SILICON SPIKE-BASED SYNAPTIC ARRAY AND ADDRESS-EVENT TRANSCEIVER

R. Jacob Vogelstein[1]*, Udayan Mallik*[2]*, and Gert Cauwenberghs*[2]

[1]Department of Biomedical Engineering
[2]Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore, Maryland 21218

ABSTRACT

An integrated array of 2,400 spiking silicon neurons, with reconfigurable synaptic connectivity and adjustable neural spike-based dynamics, is presented. At the system level, the chip serves as an address-event transceiver, with incoming and outgoing spikes communicated over an asynchronous event-driven bus. Internally, every cell implements a spiking neuron that models general principles of synaptic operation as observed in biological membranes. Synaptic conductance and synaptic reversal potential can be dynamically modulated for each event. The implementation employs mixed-signal charge-based circuits to facilitate digital control of system parameters and minimize variability due to transistor mismatch. In addition to describing the structure of the silicon neurons, we present experimental data characterizing the operation of the 3mm $\times$ 3mm chip fabricated in 0.5μm CMOS technology.

1. INTRODUCTION

An increasing number of experimental microchip designs are inspired by biological structures, particularly those found in the vertebrate and invertebrate nervous systems. Neuromorphic systems engineering [1, 2] emulates both function and structure of biological neural systems in silicon, and correspondingly achieves high levels of efficiency in the implementation of sensory systems for vision [3] and audition [4]. The complexity of neural computation, beyond sensory perception, requires a multi-chip approach and a proper communication protocol between chips to implement higher levels of processing and cognition.

The common language of neuromorphic chips is the "Address-Event Representation" (AER) communication protocol [5], which uses time-multiplexing to emulate extensive connectivity between neurons. In its original formulation, AER implements a one-to-one connection topology; to create more complex neural circuits, convergent and divergent connectivity is required. The AER framework has proved essential in enhancing the connectivity between multi-chip neuromorphic modules [6, 7, 8, 9, 10, 11]. AER "transceivers" [6, 10, 11] call for a memory-based projective field mapping that enables routing an address-event to multiple receiver locations. Accordingly, the chip described in this paper contains 2,400 neurons but no hardwired connections between cells, rather depending on an external infrastructure to route events to their appropriate targets.

There are a few examples of reconfigurable neural array transceivers in the literature [10, 11], but the one presented here differs in some important aspects. This paper concentrates on the design of the neural cell, which has two novel attributes. First, the silicon neuron implements a discrete-time version of the single compartment, conductance-based *membrane equation*—a standard model describing current flux through biological neural membranes—which enables more sophisticated simulations than a standard integrate-and-fire model allows. Second, this design permits an unlimited number of connections between neurons, with independent control of connection strength and synaptic reversal potential on a per-connection basis.

This work is partially funded by NSF Award #0120369 and ONR Award #N00014-99-1-0612. RJV is supported by an NSF Graduate Research Fellowship.

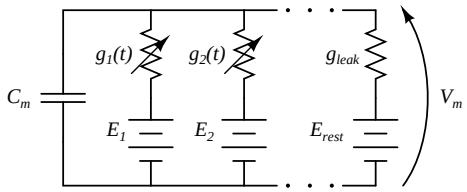


Fig. 1. Single-compartment model of a biological neuron with multiple synapses and a static leak conductance.

2. NEURAL MODEL

A number of silicon neurons have been presented in the literature with varying degrees of biological accuracy. The most detailed and accurate silicon models feature many parameters and are very flexible [12], but occupy a large on-chip area and therefore limit the number of cells that can be fabricated on a single chip. The simplest models contain only a few transistors and are well-suited for implementation in a large-scale network, but deviate significantly from the biology and have few adjustable parameters. Many applications would benefit from a balance between these two extremes: a more biologically accurate neural model allows for more sophisticated simulations of cognitive functions, but only in the context of a sophisticated network architecture. We have therefore designed a small-footprint, highly configurable, "general-purpose" silicon neuron that implements a standard model of biological neural membranes.

A popular model used in computational neuroscience to describe the current flux through biological neural membranes is the single-compartment model (Fig. 1).

The model is specified by the membrane equation:

$$C_m \frac{dV_m}{dt} = g_1(t) \cdot (E_1 - V_m) + g_2(t) \cdot (E_2 - V_m) + \cdots + g_{leak} \cdot (E_{rest} - V_m) \quad (1)$$

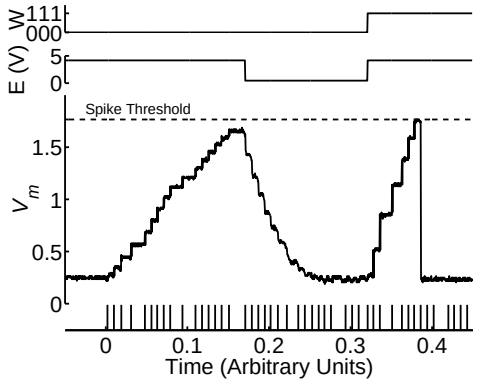


Fig. 2. Data captured from an oscilloscope during operation of the chip. The lower trace illustrates the membrane potential (V_m) of a single neuron in the array as a series of events are sent at times marked at the bottom of the figure. The synaptic reversal potential (E) and synaptic weight (W) are drawn in the top two traces.

where C_m is a large membrane capacitance, V_m is the membrane potential, $g_i(t)$ is a time-varying synaptic conductance, and E_i is a synaptic reversal potential. Although biology operates in continuous time, most neural interactions occur on the millisecond time scale, so it is possible to simulate the internal dynamics of a neuron using fast, discrete time steps. Similarly, while multiple synaptic inputs can be active simultaneously in a real neuron, it is essentially equivalent to activate a group of synapses in rapid succession due to biology's low precision in the time domain. We exploit both of these observations in the design of the silicon neuron.

The neural cell schematic is shown in Figure 3, along with event generation circuitry to trigger and communicate output spikes (see Sec. 3). The cell size, including the event generation circuitry, is 40μm $\times$ 60μm. Using a simple switched-capacitor architecture, this circuit implements a discrete-time version of the membrane equation:

$$C_m \frac{\Delta V_m}{\Delta T} = g(T) \cdot (E - V_m) \tag{2}$$

By sequentially activating switches X1 and X2, a packet of charge proportional to the externally supplied and dynamically modulated reversal potential E is added to (or subtracted from) the membrane capacitor C_m. The amount of charge transferred is also dependent on which of the three binary-sized "synaptic weight" capacitors (C0-C2) are enabled: these elements are dynamically switched on and off by applying voltage to the gates of transistors M1-M3. When sufficient charge has been integrated on C_m to cause V_m to exceed a "spike threshold", the neuron generates an event (see Sec. 3). Figure 2 illustrates the functionality of one neuron in the array as it receives a sequence of events with both the synaptic reversal potential and the synaptic weight dynamically varied.

There are a few advantages of this architecture over previous AER transceiver designs [11]. First, it allows for simulation of an unlimited number of synapses on every cell, as each incoming event can be assigned a unique weight and reversal potential. Second, it simulates biologically realistic conductance-based synapses. The use of conductance-based synapses in a neural model can have important implications: unlike standard integrate-and-fire models, the order of events in a conductance-based model is an essential factor in determining the neural output (Fig. 2). Third, charge-based circuits exploit better matching between capacitors than between MOS transistors due to threshold variations, which results in greater uniformity of operation across the chip. Finally, there is very little charge leakage off the membrane capacitor, allowing for large dynamic range in the implementation of neural dynamics on various time scales. Since neural integration is discrete-time, it is also possible to decouple event timing from emulated time, and dynamically warp the time axis [13].

3. EVENT GENERATION

Information encoded by neurons in the array is represented by the time between successive events. Therefore, event generation and communication is an essential element of the design. The event generation circuitry of [11] is embedded in every cell (Fig. 3, right). An event—signaled by a low voltage on $\overline{R_{req}}$—should be generated each time a neuron's membrane potential exceeds the spike threshold. In the circuit, charge integrated on the membrane capacitor (C_m) of a cell (see Sec. 2) results in an increase in potential applied to the gate of M4, the input terminal of a comparator. Since V_m can rise very slowly, the comparator is implemented as a current-starved inverter, with M5 biased in weak inversion, for reduced power dissipation. The spike threshold is set by V_{thresh}, the voltage applied to the source of M4; this value is shared by all cells in the array and is externally controlled. The corresponding input-referred threshold is approximately equal to $V_{thresh} + V_{T_n}$, where V_{T_n} is the threshold voltage of M4. When V_m exceeds this value, a positive-feedback loop implemented by transistors M6-M8 is activated, triggering a spike event by driving V_m to the positive rail and V_{comp} to ground.

A high voltage on M15 activates $\overline{R_{req}}$, the output node of a row-wise wired-NAND, and indicates to the row arbitration circuitry that a cell in that row has generated an event and needs to be serviced. Until this occurs, the row and column acknowledge signals R_{ack} and C_{ack} will remain low, maintaining the positive feedback loop and preventing any further inputs from affecting the cell. The row arbitration circuity indicates it has selected a row by driving one pair of R_{scan} and R_{ack} signals high. All cells in that row with pending events will then pull their $\overline{C_{req}}$ signals low, indicating to the column arbitration circuitry that they have generated events and need to be serviced. Finally, the column arbitration circuitry indicates which column it has selected by driving one column's C_{req} signal high. At that point, both R_{ack} and C_{ack} are asserted (for one cell only) so the positive-feedback loop is inactivated and the reset circuit implemented by nMOS transistors M9 and M10 causes V_m to become V_{reset} (like V_{thresh}, V_{reset} is shared by all cells in the array and is externally controlled). As V_m drops below the comparator's threshold voltage, V_{comp} is pulled high by M5 and the column and row requests ($\overline{C_{req}}$ and $\overline{R_{req}}$) are removed. This completes the handshaking sequence between a cell and the arbitration circuitry.

4. EXPERIMENTAL RESULTS

Every incoming event is routed to one or more neurons and is associated with a particular binary weight and an analog reversal potential (E). Additionally, a spike threshold voltage (V_{thresh}) and resting potential (V_{reset}) are set globally for the entire chip. To quantify neurons' dependence on these parameters, we have performed three experiments. First, to determine the effect of the

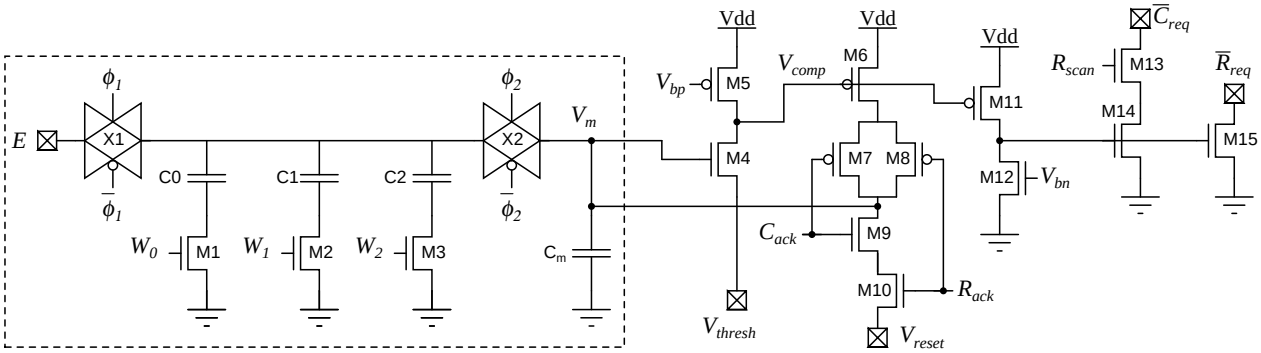


Fig. 3. Silicon single-compartment neuron (inside dashed box), with event generation circuitry (shown right, [11]).

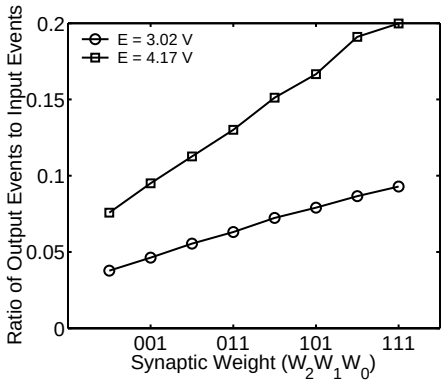


Fig. 4. Average ratio of output events to input events versus synaptic weight. Data are averaged across 10 trials per cell at each value of synaptic weight, and then averaged over all 2,400 cells in the array. The weight capacitors C_0, C_1, and C_2 are designed to achieve (ideally) the size ratio 1:2:4.

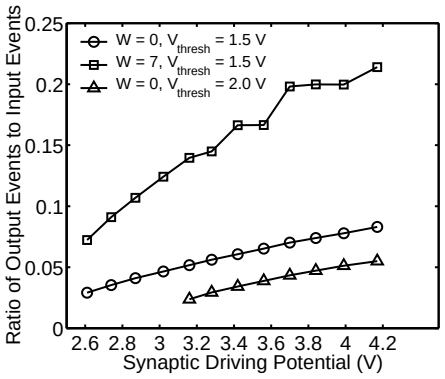


Fig. 5. Average ratio of output events to input events versus synaptic reversal potential (E). Data are averaged across 10 trials per cell at each value of synaptic weight, and then averaged across all 2,400 cells in the array.

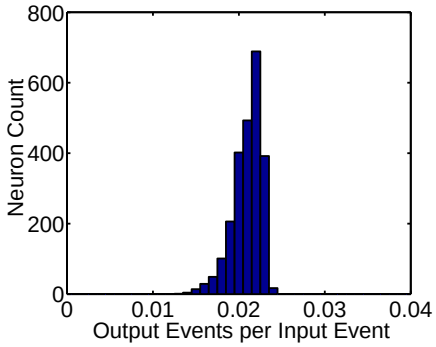
weight capacitors C0-C2 (see Fig. 3), each neuron's membrane potential was reset to a fixed voltage and a series of excitatory events at a fixed reversal potential and a given synaptic weight were sent to the cell. The number of events required to elicit a spike was recorded and after ten measurements at each synaptic weight, another cell in the array was tested. The results over all 2,400 cells were summarized by plotting the average ratio of output events to input events versus synaptic weight (Fig. 4). The design called for a greater slope for the lines in Figure 4, but this was limited by a large parasitic capacitance when a weight capacitor was in the "off" state.

The second experiment was designed to quantify the effect of the synaptic reversal potential (E). Here, each neuron's membrane potential was reset to a fixed voltage and a series of events at a given excitatory synaptic reversal potential and a fixed synaptic weight were sent to the cell. Again, the number of events required to elicit a spike was recorded. However, instead of varying the weight (as in the first experiment), the same cell was then re-tested with a different value of E. The results over all 2,400 cells were summarized by plotting the average ratio of output events to input events versus synaptic reversal potential (Fig. 5). Although in normal operation the spike threshold voltage V_{thresh} is likely to be fixed, in some cases it may be desirable to dynamically vary this parameter. Therefore, we repeated the experiment described above using two different values of V_{thresh} (Fig. 5).

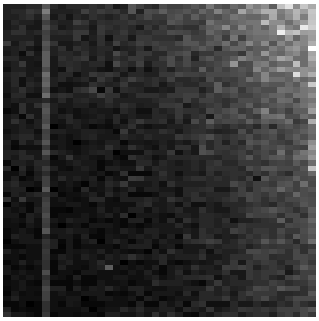
As discussed in Section 2, one of the advantages of implementing the neurons in charge mode is that it minimizes variability across the chip. To quantify the mismatch between neurons, we conducted a third experiment wherein all of the event parameters (synaptic weight, synaptic reversal potential, spike threshold voltage, and resting potential) were held constant at values that were barely supra-threshold. The average ratio of the number of output events to input events was measured as before, and the distribution was plotted as a histogram (Fig. 6a). The variability is small, with a standard deviation of 0.0017 around the mean of 0.0210. To see if there was any systematic variation, the number of input events required to elicit an output event was converted into a normalized gray-scale value and the value for each neuron was plotted as a single pixel in a 60×40 bitmap, where darker pixels correspond to a larger number of output events per input events (Fig. 6b). Evidently, there is a gradient toward lower response rates in the upper-right quadrant of the array.

5. CONCLUSION

We have presented a novel neural array transceiver consisting of 2,400 spiking neurons that each implement a discrete-time version



(a)



(b)

Fig. 6. (a) Histogram showing distribution of the average ratio of output events to input events. (b) Bitmap illustrating spatial trends in the variation of the average ratio of output events to input events (see text for details).

of the biological membrane equation and include a single "general-purpose" synapse with dynamically configurable weight and reversal potential. The data shown verify the functionality of the chip. Future work will focus on embedding multiple chips in a large system containing digital memory to store connection patterns and synaptic parameters, a microcontroller to manage a shared AER bus, and a high-speed interface with a neuromorphic sensor or personal computer. A previous version of this system was used for a variety of applications, including implementing Laplacian filters to isolate vertical edges on static images, a task that ran two orders of magnitude faster in hardware than in simulation [11]. The new system will permit much larger networks and will operate approximately 100 times faster than the old hardware, to allow rapid prototyping and real-time emulation of realistic neural circuits and, ultimately, interface with real biological neurons in computer-controlled wetware experiments. An interesting extension of the functionality of these networks is to incorporate silicon models of spike-based learning [14] in the address-domain [13].

6. REFERENCES

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

[2] T.S. Lande, Ed., *Neuromorphic Systems Engineering— Neural Networks in Silicon*, Kluwer Academic, Norwell, MA, 1998.

[3] C. Koch and H. Li, Eds., *Vision Chips: Implementing Vision Algorithms with Analog VLSI Circuits*, IEEE Computer Press, 1995.

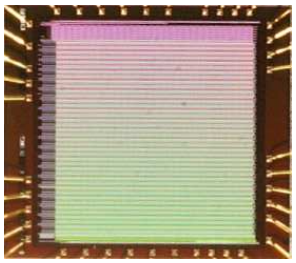


Fig. 7. Chip micrograph. Center: 60×40 neuron array. Periphery: row and column address-event decoders and arbitrating encoders.

[4] A. van Schaik, E. Fragniere, and E. Vittoz, "Improved silicon cochlea using compatible lateral bipolar transistors," in *Adv. Neural Info. Proc. Sys. 8*, D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, Eds., pp. 671–677. MIT Press, Cambridge, MA, 1996.

[5] M. Mahowald, *An analog VLSI system for stereoscopic vision*, Kluwer Academic Publishers, Boston, MA, 1994.

[6] K. A. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Trans. Circ. Sys.-II*, vol. 47, no. 5, pp. 416–434, 2000.

[7] C. M. Higgins and C. Koch, "Multi-chip neuromorphic motion processing," in *Proc. 20th Ann. Conf. Adv. Res. VLSI*, D. S. Wills and S. P. DeWeerth, Eds., Los Alamitos, CA, 1999, pp. 309–323, IEEE Computer Society.

[8] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbrück, and R. Douglas, "Orientation-selective aVLSI spiking neurons," in *Adv. Neural Info. Proc. Sys. 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, Cambridge, MA, 2002.

[9] N. Kumar, W. Himmelbauer, G. Cauwenberghs, and A. Andreou, "An analog VLSI chip with asynchronous interface for auditory feature extraction," *IEEE Trans. Circ. Sys.-II*, vol. 45, no. 5, pp. 600–606, 1998.

[10] G. Indiveri, A. M. Whatley, and J. Kramer, "A reconfigurable neuromorphic VLSI multi-chip system applied to visual motion computation," *Proc. MicroNeuro'99*, Apr. 1999.

[11] D. H. Goldberg, G. Cauwenberghs, and A. G. Andreou, "Probabilistic synaptic weighting in a reconfigurable network of VLSI integrate-and-fire neurons," *Neural Networks*, vol. 14, no. 6-7, pp. 781–793, 2001.

[12] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, pp. 515–518, 1991.

[13] R. J. Vogelstein, F. Tenore, R. Philipp, M. S. Adlerstein, D. H. Goldberg, and G. Cauwenberghs, "Spike timing-dependent plasticity in the address domain," in *Adv. Neural Info. Proc. Sys. 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, Cambridge, MA, 2003.

[14] P. Hafliger and M. Mahowald, "Spike based normalizing Hebbian learning in an analog VLSI artificial neuron," in *Learning On Silicon*, G. Cauwenberghs and M. A. Bayoumi, Eds., pp. 131–142. Kluwer Academic, Norwell, MA, 1999.