Active Memristor Neurons for Neuromorphic Computing

Wei Yi
Sensors and Electronics Laboratory
HRL Laboratories, LLC.
Malibu, CA 90265-4797
Email: wyi@hrl.com

Abstract*—To rival mammals' intelligence, future artificial intelligence (AI) needs to be both strong – capable of self-learning with raw inputs, and broad – capable of handling multiple contexts rather than narrowly restricted problems. It is unlikely that AI implemented on conventional computing platforms will eventually achieve both goals. This is because even if the brain's connectivity is fully understood and reproduced, the neuron and synapse computing primitives, if built with non-biomimetic CMOS circuits, are not capable to emulate the rich dynamics of biological counterparts without sacrificing the energy consumption and size. Memristors may provide an alternative approach to realize strong and broad AI. The nonvolatile, stochastic and adaptive switching dynamics make the passive memristor a close electronic analogue to biological synapses. Owing to their superb scalability, memristor crossbars can potentially reach the synapse density of the brain. A complementary device, active memristor, can be used to construct electronic equivalent of the biological neurons. Active memristors show volatile resistive switching and are locally active within a hysteretic negative differential resistance (NDR) regime in current-voltage characteristics. The NDR regime endows active memristors a.c. signal gain, a must have for neural signal processing. Previous demonstrations of neuron circuits based on active memristors only showed basic spiking behaviors exhibited by leaky integrate-and-fire (LIF) models, and thus offer limited computational complexity. Using scalable VO_2 active memristors, we show that memristor neurons actually possess most of the known biological neuronal dynamics and all three classes of neuron excitability. Furthermore, our simulations show that the size and power scaling of memristor neurons project toward biologically competitive neuron density and energy efficiency. Our work indicates that scalable and biomimetic active memristor neurons, in combination with passive memristor synapses, will provide a self-sufficient path toward biologically competitive neuromorphic hardware that could disrupt the paradigm of data analysis and sensemaking.

Keywords—memristor; neuromorphic computing; neuron; synapse; biomimetic; vanadium dioxide; artificial intelligence

I. Introduction

At the building block level, today's neuromorphic hardware almost exclusively use silicon CMOS transistors to emulate certain aspects of neuron and synapse dynamics. However, as depicted in Fig. 1, significant gaps in energy efficiency (EE), size, and biological fidelity exist between today's state-of-the-art (SOA) Si neurons [1–9] and biological neurons. These gaps are not likely to be closed even after Moore's law expires. Silicon neurons are built with non-biomimetic CMOS transistors and suffer from the so-called cost-fidelity dilemma, i.e., scalability and biological fidelity are not simultaneously achievable. High-fidelity Si neurons (data points near bottom left) are not scalable. Fig. 1 also shows the trend of higher EE in smaller biological neurons in simulated Hodgkin-Huxley (HH) cells [10]. A record EE of 0.4 pJ/spike was achieved in an analog silicon neuron [5], but it is based on an LIF model that is known to show only 3 out of 20+ neuro-computational properties [11], and thus offer limited computational capability. For synapses, Silicon CMOS does not offer a scalable path to achieve the brain's synapse density (10^{10}/cm^2) or synapse-neuron ratio (>1,000:1). At the network or system level, neuromorphic

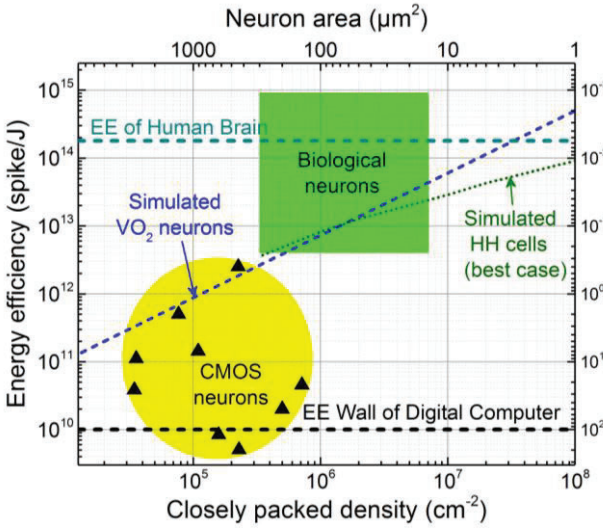

Fig. 1. Comparison of neuron energy efficiency (EE, in spike/J) versus neuron area. Data points (triangles) are from CMOS neurons published in 2008-2014. The estimated domain of biological neurons is outlined by the green rectangle. Green dotted line represents the best-case simulated Hodgkin-Huxley (HH) cells at ion channel density of 0.5/μm^2. Blue dashed line represents simulated VO_2 active memristor neurons at specific membrane capacitance of 1 fF/μm^2. VO_2 neurons show superior EE-area scaling than the best-case HH cells at neuron sizes smaller than 70 μm^2, and can surpass the estimated human brain EE (horizontal dark cyan dashed line) of 1.8×10^{14} spike/J (or 5.6 fJ/spike energy use) at neuron sizes smaller than 3 μm^2. Black dashed line is the conceived fundamental limit of EE for digital computers.

hardware utilize sparse coding in spike domain to excel in EE. Taking deep-learning inference for example, by mapping a convolutional neural network (CNN) into a neuromorphic spike-domain processor (TrueNorth), a record EE of 6.7 nJ/bit in image classification was achieved [12]. However, the realized chip-level throughput was only ~30.7 Mbit/s, far below mainstream graphic processing units (GPUs). The throughput of a silicon neuromorphic processor is likely limited by the poor scalability of Si neurons and synapses that handicaps the network scale and connectivity and hence the performance. A new paradigm in neuromorphic hardware is needed to break the dilemmas at both the building block and the system levels.

Fig. 2 shows the basic building blocks for a self-sufficient all-memristor neural network as an alternative approach for neuromorphic computing. The memristor is the 4th basic passive 1-port (two-terminal) circuit element and exhibits a pinched hysteresis I-V characteristic and nonvolatile memory [13]. The nonvolatile, stochastic and adaptive nature of its switching dynamics make the passive memristor a close electronic analogue to biological synapses. Owing to the superb $4F^2/N$ (F: half pitch, N: number of layers) scalability and stackability, memristor crossbar networks can potentially reach the synapse density of the brain with a single layer crossbar array at 100-nm pitch. A complementary device, active memristor, provide a biomimetic and scalable approach to construct electronic neurons. Active memristors are hallmarked by a locally-active hysteretic NDR regime in the otherwise volatile switching I-V. Among various mechanisms, Mott insulator-to-metal transition induced NDR can be realized at a switching energy less than 1 fJ and switching time faster than 1 ps in VO_2 material [14]. It can provide signal gain in a.c. domain, which can be exploited to create compact circuits to emulate the action potential (spike) generation in real neurons. Active memristors have the same $4F^2/N$ scalability and stackability as passive memristors. The biomimeticity lies in the fact that one can use two active memristors to form direct electronic analogues of Na^+ and K^+ membrane ion channels that mediate neuron action potential generation. Active memristor based spiking neurons, either emulating a single or two coupled membrane ion channels, have been demonstrated previously [15–17], but these prior arts did not demonstrate the full biological neuronal dynamics, or realized all three classes of neuron excitability.

Reference [14], for the first time, showed that active memristor neurons possess most of the known biological neuronal dynamics. 23 types of biological neuronal behaviors are experimentally demonstrated, including tonic spiking and bursting, phasic spiking (class 3 excitability) and bursting, mixed-mode spiking, spike frequency adaptation, class 1 and class 2 excitabilities, spike latency, subthreshold oscillations, integrator, resonator, rebound spike and burst, threshold variability, bistability, depolarizing after-potential, accommodation, inhibition-induced spiking and bursting, all-or-nothing firing, refractory period, and excitation block. The built-in stochasticity in active memristor neurons was shown by stochastic phase-locked firing, aka skipping phenomenon. SPICE simulations in [14] showed that the dynamic and static power scaling of active memristor neurons project toward biologically competitive neuron density and EE. In Fig. 1, a

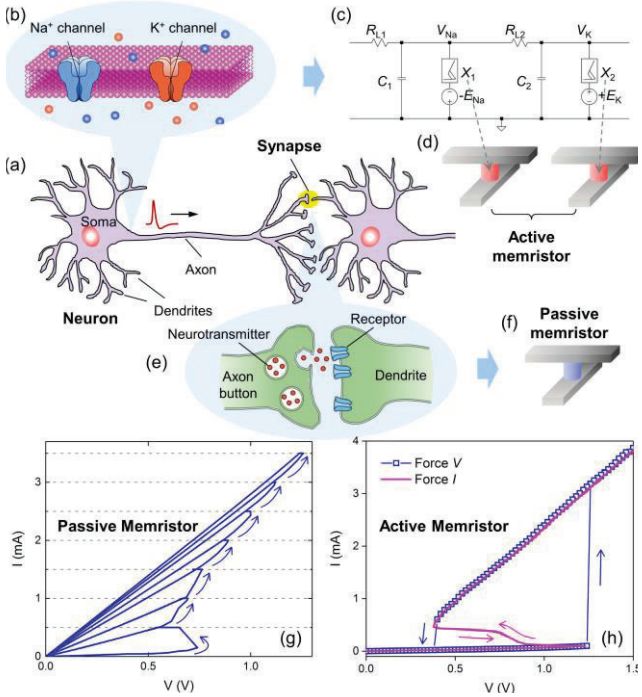

Fig. 2. Building blocks of a scalable and biomimetic all-memristor neural network. A biological neuron in (a), or more accurately its spiking membrane in (b), is emulated by a propotype circuit in (c), mimicking the voltage-gated Na^+ (K^+) ion channels on the cell membrane. Each of them is emulated by a negatively (positively) d.c. biased nanoscale active memristor device X_1 (X_2) in (d), which is closely coupled with a local membrane capacitor C_1 (C_2) and a series load resistor R_{L1} (R_{L2}), forming two resistively-coupled Pearson-Anson relaxation oscillators. Spiking mechanism of the circuit in (c) is explained in detail in [14]. A synapse in (e) is emulated by one or more nanoscale passive memristor devices in (f). Typical switching I-V of passive and active memristors are shown in (g) and (h), respectively.

simulated VO_2 neuron EE-area scaling at specific membrane capacitance of 1 fF/μm^2 projects a 5.6 fJ/spike energy use at neuron area of 3 μm^2, matching the EE of human brain.

II. Circuit Topology and Spiking Behaviors

Fig. 3 shows the three VO_2 active memristor prototype neuron circuits, tonic neuron, phasic neuron, and mix-mode neuron, and their characteristic spiking behaviors measured experimentally. Both tonic and phasic neurons play important roles in the central nervous system. Therefore it is important to be able to emulate both types of neurons using scalable active memristor circuits. Tonic neurons fire a continuous train of evenly spaced spikes (or periodic bursts of spikes) when stimulated by a steady d.c. current input. Phasic neurons, on the contrary, may fire only a single spike at the onset of the steady d.c. current input due to transient dynamics, and remain quiescent afterwards as the system reaches the steady state. Phasic firing is also known as Class 3 excitability. Before [14], there was yet no prior demonstration of phasic spiking in active memristor neurons. We found that phasic neurons can be realized simply by replacing the load resistor R_{L1} with a capacitor C_{in} (or by inserting a capacitor C_{in} before R_{L1}) in the tonic neuron circuit. In other words, replacing the resistively-

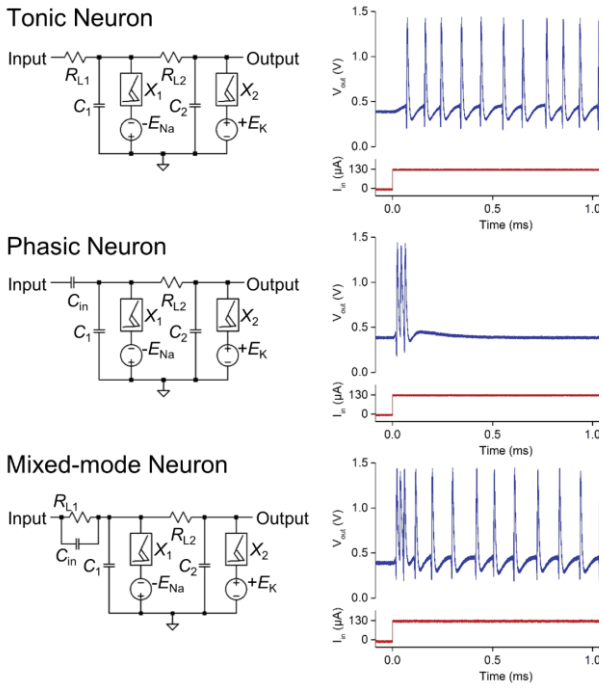

Fig. 3. Circuit diagrams and typical spike waveforms of three prototypes of active-memristor neuron circuits (from top to bottom): tonic neuron, phasic neuron, and mixed-mode neuron.

coupled dendritic input in a tonic neuron with a capacitively-coupled dendritic input turns it into a phasic neuron. Other than the input stage, both types of neurons share the same circuit topology. If a capacitor C_{in} is placed in parallel with the load resistor R_{L1}, the circuit becomes a mixed-mode neuron, and fires a phasic burst followed by a train of tonic spiking when stimulated by a steady d.c. current input.

As detailed in [14], we found that four types of basic biological neuron spiking behaviors are shared properties of both tonic and phasic neurons, including all-or-nothing firing, refractory period, spike frequency adaptation, and spike latency. In tonic neurons, ten unique spiking behaviors are observed, including tonic spiking (see Fig. 3), tonic bursting, Class 1 excitability, Class 2 excitability, subthreshold oscillations, integrator, bistability, inhibition-induced spiking, inhibition-induced bursting, and excitation block. In phasic neurons, eight unique spiking behaviors are observed, including phasic spiking, i.e. Class 3 excitability, phasic bursting, rebound spike, rebound burst, resonator, threshold variability, depolarizing after-potential, and accommodation. Together with the mixed-mode spiking behavior observed in mixed-mode neurons, we have observed 23 types of known biological neuron spiking behaviors.

In Hodgkin's classification, there are three basic classes of neuron excitability that can be discerned by spiking patterns: Class 1, Class 2, and Class 3 [18]. The nonlinear dynamical mechanism responsible for each class of excitability is reasonably well understood. Biological tonic neurons can exhibit Class 1 or Class 2 excitability, and in some cases can convert between the two excitabilities. Class 1 neurons have relatively low spike onset threshold and initial frequency, and the spike frequency increases with the strength of input current with a pronounced slope. Class 2 neurons exhibit much larger spike onset current thresholds and initial frequency, and the spike frequency is relatively insensitive to changes in the input strength. We found that tonic memristor neurons can be converted between being Class 1 and Class 2 excitable, simply by adjusting the Na^+ and K^+ membrane time constants determined by the load resistors and membrane capacitors (see Fig. 5 in [14]).

To study the neuronal stochasticity, we applied the classic stochastic spike train analysis, i.e. the joint interspike interval (JISI) analysis, to study stochasticity and correlation in spike patterns. The results show that VO_2 active memristor neurons exhibit input-noise sensitive stochastically phase locked firing, aka skipping, in a manner similar to biological neurons (see Fig. 6 in [14]).

III. Discussion: All-Memristor Electronic Cortex

Since both active memristor neurons and passive memristor synapses are fabricated from deposited thin film structures, it is in principle feasible to vertically stack repeated pairs of memristive neurosynaptic cores to directly map to the brain cortical layers. Fig. 4 illustrates the concept of vertically stacked multilayer all-memristor neurosynaptic cores. For simplicity, recurrent connections as well as inter-core connections are not

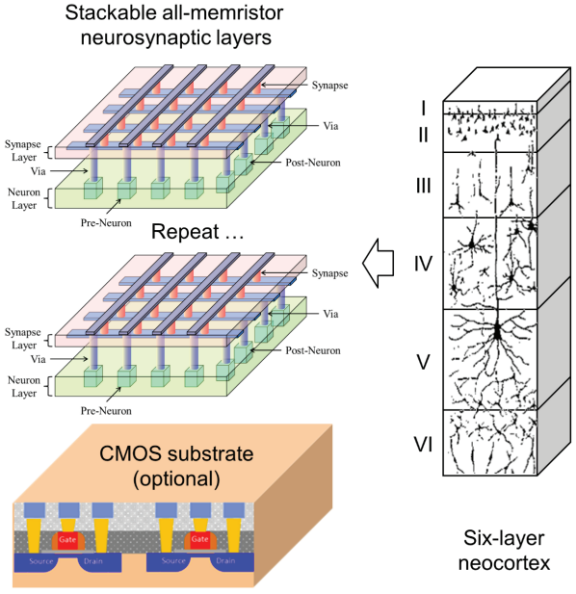

Fig. 4. Scalable and stackable all-memristor neurosynaptic layers could possibly map the connectivity of mammal's multilayered cerebral cortex with similar network scales.

shown, but can be included. Such a high-density pseudo-three-dimensional neural network connectivity cannot be easily achieved using conventional CMOS technology. The all-memristor neuromorphic processor can be integrated onto a CMOS substrate using, e.g. backend-of-the-line (BEOL) processes. A CMOS substrate is not a must-have, but can provide functions such as power regulation and communication.

If the conceived trade-off between neuron/synapse scalability and biological fidelity were eliminated, could it be possible to eventually realize mammal-level intelligence at similar levels of size, weight and power consumption (SWaP)? It is worth pointing out that most of today's artificial neural networks use simplified neuron emulators (e.g. LIF models) that only have a small subset of neurocomputational properties compared with real neurons. An important research topic is to understand and exploit the rich biological neuronal dynamics at the network level to see if the intelligent processing is rooted not only upon the connectivity, but also the complexity of building blocks.

ACKNOWLEDGMENT

This work was supported by HRL Laboratories, LLC. Kenneth K. Tsang, Stephen K. Lam, Xiwei Bai, Jack A. Crowell, Elias A. Flores have contributed to the work. We also acknowledge P. A. King and S. J. Kim for fabrication and electrical test support, T. C. Oh for electrical tests at the early stage of this project, and J. M. Chappell for electrical characterizations of passive memristors.

REFERENCES

[1] J. H. B. Wijekoon and P. Dudek, "Compact silicon neuron circuit with spiking and bursting behavior," Neural Netw. vol. 21, pp. 524–534, 2008.

[2] P. Livi and G. Indiveri, "A current-mode conductance-based silicon neuron for address-event neuromorphic systems," IEEE Int. Symp. Circ. Sys., 2009.

[3] J. Schemmel, et al. "A wafer-scale neuromorphic hardware system for large-scale neural modeling," IEEE Int. Symp. Circ. Sys., 2010.

[4] T. Yu, J. Park, S. Joshi, C. Maier, and G. Cauwenberghs, "65k-neuron integrate-and-fire array transceiver with address-event reconfigurable synaptic routing," IEEE Biomed. Circ. Sys. Conf., 2012.

[5] J. M. Cruz-Albrecht, M. W. Yung, and N. Srinivasa, "Energy-efficient neuron, synapse and STDP integrated circuits," IEEE Trans. Biomed. Circuits Syst. vol. 6, pp. 246–256, 2012.

[6] A. Joubert, B. Belhadj, O. Temam, and R. Heliot, "Hardware spiking neurons design: analog or digital?" IEEE Int. J. Conf. Neural Netw., 2012.

[7] J. Park, S. Ha, T. Yu, E. Neftci, and G. Cauwenberghs, "A 65k-neuron 73-Mevents/s 22-pJ/event asynchronous micro-pipelined integrate-and-fire array transceiver," IEEE Biomed. Circ. Sys. Conf., 2014.

[8] B. V. Benjamin, et al. "Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations," Proc. IEEE vol. 102, pp. 699–716, 2014.

[9] P. A. Merolla, et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science vol. 345, pp. 668–673, 2014.

[10] B. Sengupta, A. A. Faisal, S. B. Laughlin, and J. E. Niven, "The effect of cell size and channel density on neuronal information encoding and energy efficiency," J. Cerebral Blood Flow & Metabolism vol. 33, pp. 1465–1473, 2013.

[11] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" IEEE Trans. Neural Netw. vol. 15, pp. 1063–1070, 2004.

[12] S. K. Esser, et al. "Convolutional networks for fast, energy-efficient neuromorphic computing," Proc. Nat. Acad. Sci. vol. 113, pp. 11441–11446, 2016.

[13] L. O. Chua and S. M. Kang, "Memristive devices and systems," Proc. IEEE vol. 64, pp. 209–223, 1976.

[14] W. Yi, et al. "Biological plausibility and stochasticity in scalable VO_2 active memristor neurons," Nat. Commun. vol. 9, pp. 4661, 2018.

[15] M. D. Pickett, G. Medeiros-Ribeiro, and R. S. Williams, "A scalable neuristor built with Mott memristors," Nat. Mater. vol. 12, pp. 114–117, 2013.

[16] P. Stoliar, et al. "A leaky-integrate-and-fire neuron analog realized with a Mott insulator," Adv. Funct. Mater. vol. 27, pp. 1604740 (2017).

[17] Z. Wang, et al. "Fully memristive neural networks for pattern classification with unsupervised learning," Nat. Electron. vol. 1, pp. 137–145, 2018.

[18] A. L. Hodgkin, "The local electric changes associated with repetitive action in a non-medullated axon," J. Physiol. vol. 15, pp. 165–181, 1948.