

Chapter 5

Building Silicon Nervous Systems with Dendritic Tree Neuromorphs.

5.1 Introduction

Essential to an understanding of a brain is an understanding of the signaling taking place within it [Koch, 1997]. Questions as to the nature of neural codes are being tackled by collecting electrophysiological data and analyzing trains of action potentials or spikes for the information that they convey about sensory stimulation [Optican & Richmond, 1987; Geisler et al., 1991]. As a result of these efforts, we can now see that spike trains encode information not just in their average frequency but also in their temporal patterning [Cariani, 1994; Rieke et al., 1997](See Chapter 4, this volume).

The question as to how neurons, individually and in networks make use of and interpret such patterns is largely unanswered. Because the requisite experiments to trace information processing in living neural tissue are so difficult, simulation can be very helpful in bridging the gaps in our knowledge and developing insights. Realistic models of neurons connected as networks, can be simulated in computer software (e.g. Chapter 10) and studied for their ability to process spike-borne information. The justification for simulating networks of realistic neurons in VLSI or other special purpose hardware is that significant speed advantages can be gained. However, in the hardware developments to be described, we have intentionally restrained processing speed to physiological rates, in keeping with an interest to pursue the behavior-based or animat approach to artificial intelligence [Maes, 1993; Steels & Brooks, 1995]. The idea is that by building working systems that behave in the real world in real time one acquires an understanding of what nervous systems must achieve using fallible and variable components in a messy environment. To be informative about nervous system design, the artificial system controlling behavior should be a reasonably faithful imitation of its biological prototype. It is our view that life-like behavior is likely to be generated by life-like components. Therefore one important requirement of the neuron-like elements, or neuromorphs, composing the artificial nervous system should be that they process information in the form of spike trains.

5.1.1 Why spikes?

Signaling over long distances in nervous systems has to be carried out by the energy expensive process of propagated action potentials. Given the cost, it would be surprising if evolution had not done more with action potentials than simply encode analog variables like pressure on skin, or the wavelength of visible light into spike frequency for transmission. Sensory stimulation typically has important temporal qualities in addition to intensive qualities (e.g. the modulations of speech sounds and the roughness of surface texture to the touch), and there would be advantages to capturing this information rather directly in the form of correspondingly patterned spike trains. The physical processes underlying sensory transduction also impose temporal structure on the signals generated, as will the subsequent neural processing. Even the visual system, which is inherently slow, deals in spike trains that convey substantially more information in their temporal patterning than in their average frequency [Optican & Richmond, 1987; McClurkin et al., 1991]. Moreover, spike trains can multiplex information [Gawne et al., 1991; Victor & Purpura, 1996]. In the auditory system, for example, spike frequency can carry information about sound intensity, while the phasing of spikes indicates sound location [Geisler et al., 1991]. The relative timing of spikes from widely separated neurons in the brain may also carry meaning: if there is synchronization between them, the neurons could be dealing with the same stimulus object [Singer, 1995]. Thus, an emulated nervous system *should* traffic in spikes, not only to be realistic, but also to be efficient. Fortunately, spike communication is well suited to VLSI implementations. As pulses of very short duration, spikes can be sent rapidly to numerous, arbitrary destinations through

conventional digital circuitry [Mahowald, 1992; Elias, 1992].

5.1.2 Dendritic processing of spikes

The means of interpreting trains of spikes, and their temporal patterning is to be found in the rich variety of mechanisms available to biological neurons. The structure that must play a critical role is the dendritic tree, so prominent and elaborate in neurons like the pyramidal cells of cerebral cortex and the Purkinje cells of the cerebellum. The classical conception of the neuron held that dendrites are inexcitable, extensions of the neuronal cell body [Eccles, 1957], expanding the neuron's surface area for multiple synaptic connections and extending its reach to create a receptive field. At the same time, passive dendrites could also provide a means of differentially weighting and delaying synaptic inputs. A line of theoretical work, pioneered by Rall, has investigated the kinds of spatial and temporal processing that could be performed by dendritic trees. Even modeled simplistically as passive cables, dendrites can perform nontrivial functions such as responding preferentially to direction of movement [Rall, 1964; Northmore & Elias, 1993]. Neurons vary enormously in the complexity of their dendrites, ranging from elaborately branched trees with numerous synaptic spines to cells with no dendrites at all. Dendritic morphologies appear to reflect the kinds of temporal processing that neurons carry out [Rose & Call, 1993].

It is clear from a growing body of physiological work on neurons from many areas of the brain that dendritic membranes contain ionic channels that are voltage-dependent or influenced by intracellular second messenger systems [Hille, 1992]. Such mechanisms allow for non-linear operations, such as the amplification of postsynaptic potentials and the generation of action potentials on the dendrites. Synaptic inputs could then sum their effects linearly, but only until a local action potential is fired. In this way the results of local processing in the distal branches of a dendritic tree could have significant effects on the spike firing of the cell's output axon. Non-linearities of this kind could greatly increase dendritic processing power [Mel, 1993]. While spiking events in dendrites have been known for some time, new electrophysiological experiments show that dendrites respond to axonal firings by conducting spikes antidromically up the dendritic tree, causing changes in the efficacy of recently activated synapses [Markram et al., 1997]. We tend to think of the passive silicon dendrites described here as skeletons, eventually to be fleshed out with active membrane channels. Until then, we prefer to mimic active processes using software in conjunction with the "Virtual Wires" spike distribution system presented below [Westerman et al., 1998].

One objective is to produce a set of neuromorphs as general purpose building blocks for networks of up to a few thousand units. To make the neuromorphs sufficiently flexible for use in specific applications, their operating characteristics should be tunable. Because neurons are capable of operating in different modes at different times (e.g. thalamic neurons [McCormick et al., 1992]), they should be tunable by means of their own activity, or by that of others in a network.

In describing the hardware implementation of our neuromorphs, we discuss the design of the various components: dendrites, synapses, spike generators, and the spike distribution system. The design choices are necessarily compromises between physiological verisimilitude on the one hand, and factors such as simplicity and compactness of implementation, low power consumption, and temperature insensitivity, on the other. If we seem to prefer expedience to physiological correctness, it is that we are eager to build networks large enough to perform capably in the real world. The deficiencies in performance due to the simplifications will show up soon enough, teaching us the important features to incorporate, and something of the design principles of nervous systems.

In the next section, we introduce each of the main components with discussion of what they are intended to accomplish, and the factors that influenced the design. We briefly describe the electronic circuitry and give references to papers for details. The subsequent section shows examples of neuromorphs in action, both singly and in small networks.

5.2. Implementation in VLSI

5.2.1 Artificial dendrites

Like most computational models of neurons [Rall, 1964], our VLSI dendrites are composed of a series of identical compartments, each containing a membrane capacitance (C_m), and two resistors, R_a and R_m , the axial cytoplasmic and membrane resistances (see Fig. 5.1). Each branch, typically 16 compartments long is connected to form a tree-like structure. In the

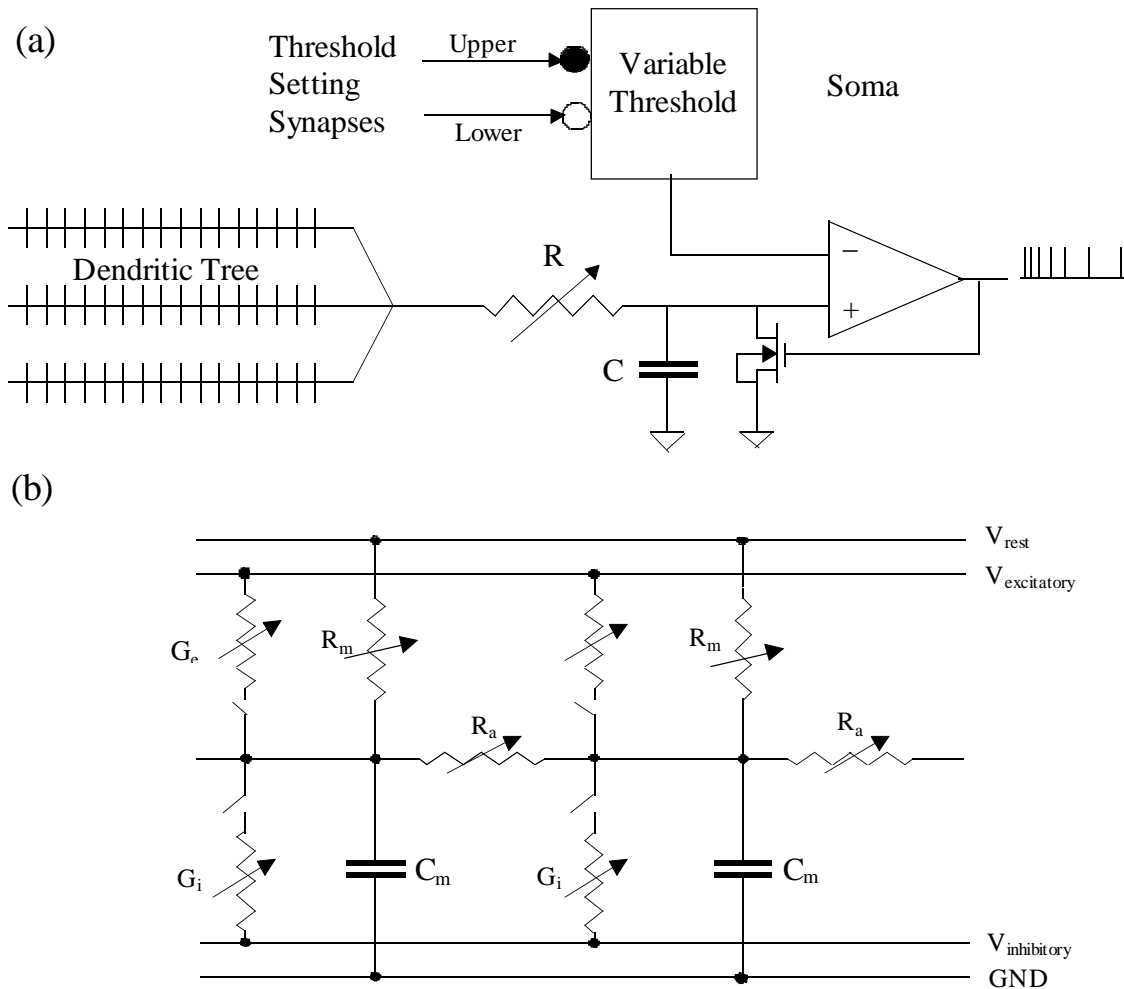


Figure 5.1. (a) Silicon neuron with dendritic tree. Synapses are located at the cross points and at the soma. Activating soma synapses sets the spike firing threshold for the integrate-and-fire soma. The integration time constant is determined by a programmable resistor, R , and a fixed capacitor, C . The integration capacitor is discharged whenever a spike is generated by the soma. (b) A two-compartment section of dendrite. Each compartment contains a storage capacitor (C_m) representing the membrane capacitance of an isopotential length of dendrite and a set of conductance paths from the capacitor to various potentials. The switched capacitor membrane (R_m) and axial (R_a) resistors, connect the compartment to a resting potential and adjacent compartments. The excitatory (G_e) and inhibitory (G_i) synaptic conductances, which turn on momentarily when a synapse is activated, pull the compartment capacitor voltage towards their respective synapse supply potentials.

examples described, we use trees with 3 – 8 primary branches connected to a "soma", but higher order branching structures more complex than that of Figure 5.1 are possible.

In order to make the dendrites reasonably compact, the membrane capacitors, C_m , should be physically small, which means their capacitance will be less than 1 pF. To obtain time constants of physiologically realistic values (e.g. 50 msec), relatively large values are required for R_a and R_m . A problem in VLSI is making high-valued resistors that are linear over wide ranges of voltage at the resistor terminals. A convenient and well-behaved solution is the switched capacitor circuit [Elias & Northmore, 1995; Allen & Sanchez-Sinencio, 1984] (see also Chapter 8, this volume). It consists of two switches in series, with a charge storage capacitor, C_h , at the junction of the two switches (See Fig. 5.2). The switches are minimum-size transistors that are driven by trains of gate pulses (ϕ_1 , ϕ_2) that do not overlap, ensuring that only one switch is on at a time. In operation, the first switch to close charges C_h to the potential of terminal A, the second switch transfers a packet of charge from C_h to terminal B. Because charge transferred is proportional to the potential difference between A and B, the device works like a resistor whose value is inversely proportional to the rate of switching. In order to achieve smooth compartment voltages, C_m must be considerably larger than C_h . This condition is satisfied by using parasitic capacitances for C_h , making the device very compact while achieving effective resistances of 500 K Ohms to 1000 G Ohms. Control of the switching signal frequencies makes it possible to change the values of R_a and R_m very easily, thereby changing the length constant, and the membrane time constant. We are also developing a system of controlling the values of R_a and R_m by spiking activity originating anywhere from the network, thereby exerting "modulatory control" over these key neuromorph parameters. Such a capability is important to incorporate because a neuron's membrane time constant may change substantially under normal operating conditions [Bernander et al., 1991].

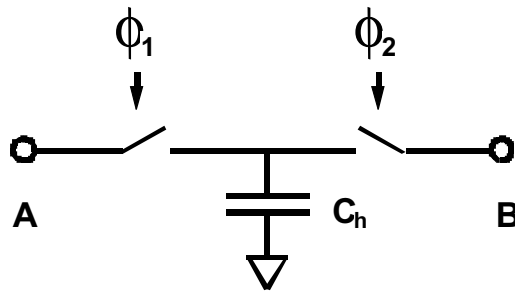


Figure 5.2. Switched capacitor circuit emulating a resistor between terminals A and B. C_h is a charge-holding capacitor formed by parasitic capacitances. ϕ_1 and ϕ_2 are non-overlapping switching signals whose frequency determines the effective resistance [Allen & Sanchez-Sinencio, 1984].

5.2.2 Synapses

A spike that is afferent to the artificial dendritic tree activates either an excitatory or an inhibitory synapse, one of each type being provided in every dendritic compartment. In early neuromorphs, synapses were emulated by transistors that acted as large, fixed conductances that briefly (50 ns) connected the compartmental capacitor to +5v in the case of excitatory synapses, or to 0v in the case of inhibitory synapses (Fig. 5.1). When one of these synapses is activated, the compartment voltage almost immediately moves to its respective supply voltage and then decays as the charge diffuses in both directions along the dendrite. The postsynaptic potential (PSP) as it appears at the soma resembles an

alpha function [Jack et al., 1975] with an amplitude that decreases exponentially with the distance of synaptic activation from the soma (Fig. 5.3). At the same time the latency of the peak of the PSP increases, roughly in proportion to the distance from the soma [Jack et al., 1975]. Thus, one of the basic functions that could be performed by the artificial dendrite is to weight and delay an input signal by amounts depending upon the location of the activated synapse.

In order to control the amplitude and the delay of PSPs independently, we have fabricated neuromorphs with variable efficacy synapses [Westerman et al., 1997]. Although it is easy to make variable current sources in CMOS technology, fast chemical synapses in the nervous system should be emulated with variable conductances. Because the amount of charge a conductance carries depends upon the potential difference between the membrane potential and the synaptic driving potential, the efficacy of a single synaptic activation will depend upon recent events that have influenced the membrane potential. This effect is important to mimic in neuromorphs because it leads to sublinear summation of synaptic activations [Rall, 1964; Koch & Poggio, 1987; Shepherd & Koch, 1990; Northmore & Elias, 1996].

Instead of making a variable conductance for each synaptic site on all the dendrites we made arrays of switchable conductances that are shared by many synapses on a chip. For the excitatory synapses, one array of 15 conductances was connected to +5v (the excitatory "reversal potential") and the conductance steps were arranged to produce equally spaced depolarizations of the compartment. Inhibitory synapses were similarly supplied by another array of 15 conductances connected to 0 v (the inhibitory "reversal potential"). Because each synapse activation is brief (50 ns), the address of the conductance selected (i.e. the weight) can be time multiplexed with the synapse address. This allows spikes coming from different sources to activate one synaptic site with different efficacies. For details of the design see [Westerman et al., 1997].

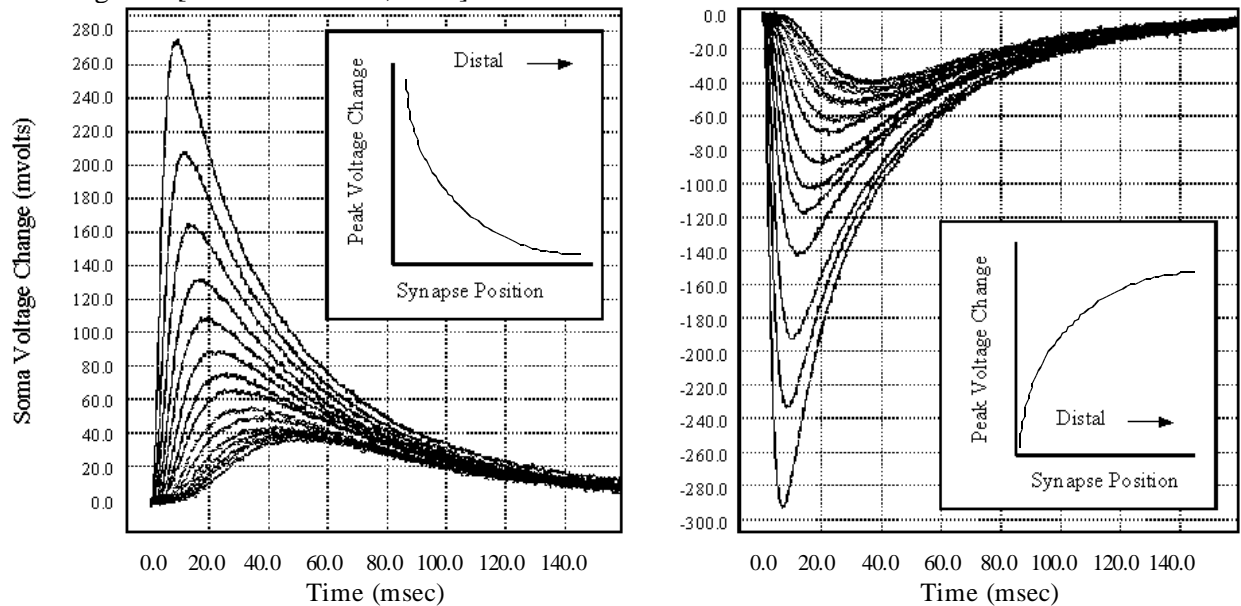


Fig 5.3. Measured impulse responses at the soma due to activating a single synapse at different locations for 100 msec. Note that the responses last over 1,000,000 times longer than the signals that caused them. Activating progressively more distal compartments evokes responses with progressively lower amplitudes and longer peak latencies.

5.2.3 Dendritic non-linearities

Because the artificial dendrite tree is a passive structure composed of linear components, a basic function that we should expect it to perform is summation. The classic modes of temporal and spatial summation of postsynaptic potentials are illustrated in Figure 5.4a–d by the activation of maximum strength synapses at different times on different branches and observing complete summation of the individual PSPs at the soma. However, the utility of the dendritic tree as a spike processor mainly derives from the fundamental non-linearity alluded to in the last section. Because synaptic activation opens a conductance to a voltage source, the charge flowing into the compartment is proportional to the difference between the present compartment potential and the driving potential. This means that if the compartment potential is already polarized by recent activations, a subsequent activation of the compartment by the same sign synapse will deliver a lesser amount of charge than if it had occurred to the compartment in a quiescent state. Since charge diffuses along the dendritic branches to the spike generator at the soma, the more recent activation will have a lesser effect on the output spike firing. This dendritic saturation effect results in sublinear summation of the two PSPs, and is therefore most severe when two activations of the same type, either both inhibitory or both excitatory occur close together in time, and in compartments that are close together (Fig. 5.4e–h). As Figure 5.4g shows, simultaneous activation of maximum strength excitatory synapses results in a PSP that is only as big as the PSP produced by one activation. As the time interval between the two activations increases, the summed PSP grows because saturation of the second impulse response diminishes. Still further increases in the interval result in less temporal summation of the individual PSPs and the summed response shrinks, the overall effect being to tune the response for activation interval. If spike threshold is set appropriately, firing of output spikes will only occur for a range of synapse activation intervals (Fig 5.4 h).

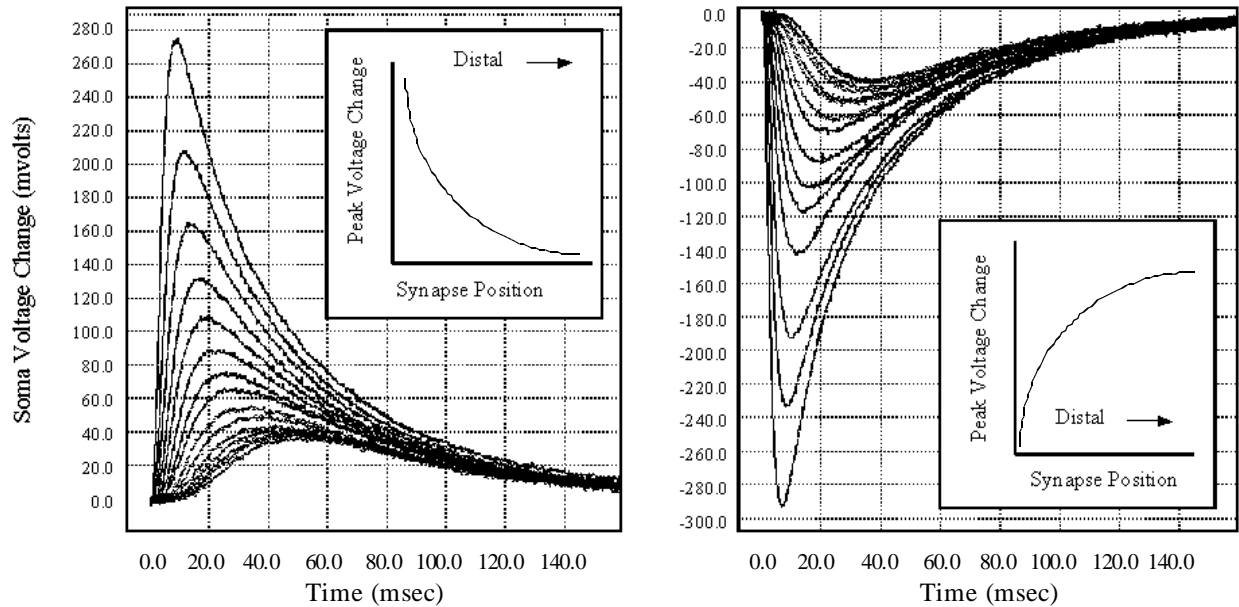


Fig 5.3. Measured impulse responses at the soma due to activating a single synapse at different locations for 100 msec. Note that the responses last over 1,000,000 times longer than the signals that caused them. Activating progressively more distal compartments evokes responses with progressively lower amplitudes and longer peak latencies.

Conditions that retard the diffusion of charge from the activated compartment stretch the time period over which the sublinear summation occurs and lengthen the preferred time interval. One way of bringing about this retardation is by increasing the axial resistance, R_a ; another is by contemporaneous activation of neighboring compartments with synapses of the same sign. The opposite effect, an acceleration of charge diffusion, can be brought about by activating neighboring compartments with synapses of opposite sign. Thus, the tuning characteristics of a segment of dendrite depend upon the configuration of synapses. Clusters of synapses on the same branch produce non-linearities and hence spike interval selectivities, whereas inputs distributed to distant synapses, especially on different branches of the dendritic tree will tend to be summed linearly [Northmore & Elias, 1996]. Similar effects may well occur in neurons, because there is good evidence that saturating non-linearities limit the response of pyramidal cells in visual cortex under normal conditions of visual stimulation [Ferster and Jagadeesh, 1992].

5.2.4 Spike Generating Soma

In the classical neuron, spike generation does not occur in the dendrites, but only in the axon initial segment and soma. Accordingly, our neuromorphs have a single spike generator in a "soma" (Fig. 5.1). Rather than attempt to model the ionic channel mechanisms responsible for firing action potentials in neurons, we employ a leaky integrate-and-fire spike generator. This is simple and compact to implement electronically, its behavior is well understood [Knight, 1972], and is widely used in simulations of spiking neurons [e.g. Gerstner et al., 1996].

The potential, V_s , appearing at the soma end of the dendritic tree is applied via a source follower to an RC integrator (see Fig 5.1). When the voltage on C exceeds V_{th} , the comparator fires a spike that also discharges C. The gain of the spike generator is determined by R, which is a switched capacitor controlled by its own clocking signal. Output spike frequency is given by:

$$F_{out} = - \left(RC \cdot \log \left(1 - \frac{V_{th}}{V_s} \right) \right)^{-1} \quad (1)$$

Note that while spike frequency is not limited by a refractory period, as it is in neurons, refractoriness could be implemented easily by having the comparator trigger a one-shot to switch on the discharge transistor for the requisite period. Although we have explored this and other methods of producing refractoriness, we have not felt compelled to introduce refractoriness for our experimental models, other forms of non-linearity, particularly at the synapses, seem to offer more immediate computational advantages [Northmore & Elias, 1996].

5.2.5 Excitability control

In addition to the fast synaptic inputs that we have considered so far, neurons are subject to slower modulatory influences that alter their excitability [Hille, 1992]. Because such mechanisms play an important part in the feedback loops that regulate neural activity, an essential feature of a neuromorph should be a capability to adjust certain parameters using the neuromorph's own activity, or the activity originating in a network of neuromorphs. Our solution was to bring these parameters under the control of spiking activity by exploiting the flexibility of the spike distribution system (see next section). The first implementation of such a capability was the control of V_{th} . The mechanism we used to control it, the "flux capacitor", turned out to have other applications, including a short-term, analog memory [Elias et al., 1997].

The flux capacitor (see Fig. 5.5) uses a conventional MOS capacitor, C_h , to hold charge. The voltage upon C_h , used for V_{th} in this case, is set by two opposing pairs of switches (referred to as Upper & Lower synapses in Figs 5.1a and 5.5). Incoming spikes addressed in the same fashion as synapses independently operate each switch pair. On every activation of the "upper synapse" by a spike, the switch closest to the supply voltage source, V_u turns on briefly, charging C_u . Then the other switch of the pair turns on briefly, transferring a small packet of charge to C_h , raising its voltage. Similarly, activation of the "lower synapse" removes charge from C_h , lowering its voltage. The device can be made very compact because only two transistors are used for each synapse and the capacitors C_u and C_L are formed by the parasitic diffusion capacitances between the transistors. The voltage stored on C_h depends upon the ratio of spike frequencies delivered to upper and lower synapses. Because the decay time constant is relatively long (ca. 700 secs with $C_h = 0.5$ pF), low frequency spike trains are sufficient to maintain V_{th} or any other control voltage used to set neuromorph parameters.

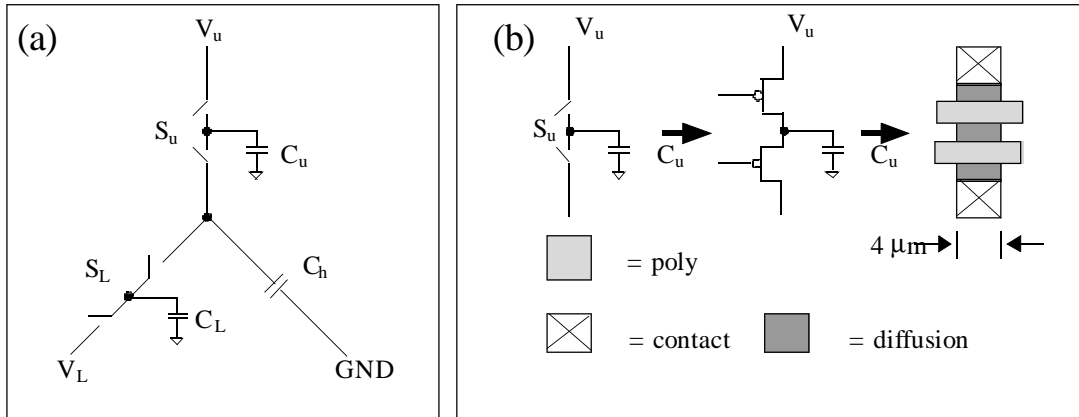


Figure 5.5. (a) Circuit for flux capacitor analog memory. S_u and S_L are the upper and lower synapses, respectively, V_u and V_L are the corresponding synaptic supply potentials. State is held on C_h . The effective synaptic weights are determined by C_u and C_L relative to C_h . (b) Synapse circuit schematic and VLSI layout. Each synapse is made up of two MOS transistors in series. The physical layout is minimum size and C_u and C_L are formed by the parasitic diffusion capacitance between the transistors.

5.2.5 Spike distribution – Virtual wires

Building networks of neuromorphs places a heavy responsibility on the system that interconnects the units. Like an axonal pathway, it should distribute spikes generated by one neuron to a potentially large number of recipient synapses. Inevitably, axons incur conduction delays, and the resulting differential delays, together with synaptic and dendritic delays can be important in neural information processing. Therefore, a prime requirement of the system should be an ability to fan out spikes from a neuromorph source, delivering them with delays specific to each destination. A practical system should also allow synaptic connections and network architecture to be set up and altered with speed and flexibility. This is especially important when searching for connection patterns by simulated learning, development or evolution. The Virtual Wire system (Fig. 5.6) is a solution that satisfies these requirements.

A number of neuromorphs, say up to 1000, residing on some 16 chips, can be accommodated on one circuit board that we call a domain. Within a domain, neuromorphs may

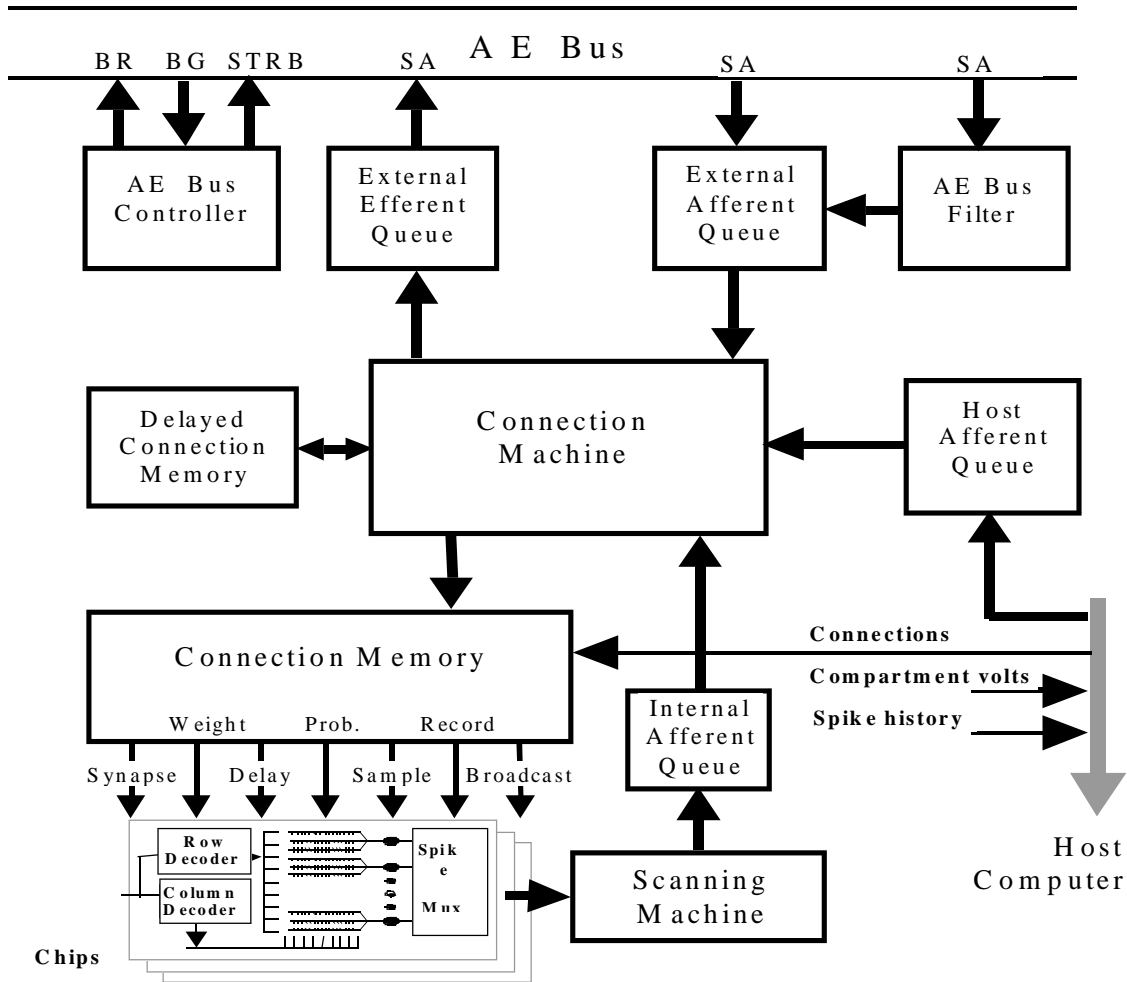


Figure 5.6. Virtual Wires System. The Scanning Machine scans through all on-board neuromorphs (currently set at 1000) in 100 μ sec, storing the addresses of spiking neuromorphs in the Internal Afferent Queue. Spike Addresses of neuromorphs from other Domains travel on the Address Event Bus and are stored in the External Afferent Queue. Spike Addresses generated by the host computer are stored in the Host Afferent Queue. The Connection Machine interrogates the various queues for addresses of spike sources. When it finds one, it activates all the target synapses of that source, either immediately, or after a delay specified in the Delayed Connection Memory.

be densely interconnected rather as in a cortical region where most of the interneuronal connections are short. The Scanning Machine (see Fig. 5.6) continuously polls all potential sources of spikes within its domain. When a spike occurs, the Connection Machine uses the spike source to look up the addresses of the destination synapses in the Connection Memory. It then activates those synapses that are due to be activated by sending address information to the neuromorph chips where row and column decoders direct a pulse to the appropriate synaptic destination. The Connection Memory also stores, together with the synapse address, a delay for activation, the synaptic weight, a synaptic activation probability, and several other bits used for sampling compartmental voltages. If an activation is to be delayed, the address and weight of its synaptic destination are stored temporarily in the Delayed Connection Memory. During its

polling, the Connection Machine also checks this memory executing any pending activations that are due.

A feature of this scheme is that a single dendritic compartment can be synaptically activated from any number of sources with different delays, weights and probabilities. In this way a dendritic compartment with its excitatory and inhibitory synapse circuits can do the duty of a segment of dendrite with multiple synapses. The scanning machine runs fast enough to introduce, at most, 0.1 msec asynchrony between activations that are supposed to be simultaneous. Longer-range connections between domains are made via an Address-Event bus [Mahowald, 1992].

5.3 Neuromorphs in action

The following experimental results were obtained from neuromorphs with dendritic trees composed of 4 primary branches of 16 compartments. Chips, with 2-4 neuromorphs, were fabricated with a 2 μ m CMOS double-poly n-well process on a 2 x 2 mm MOSIS Tiny Chip format.

5.3.1 Feedback to threshold-setting synapses.

Figure 5.7a shows connections allowing a neuromorph to regulate its own excitability via the upper and lower threshold-setting synapses on the soma. A continuous train of spikes is supplied to the lower synapse, while the neuromorph's output spikes are fed back with zero delay to the upper synapse. The neuromorph then generates spikes spontaneously at a rate, F_{out} , balancing the train to the lower synapse. The application of a train of spikes activating an excitatory synapse on the dendritic tree raises V_s well above V_{th} and the neuromorph starts to fire at a high rate, which raises V_{th} via the feedback connection. Output firing then drops, eventually to a rate slightly above its spontaneous rate. When the excitatory input train ceases, V_{th} now exceeds V_s and firing stops. Only when V_{th} has fallen to its original level does spontaneous firing resume.

One useful effect of this arrangement is to generate spike trains with onset and offset transients, very much like those generated by neurons in sensory pathways, for example. These temporal filtering characteristics which accentuate onsets and offsets can easily be changed. Making additional feedback connections from the output to the upper synapse gives a still more sharply transient response (Fig 5.7c), and other combinations can be arranged to tailor a particular response profile [Elias et al., 1997]. The output firing frequency, F_{out} , is given by

$$F_{out} = - \left(RC \cdot \log \left(1 - \frac{\left(\sum_{i=0}^N \Delta V_U + \sum_{t=0}^M \Delta V_L \right)}{V_s} \right) \right)^{-1} \quad (2)$$

where again, V_s is the "membrane potential" and RC is the time constant of the integrate-and-fire soma. V_{th} of equation (1) is replaced with the change in flux capacitor voltage due to N feedback connections to the upper synapse, and $M = F_L/F_{out}$, where F_L is the frequency of activation of the lower synapse. Since F_{out} appears as an argument in the log function, equation (2) is most easily solved numerically.

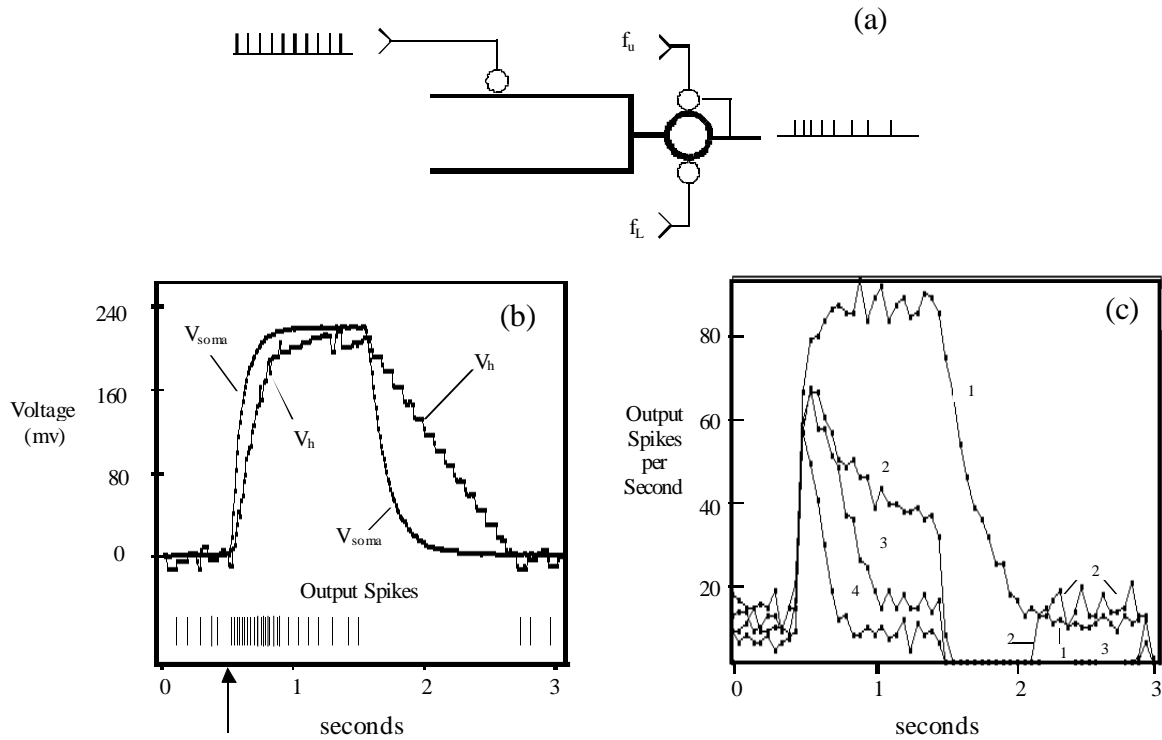


Figure 5.7. (a) Two-branched dendritic tree neuromorph. A spike train of one second duration is input to a proximal excitatory synapse on one branch. Tonic spike trains of frequency f_u and f_L are delivered to the upper and lower threshold-setting synapses along with various amounts of feedback to upper synapse from output. (b) Soma voltage (V_{soma}), and the threshold voltage (V_{th}) as a 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 one 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.

Operating neuromorphs with negative feedback connections brings other benefits. One is that it minimizes the effects of threshold variation between individual neuromorphs, which is especially noticeable between different chips. It also allows neuromorphs to translate increments in the soma potential, V_s , into increments of output spike frequency in a linear fashion; decrements in V_s , unless small, will cut off spike firing, thereby providing a thresholding non-linearity. These are typical neuronal response characteristics.

5.3.2 Discrimination of complex spatio-temporal patterns

To demonstrate the power of the dendritic tree as a processor of spike-encoded information, we built a small, feed-forward network of two layers of neuromorphs to discriminate spatio-temporal patterns that could, for example, represent the waveforms of spoken vowel sounds. The first layer consisted of five J-units and the second of three K-units. The aim was to have each of three input vowels fire a different output (K) neuromorph. Each vowel waveform was digitized into three concurrent spike trains as follows. A spike in a given train was generated at the moment that the waveform crossed a fixed threshold level designated for that train. It turned out to be easy to train the network to discriminate the three vowel sounds \a, \i and \u, probably because they were encoded by trains of different average spike frequencies. To pose a more challenging problem, therefore, we set the network to discriminate the input patterns of Fig. 5.8, each composed of three, 250 msec spike trains of the same frequency, differing only in their phase relationships. The three trains of each input pattern were distributed with delays to excitatory and inhibitory synapses on the dendrites of a first layer of neuromorphs, the J-layer. The J-unit thresholds were feedback regulated so that their somas fired spontaneously and operated in the "linear mode", whereas the somas of the K-units of the output layer operated with fixed thresholds that were high enough to ensure that they were normally silent.

The synaptic connections to the J-units and their associated delays were found by a search procedure so as to provide a partial discrimination of the input patterns. The connections onto the J-units were selected by searching through random combinations of synapse type (excitatory/inhibitory), synapse site, and transmission delay, evaluating each combination for its ability to discriminate between the three input patterns. (Only fixed weight, maximal strength synapses were available for these experiments.) Rejected were combinations that, during the 250-msec input, yielded less than 12 spikes to at least one pattern, or nearly equal numbers of spikes to each pattern. Saved for possible use were combinations that yielded at least a 75% difference in spike number between the most and the least effective input patterns. Screening of combinations was done in parallel by presenting each pattern simultaneously to five neuromorphs via different connections, allowing two combinations to be screened per second. About 2% of the combinations were useful.

Having picked the input-J connections to form "basis units", the patterns were fully discriminated by training connections to the K units, which functioned as a layer of linear threshold units. Each J unit made a single synaptic connection to the dendritic tree of each K unit. The synaptic inputs of the different J units were summed linearly by locating them on separate branches of the K dendritic trees. The J-K connections were readily trained by a Perceptron-type learning rule to make the designated K unit fire at least 10 spikes and the others as few spikes as possible. An error function, based on the total number of spikes fired during the 250 msec input period, was designed to change the effective J-K synaptic weights after each trial. Selecting an excitatory or inhibitory synapse set the sign of a weight, and moving it along the dendritic branch toward the soma increased its efficacy.

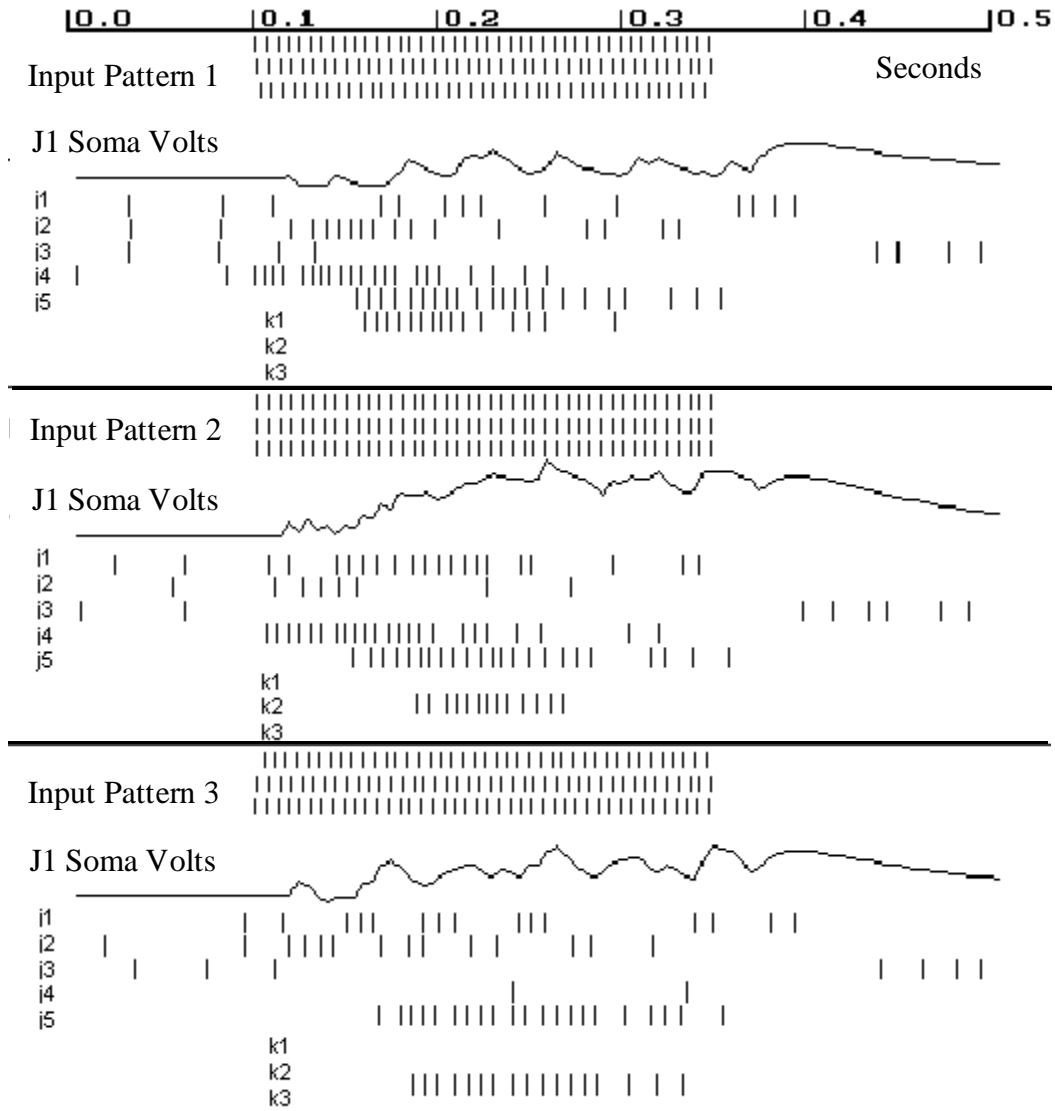


Figure 5.8. Temporal pattern discrimination. The three parts of this figure show the responses of all the neuromorphs in the network to different input patterns, each of which consists of three spike trains of the same frequency, differing only in their phase relations. "J1 Soma volts" shows an example of the potential waveforms generated at the soma. Spike trains labeled J1-J5 show the responses of all the J-units. Note the spontaneous activity before the stimulus onset. Spike trains labeled K1-K3 show the responses of the output units after training. Note that a different K unit fires to each input pattern.

As Figure 5.8 shows, the J-units, which fired spontaneously, responded to input trains with complex, temporally structured firing that depended upon the specific input pattern. It is noteworthy that much of the differentiation of the responses was due to latency differences. Because the K-layer works like a Perceptron, summing input spikes, it has to wait for several J-units to fire before it produces its

output. Consequently, the processing by this layer did not take advantage of the temporal structure of its inputs. If it were to do so, as in the following example, the speed of classification could be greatly increased.

5.3.3 Processing of temporally encoded information

Cerebral cortex performs complex discriminations very rapidly, making decisions based on only one or two spikes in about a 30 msec time window [Thorpe & Imbert, 1989; Rolls & Tovee, 1994]. For such high speed performance a recognition network needs to exploit the timing of individual spikes, and Maass [1997] (see also Chapter 2) has shown that neurons could do this by integrating the initial phases of post synaptic potentials to generate temporally coded output spikes. Here we demonstrate a neuromorph used in just this fashion.

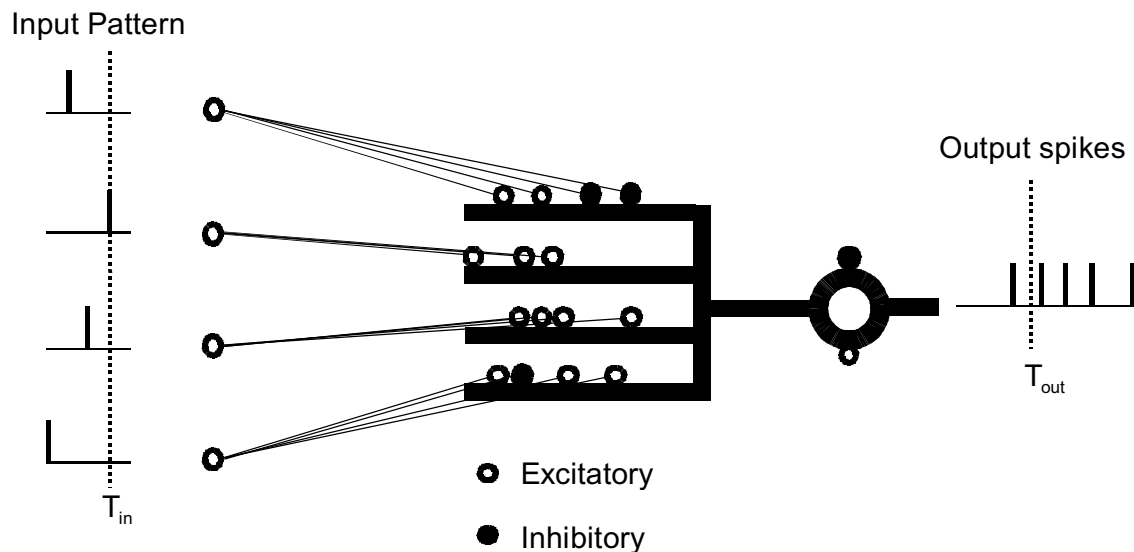


Figure 5.9. Processing spike-time coded inputs. An input pattern is represented by the time advance of four spikes relative to the reference time T_{in} . The input spikes are sent to excitatory and inhibitory synapses on separate branches of a 4-branch dendritic tree. The output value is encoded by the relative time advance of the first output spike generated by the soma relative to the reference time T_{out} .

Figure 5.9 shows four input units each sending a spike with zero delay to a set of synapses on one branch of a neuromorph, an arrangement that ensures that the PSPs generated on each branch are summed linearly at the soma. As in Maass's formulation, the input pattern vectors, with components x_j , are the relative time advances of the constituent spikes. Figure 5.10 shows three different input patterns coded by the time advances of their spikes relative to the reference time, T_{in} . Each input spike alone would generate a PSP at the soma with a rate of rise depending upon the net efficacy of all the synapses that that spike activates. For example, the synapses on the top branch of Fig. 5.9 would be net inhibitory because the inhibitory synapses are closer to the soma; those on the third branch would be strongly excitatory.

Each input spike, representing x_j by its timing, was accorded an effective weight, w_j , by looking-up in a table of synapse combinations – again, only full-strength synapses were available. The table of synapse combinations was ordered in terms of the maximum rate of rise of the PSPs that they generated at the soma when they were all activated simultaneously on one dendritic branch. Presenting a 4-spike input pattern generated a PSP at the soma (V_s in Fig. 5.10) representing the summation of each branch’s PSP. The average rate of rise of the potential, and therefore the time at which the neuromorph’s firing threshold was reached, depended upon the temporal ordering of the input spikes, and upon the weight accorded each spike. A supervised, Perceptron-type learning rule was used to adjust synaptic connections so as to generate a first output spike close to target latencies arbitrarily chosen for each input pattern. In the result shown in Figure 5.10, we used the target times of 0, 2 and 1 msec, measured as time-advances relative to a reference time, T_{out} . Training was performed by presenting all three input patterns and obtaining an error for each pattern p , $E_p = T_p - y_p$, where T_p is the target time and y_p is the first spike time, both measured as a time advance relative to T_{out} . As a practical matter, it was necessary to truncate E_p to ± 1.5 msec before use in the following formula for the change in weight for input spike j ,

$$\Delta w_j = \epsilon \sum_p E_p x_{jp}.$$

As Figure 5.10 shows, each input pattern generated a train of output spikes, the first of which occurred close to the target latencies. Achieving this performance was not easy because the timing of spikes produced by the soma is subject to some jitter, apparently caused by noise at the inputs to the spike generator. There was also the problem of selecting synapses so that their PSPs summed to the requisite rise time at the soma, given that a synapse’s PSP peak latency is inversely related to its peak amplitude. The use of variable conductance synapses will provide additional control over rise time. Nevertheless, this demonstration shows that a dendritic tree neuromorph can interpret information contained in a purely temporal code of minimal spike number. However, as we have seen in section 5.3.2, discrimination was more robust when input patterns were spike trains in which the information to be decoded lay in the time differences between multiple spikes.

5.4 Conclusions

Modeled as classical neurons with passive dendrites, our silicon neuromorphs have the dynamical properties to process spike trains much as the nervous system does. The artificial dendrites are able to sum PSPs linearly, or to exhibit a saturating non-linearity, depending on the timing and location of synaptic activations. The integrate-and-fire soma issues spikes in temporally patterned bursts and trains that not only look biologically realistic but are able to convey discriminative information for subsequent stages of neuromorphic processing [Northmore and Elias, 1997]. Real neurons elaborate upon the classical theme with a variety of voltage- and chemical-sensitive ion channels in their cell membranes, which expand their processing capabilities in ways that we have barely begun to appreciate. However, the process of building, even along classical lines, is beginning to teach us something about neural styles of computation.

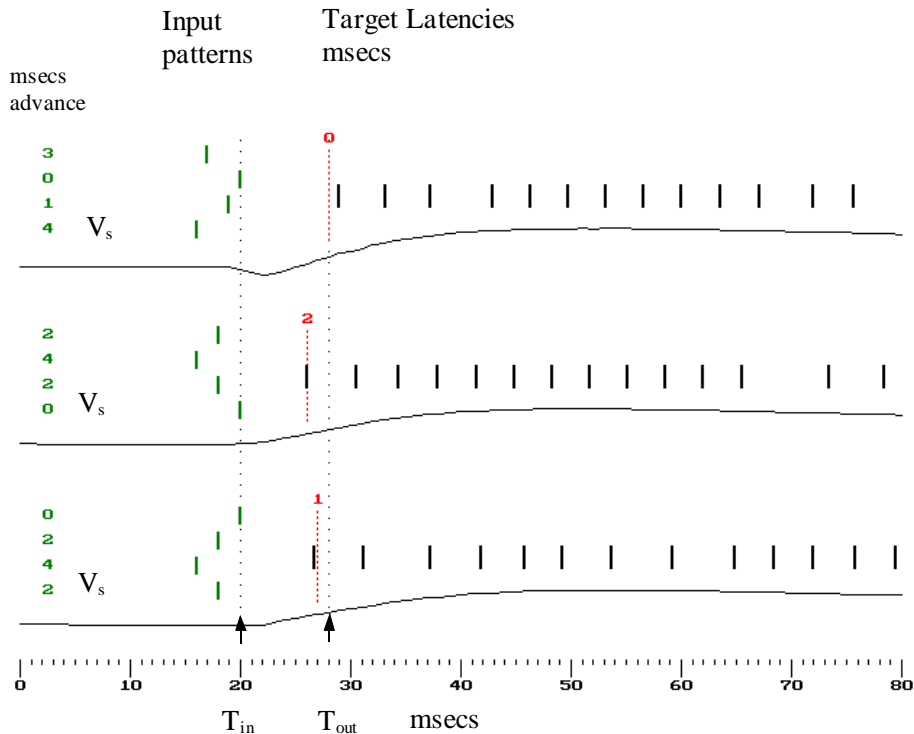


Figure 5.10. Processing three spike time-coded input patterns. Each input pattern is represented by the time advance of four spikes relative to T_{in} . The network of figure 5.9 was trained to generate a first output spike with a time advance relative to T_{out} shown by the dashed lines labelled "Target latencies". V_s is the potential appearing at the soma in response to the input spikes.

The incorporation of active channels into the artificial dendrite is a current aim, but it is feasible to emulate only a few of the known channel mechanisms if the dendrites are not to be burdened with circuit complexity. A first priority would be a channel that confers a "superlinearity" to the summation of synaptic inputs, such as the NMDA channel [Mel, 1993]. This channel is also implicated in long-term potentiation of synaptic efficacy, a mechanism for learning [Bliss & Collingridge, 1993].

Learning, in its various forms looms as a challenge to neuromorphic endeavors. A goal to strive for is activity-dependent modification with long-term setting of synaptic efficacy at multiple sites on a dendritic tree. We now know how to make dendritic trees; distributing long term storage over dendrites is more of a problem, but one that could be solved by existing technologies such as floating gates [Diori et al., 1995] or flux capacitors. At present, our approach to adding channel complexity and learning capability to artificial dendrites is one of circumspection. The understanding of how channel mechanisms contribute to the overall function of dendrites is at a primitive stage, while new experimental techniques are still yielding important discoveries (e.g. Markram et al., 1997). To explore how channel mechanisms in artificial dendrites might implement rules of learning we are currently taking a hardware-software hybrid approach before committing to silicon. Neuromorphs, with passive dendrites and integrate-and-fire somas, accept and generate spikes in real time, while a host computer samples compartment potentials and spiking activity. On the basis of these data, software executes learning rules

by changing synaptic weights, connection patterns, or global neuromorph parameters. In this fashion, the Virtual Wire system is being used to explore Hebbian association between real incoming spikes and simulated spikes back-propagated over the dendritic tree [Westerman et al., 1998].

Acknowledgments

Supported by NSF Grants BCS-9315879, and BEF-9511674. We thank Shawn Gallagher for help with the figures.

Bibliography

[Allen & Sanchez-Sinencio, 1984] Allen, P.E. and Sanchez-Sinencio, E. (1984) *Switched Capacitor Circuits*. Van Nostrand Reinhold Company, New York.

[Bernander et al., 1991] Bernander, O, Douglas, R.J., Martin, K.A.C. and Koch, C. (1991) Synaptic background activity influences spatiotemporal integration in single pyramidal cells. *Proc. Natl. Acad. Sci. USA* 88:11569-11573.

[Bliss & Collingridge, 1993] Bliss, T.V.P. and Collingridge G.L. (1993) A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* 361: 31-39.

[Cariani, 1994] Cariani, P. (1994) As if time really mattered: Temporal strategies for neural coding of sensory information. In *Origins: Brain and Self-organization*, Ed. K. Pribram. Lawrence Erlbaum, Hillsdale, NJ. Pp 208-252.

[Diori, et al., 1995] Diori, C., Hasler, P., Minch, B. and Mead, C. (1995) A high-resolution non-volatile analog memory cell. *Advances in Neural Information Processing Systems*, 7: 817-824.

[Eccles, 1957] Eccles, J.C. (1957) *The Physiology of Nerve Cells*. Johns Hopkins Univ. Press, Baltimore, Maryland.

[Elias, 1993] Elias, J.G. (1993) Artificial dendritic trees. *Neural Computation* 5: 648-663.

[Elias & Northmore, 1995] Elias, J.G. and Northmore, D.P.M. (1995) Switched-capacitor neuromorphs with wide-range variable dynamics. *IEEE Trans. Neural Networks* 6:1542-1548.

[Elias et al., 1997] Elias, J.G. and Northmore, D.P.M. and Westerman, W. (1997) An analog memory device for spiking silicon neurons. *Neural Computation* 9: 419-440.

[Ferster & Jagadeesh, 1992] Ferster, D. and Jagadeesh, B. (1992) EPSP-IPSP interactions in cat visual cortex studied with in vivo whole-cell patch recording. *J. Neuroscience* 12: 1262-1274.

[Gawne et al., 1991] Gawne, T.J., McClurkin, J.W., Richmond, B.J. and Optican, L.M. (1991) Lateral geniculate neurons in behaving primates, III. response predictions of a channel model with multiple spatial filters. *J. Neurophysiol.* 66:809-823.

[Geisler et al. 1991] Geisler, W.S., Albrecht, D.G., Salvi, R.J. and Saunders, S.S. (1991) Discrimination performance of single neurons: Rate and temporal-pattern information. *J. Neurophysiol.* 66: 334-362.

[Gerstner et al., 1996] Gerstner, W., Kempter, R., van Hemmen, J.L. and Wagner, H. (1996) A neuronal learning rule for sub-millisecond temporal coding. *Nature* 383: 76-78.

- [Hille, 1992] Hille, B. (1992) *Ionic Channels of Excitable Membranes*. Sinauer, Sunderland, Massachusetts.
- [Jack et al, 1975] Jack, J.J.B., Noble, D. & Tsien, R.W. (1975) *Electric Current Flow in Excitable Cells*. Oxford University Press, London.
- [Knight, 1972] Knight, B. (1972) Dynamics of encoding in a population of neurons. *J. General Physiology* 59:734–766.
- [Koch, 1997] Koch, C. (1997) Computation and the single neuron. *Nature* 385: 207–210.
- [Koch & Poggio, 1987] Koch, C. and Poggio, T. (1987) Biophysics of computation: neurons, synapses, and membranes. In *Synaptic Function*, ed G. Edelman, W. Gall, & W. Cowan. Wiley-Liss, New York, Pp 637–697.
- [Maass, 1997] Maass, W. (1997) Fast sigmoidal networks via spiking neurons. *Neural Computation* 9: 279–304.
- [Maes, 1993] Maes, P. (1993) Behavior-based artificial intelligence. In *From Animals to Animats 2*. Eds. Meyer, J., Roitblat, H., and Wilson, S. MIT Press, Cambridge, Massachusetts.
- [Mahowald, 1992] Mahowald, M.A. (1992) Evolving analog VLSI neurons. In *Single Neuron Computation*, Eds: T. McKenna, J. Davis, S. Zornetzer. Academic Press, San Diego, California.
- [Markram et al., 1997] Markram, H., Lhbke, J., Frotscher, M. and Sakman, B. (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275: 213–215.
- [McClurkin et al., 1991] McClurkin, J.W., Gawne, T.J., Optican, L.M. and Richmond, B.J. (1991) Lateral geniculate neurons in behaving primates. II. Encoding of visual information in the temporal shape of the response. *J Neurophysiol.* 66: 794–808.
- [McCormick et al., 1992] 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. Zornetzer. Academic Press, San Diego, California.
- [Mel, 1993] Mel, B.W. (1993) Synaptic integration in an excitable dendritic tree. *J. Neurophysiol.* 70: 1086–1101.
- [Mel, 1994] Mel, B.W. (1994) Information processing in dendritic trees. *Neural Computation* 6: 1031–1085.
- [Northmore & Elias, 1993] Northmore, D.P.M. and Elias, J.G. (1993) Directionally selective artificial dendritic trees. In *Proceedings of World Congress on Neural Networks, vol IV*, Lawrence Erlbaum Associates, Hillsdale, NJ, Pp 503–508.
- [Northmore & Elias, 1996] Northmore, D.P.M. and Elias, J.G. (1996) Spike train processing by a silicon neuromorph: The role of sublinear summation in dendrites. *Neural Computation* 8:1245–1265.
- [Northmore & Elias, 1997] Northmore, D.P.M. and Elias, J.G. (1997) Discrimination of phase-coded

spike trains by silicon neurons with artificial dendritic trees. In *Computational Neuroscience: Trends in Research 1997*. Ed. J.M. Bower, Plenum Press, New York. Pp 153–157.

[Optican & Richmond, 1987] Optican, L.M. and Richmond, B.J. (1987) Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *J Neurophysiol.* 57:162–178.

[Rall, 1964] Rall, W. (1964) Theoretical significance of dendritic trees for neuronal input–output relations. In *Neural Theory and Modeling*. Ed. R.F. Reiss, Stanford University Press, Pp 73–79.

[Rieke et al., 1997] Rieke, F., Warland, D., de Ruyter van Steveninck, R. Bialek, W. (1997) *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, Massachusetts.

[Rolls & Tovee, 1994] Rolls, E.T. and Tovee, M.J. (1994) Processing speed in the cerebral cortex and the neurophysiology of visual masking. *Proc Roy. Soc. (Lond.) B.* 257:9–15.

[Rose & Call, 1993] Rose, G.J. and Call, S.J. (1993) Temporal filtering properties of midbrain neurons in an electric fish: implications for the function of dendritic spines. *J. Neurosci.* 13:1178–1189.

[Steels & Brooks, 1995] Steels, L. and Brooks, R. (1995) *The Artificial Life Route to Artificial Intelligence: Building Embodied, Situated Agents*. Lawrence Erlbaum Associates, Hillsdale, NJ.

[Singer, 1995] Singer, W. (1995) Synchronization of neuronal responses as a putative binding mechanism. In *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib, ed. MIT press, Cambridge, Massachusetts. Pp 960–964.

[Thorpe & Imbert, 1989] Thorpe, S.J. and Imbert, M. (1989) Biological constraints on connectionist models. In *Connectionism in Perspective* (ed. R.Pfeifer, Z. Schreie & F. Fogelman–Soulie). Pp 63–92. Elsevier, Amsterdam.

[Victor & Purpura, 1996] Victor, J.D. and Purpura, K.P. (1996) Nature and precision of temporal coding in visual cortex: A metric–space analysis. *J. Neurophysiol.* 76, 1310–1326.

[Westerman et al., 1997] Westerman, W., Northmore, D.P.M. and Elias, J.G. (1997) Neuromorphic synapses for artificial dendrites. *Analog integrated circuits and signal processing*, 13, 167–184.

[Westerman et al., 1998] Westerman, W., Northmore, D.P.M. and Elias, J.G. (1998) Antidromic spikes drive Hebbian learning in an artificial dendritic tree. *Analog integrated circuits and signal processing*. In press.