response component relative to the transient component can be emphasized by increasing the rate of lower threshold synapse activation with respect to the feedback activation rate of the upper threshold synapse. With no feedback, a pure sustained response is observed (curve 1). With a single feedback activation, increased transient response occurs (curves 2 and 3). Increasing the number of feedback activations of the upper synapse from each output spike generates sharper transient responses (curve 4). In the recordings of curves 1 and 2, an additional tonic input to the upper threshold synapse was employed to sustain approximately the same average firing rate in the absence of input as with the recording of curve 3 (see Eq. 10).

Bistable behavior The connections to the threshold-setting synapses shown in Figure 7a cause V_{th} to move to one of two defined levels, allowing the neuromorph to exhibit bistable behavior. Tonic spike trains of 20 and 25 spikes/sec activating the upper and lower synapses alone move V_{th} to ~ 2.22 volts (Eq. 5, $V_u = 5$). This threshold value prevents the neuromorph from firing spontaneously. However, when the neuromorph receives a brief burst of excitatory dendritic input, thus causing it to fire, V_{th} moves low enough (~ 1.7 volts) due to feedback activations to enable continuous firing in the absence of excitatory input. The soma firing frequency is then sufficient to hold the threshold low and overcome the tendency of the tonic spike trains to raise threshold. Figure 7b shows the neuromorph firing at a spontaneous rate of about 82/sec (cf. Eq. 10 with $RC = .005$, $V_u = 5$, $V_{soma} = 1.88$, $a = 2$, $b = 4$, $\Sigma f_{upper} = 20$, $\Sigma f_{lower} = 25$), the rate rising temporarily under the influence of a 1 second train of spikes to an excitatory synapse on the dendrite. In Figure 7c, a strong inhibitory input to the dendrite shuts off firing, allowing V_{th} to rise to its upper level under the influence of the tonic input spikes. It remains in the high-threshold state (Fig. 7d), until a strong excitatory input on the dendrite generates sufficient output spiking to switch the neuromorph to the low-threshold state again (Fig. 7e). The neuromorphic circuit is capable of processing synaptic inputs in either state, provided that these are not strong enough to cause a change of state. In the high-threshold state, inputs are subjected to a thresholding nonlinearity. Excitatory input trains that are brief and relatively weak in their depolarizing effects (e.g., 50 spikes/sec for 100 msec activating a proximal synapse) elicit just a few output spikes without causing a state change. In the low-threshold state, there is a much more linear relationship between output and input.

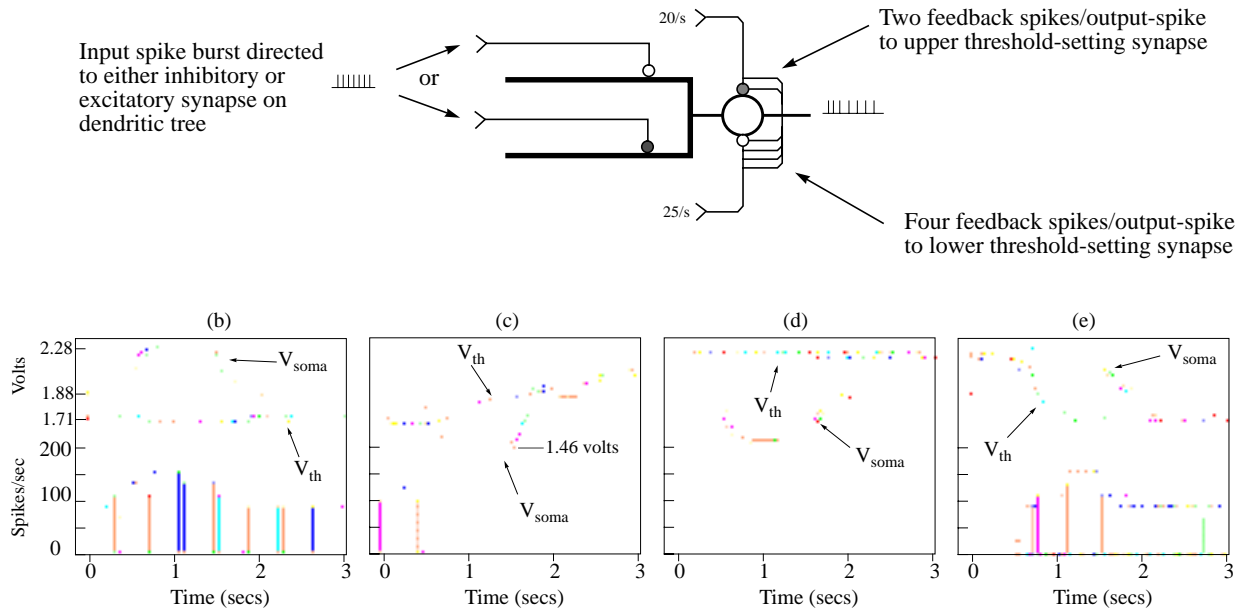


FIGURE 7. (a) A two-branch dendritic tree neuromorph with connections for bi-stable behavior. A one-second-long input spike train of 100 spikes/sec was delivered to either a strong excitatory synapse shown by open circle on dendrite, or to a strong inhibitory synapse, filled circle. Tonic spike trains of 20/sec and 25/sec activated upper and lower threshold-setting synapses, respectively. Each output spike activated the upper threshold synapse twice, and the lower threshold synapse four times. (b-e) Soma voltage (V_{soma}), threshold (V_{th}), and peri-stimulus time histogram of output spike firing. (b) Response to excitatory input in low threshold state. Excitatory input does not cause low-to-high threshold state transition; (c) Inhibitory input effects change from low to high threshold state; (d) Inhibitory input in high threshold state does not affect state; (e) Excitatory input changes state from high to low threshold.

Classical conditioning As a further illustration of the application of the flux capacitor, Figure 8a shows a neuromorphic network that exhibits the main features of classical conditioning. The unconditioned (US) and conditioned stimulus (CS) are input trains of 100 Hz each lasting 0.5 seconds. Presentation of the US alone reliably excites the "Output" neuromorph leading to an unconditioned response. The CS by itself is normally unable to excite "Output", but may do so via the "Associator" neuromorph, if the "Associator" threshold is sufficiently low. Lowering this threshold is the function of the "Correlator" that fires when coincident spikes occur in the US and CS trains (Northmore and Elias, 1996a). Thus during training, when CS and US are paired as shown in Figure 8b, "Correlator" spikes delivered to the lower synapse of the "Associator's" flux capacitor gradually decrement its V_{th} , until the CS alone is able to excite "Output".

A mechanism to limit threshold lowering in the "Associator" is required to prevent it firing excessively and to raise its threshold when CS and US are uncorrelated, i.e., to extinguish the conditioned response. The threshold synapses of the "Associator" are not driven by tonic spike trains as this would shorten the time constants of V_{th} changes. The role of the "Differentiator" then, is to convert trains of spikes from several sources into brief bursts that raise the "Associator's" threshold. This is achieved by feeding back the "Differentiator's" output spikes to its own threshold-setting synapses in the ratio 4:2 (upper:lower). The effect of six threshold setting synaptic activations per output spike ensures that the "Differentiator's" threshold moves quickly to a higher level, this tends to shut off further firing. The same principle was used to produce the transient responses shown earlier in Figure 6.

Figure 8b shows a typical sequence of spike trains before and during conditioning, and during extinction with CS alone. Prior to training, the output response is evoked by the US but not the CS. As training proceeds by pairing of CS and US, so that they overlap in time, the "Correlator" fires, lowering the "Associator's" threshold. After a few pairing of CS and US, the "Associator's" response gains in vigor and comes to anticipate the US. The CS alone is then sufficient to elicit a strong output response. Then, however, continued firing of the "Differentiator" drives up the "Associator's" threshold, leading to extinction of the conditioned response. In the absence of input activation, the persistence of the conditioned association, once formed, depends only upon the decay of flux capacitor state.

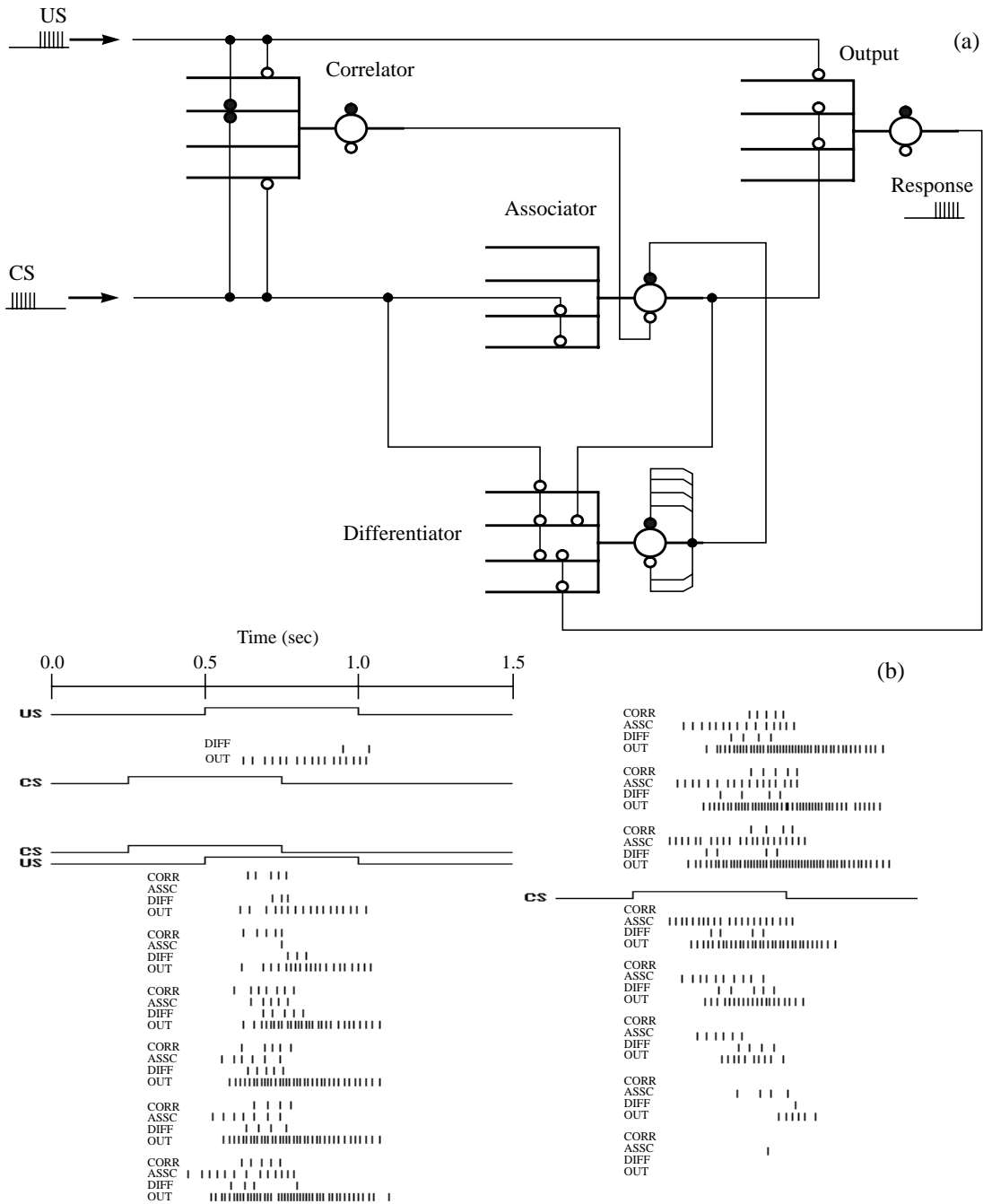


FIGURE 8. Classical conditioning in a four-neuromorph network. (a) Inputs to the network are the unconditioned stimulus (US) and the conditioned stimulus (CS), both 100 Hz trains of 0.5 sec duration. All neuromorphs have four equal dendritic branches. Open circles on the dendrites show excitatory synaptic connections; filled circles, inhibitory. Circles on the neuromorph somata are the upper and lower threshold setting synapses. The Correlator and Output neuromorph thresholds are set by tonic spike trains to their upper and lower synapses (connections not shown). Feedback to the Differentiator threshold synapses are in the ratio 4:2, upper to lower. Spike firing activity of all four neuromorphs are shown in (b). Left column, upper: before training, US alone evokes output response; CS alone evokes no activity. Training occurs by overlapped pairing of CS and US (lower left and upper right columns). The Correlator fires during the overlap of the CS and US. The Associator gradually starts responding and, with the Output response, comes to anticipate the US. Lower right column: CS presented alone. Output fires but gradually extinguishes because Differentiator acts to raise Associator's threshold.

Discussion

Negative feedback, provided by routing output spikes to the upper threshold synapse, allows neuromorphs to adjust their own excitability, thus stabilizing the spike firing frequency of the integrate-and-fire soma. Negative feedback also minimizes the effects due to variation between individual neuromorphs. However, it is control over the flux capacitor dynamics that provides temporal processing capabilities of generally longer duration than those obtained through the dendritic tree alone (Northmore and Elias, 1996b). If threshold is made to rise rapidly with output firing, the strongly transient responses shown in Fig. 6 result. With the addition of inter-neuromorphs, feedback spike frequency can be divided down, making for much slower rates of change in excitability. Thus, on a relatively short time scale, rapidly adapting effects giving temporal edge enhancement can be obtained; with longer dynamics, habituating type effects may be obtained.

Sensory systems commonly contain neurons with different temporal signaling characteristics; some are transient in nature, responding best to stimulus onsets and offsets, while others respond in a more sustained fashion. Such differences may be evident at the level of sensory receptors (e.g., mechanoreceptors in the skin), or they may be accentuated by subsequent neural filtering. Thus, for example, high-pass filtering in the insect compound eye, (McLaughlin, 1981) and in the vertebrate retina (Dowling, 1987) generate signals that mark only onsets or offsets of stimulation. Sensory systems generally transmit both sustained and transient information in parallel pathways (e.g., M- and P-cell systems; RA and SA in somatosensory system), probably as a way of maximizing channel capacity. Transients convey information that commands attention and possibly rapid response, whereas sustained responses signal the continued presence of stimuli, allowing a slower, more discriminative processing. The feedback control of neuromorph excitability makes possible a range of sustained and transient signalling that could be as important for robotics as it is for biological sensory systems.

A useful property of the flux capacitor is that it can be set to particular voltage levels depending upon the ratios of activations of the upper and lower synapses. Moreover, the level can be set by prolonged trains or by very brief bursts of activations, the level depending almost entirely upon the ratio of the number of upper to lower synapse activations (see Eq. 5). This property allows the threshold to settle at various levels depending upon which spike source dominates the threshold-setting mechanism at any given time. The resulting bistable behavior is reminiscent of the two or more modes exhibited by neurons at various sites in the CNS, most notably the thalamus (McCormick, Huguenard, and Strowbridge, 1992) where input spikes are processed differently, according to the threshold state. In the neuromorphic circuit illustrated in Fig. 7, processing is approximately linear for small signals in the low threshold state, and non-linear in the high threshold state.

The classical conditioning network was intended as a small-scale demonstration of how flux capacitors in a

network can be used for local memory storage. In this illustration, they are used either as an effective weight storage element that may be altered by training (e.g., in the Associator), or as a memory for determining the temporal characteristics of neuromorph response (as in the Differentiator). We envision that large numbers of flux capacitors, being compact, could be distributed throughout networks as short to intermediate term storage to control synaptic weights and other neuromorph parameters. The "Virtual wire" interconnection system (Elias, 1993) that made the tiny classical conditioning network possible, is designed to route spikes generated by thousands of sources, including neuromorphs and external sensors, to even larger numbers of synaptic destinations, many of which will be flux capacitor storage elements.

Conclusions

We have presented a design for a new type of simple, four-transistor integrated analog memory circuit, the flux capacitor. Its state and dynamics are set by the average charge flux through its two synapses, which are directly activated with pulsatile signals originating from an arbitrary number of neuromorph sources. The principal use of the flux capacitor is intended to be as a network dependent dynamic memory in artificial nervous systems. Its state may be incremented and decremented over a wide voltage range, as well as being set to particular levels, all by means of the spiking activity generated in a network. When its state is used as a threshold voltage for an integrate-and-fire neuromorph, output firing adaptation is possible, as demonstrated here. Work is underway to use arrays of flux capacitors to bring dendrite dynamics, soma integration time constants and synaptic conductances under the control of activity in a network.

Acknowledgments

Research supported by National Science Foundation Grants (BCS-9315879 and BEF-9511674).

References

- Alspector, J., 1987. A neuromorphic VLSI learning system. *Proc. 1987 Stanford Conf. Advanced Research in VLSI*. pp. 313-349.
- Allen, P. E. and Sanchez-Sinencio, E. 1984. *Switched Capacitor Circuits*. New York: Van Nostrand Reinhold Company.
- AMD, 1995. Data sheet for AM29F010 Flash Digital Memory from Advanced Micro Devices. Guaranteed data retention time is 10 years @ 150 C and 20 years @ 125 C.
- Diori C., Hasler, P., Minch, B., and Mead, C. 1995. A high-resolution non-volatile analog memory cell. *IEEE Int. Symposium on Circuits and Systems*, pp. 2233-2236.
- Douglas, R. and Mahowald, M. A., 1995. personal communication
- Dowling, J. 1987. *The Retina: an Approachable Part of the Brain*. Harvard Univ. Press, Cambridge, Mass.
- Elias, J.G. 1993. Artificial dendritic trees. *Neural Computation*, 5, 648-664.
- Elias, J.G., and Northmore, D.P.M. 1995. Switched-capacitor neuromorphs with wide-range variable dynamics. *IEEE Trans. Neural Networks*, vol. 6, no. 6, pp. 1542-1548.
- Hasler, P., Diori, C., Minch, B. A., and Mead, C., 1995 Single transistor learning synapses. *Advances in Neural Information Processing Systems*, vol. 7, pp. 817-824.
- Holler, M., Tam, Si., Castro, H., and Benson, R. 1989. An electrically trainable artificial neural network (ETANN) with 10240 "floating gate" synapses. International Joint Conference on Neural Networks, Washington, D.C., June 1989, pp. II-191-196
- Frohman-Bentchkowsky, D., 1971. Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure, *Applied Physics Let*, vol. 18, no. 8.
- Lee, B. W., Sheu, B. J., and Yang, H. 1991. Analog floating-gate synapses for general-purpose VLSI neural computation. *IEEE Trans. on Circuits and Systems*, vol. 38, no. 6, pp. 654-658.
- Lenzlinger, M. and Snow, E. H., 1969. Fowler-Nordheim tunneling into thermally grown SiO₂. *J. Appl. Phys.*, vol. 40, pp. 278-283.
- Mahowald, M.A. (1991) Evolving Analog VLSI Neurons, in *Single Neuron Computation*, McKenna, T., Davis, J., and Zornetzer, S. F. (eds) Academic Press, Chap 15.
- Mahowald, M.A. and Douglas, R. (1991) A silicon neuron. *Nature*, 354: 515-518
- Mead, C. 1989. *Analog VLSI and Neural Systems*. Adaptive retina. in *Analog VLSI Implementations of Neural Systems*. C. Mead and M. Ismail (eds). pp. 239-246. Kluwer Academic Pub.
- McCormick, D.A., Huguenard, J. and Strowbridge, B.W. 1992. Determination of state-dependent processing in thalamus by single neuron properties and neuromodulators in *Single Neuron Computation*, Eds: T. McKenna, J. Davis & S.F. Zornetzer, Academic Press Inc., San Diego, California.
- McLaughlin, S. B. 1989 The role of sensory adaptation in the retina. *J. Exp. Biol.* 146, pp. 39-62.
- Murray, A. and Tarassenko, L. 1994. *Analogue Neural VLSI - A pulse stream approach*. Chapman & Hall, London. Chap 3.

- Northmore, D. P. M., and Elias, J. G. 1996a, in preparation
- Northmore, D. P. M., and Elias, J. G. 1996b. Spike train processing by a silicon neuromorph: the role of sublinear summation in dendrites. *Neural Computation* 8, 1245-1265.
- Shepherd, G.M. and Koch, C. 1990. Dendritic electrotonus and synaptic integration. In *Synaptic Organization of the Brain*, G. M. Shepherd, ed. pp 439-473, Oxford Univ. Press, New York.
- Sheu, B. J. and Hu, C. M., 1983. Modeling the switch-induced error voltage on a switched-capacitor. *IEEE Trans. on Circuits and Systems*, vol. cas-30, no. 12, pp. 911-913.
- Shoemaker, P., Lagnado, I., Shimabukuro, R., 1988. Artificial neural network implementation with floating gate MOS devices, in *Hardware Implementation of Neuron Nets and Synapses*, A workshop sponsored by NSF and ONR, January, 1988, San Diego, CA
- Wilson, W. B., Massoud, H. Z., Swanson, E. J., George, R. T., and Fair, R. B., 1985. Measurement and modeling of charge feedthrough in n-channel MOS analog switches. *IEEE J. Solid-State Circuits*, vol. SC-20, no. 6, pp.1206-1213.