Memory Technologies for Neural Networks

F. Merrikh-Bayat, M. Prezioso, X. Guo, B. Hoskins, and D. B. Strukov*
Department of Electrical and Computer Engineering
UCSB Santa Barbara, CA 93106, USA
*strukov@ece.ucsb.edu

K. K. Likharev
Department of Physics and Astronomy
Stony Brook University
Stony Brook, NY 11794, USA
konstantin.likharev@stonybrook.edu

***Abstract*— Synapses, the most numerous elements of neural networks, are memory devices. Similarly to traditional memory applications, device density is one of the most essential metrics for large-scale artificial neural networks. This application, however, imposes a number of additional requirements, such as the continuous change of the memory state, so that novel engineering approaches are required. In this paper, we briefly review our recent efforts at addressing these needs. We start by reviewing the CrossNet concept, which was conceived to address major challenges of artificial neural networks. We then discuss the recent progress toward CrossNet implementation, in particular the experimental results for simple networks with crossbar-integrated resistive switching (memristive) metal oxide devices. Finally, we review preliminary results on redesigning commercial-grade embedded NOR flash memories to enable individual cell tuning. While NOR flash memories are less dense then memristor crossbars, their technology is much more mature and ready for the development of large-scale neural networks.**

***Index Terms*—memristors, flash memory, resistive switching, hybrid circuits, CrossNets, pattern classifiers, neural networks.**

I. Introduction

One of major challenges in development of artificial neural networks for high-performance information processing is a lack of adequate hardware technology [1, 2]. Indeed, practical neural networks rely on a very large number of synapses, enabling high connectivity between neurons, i.e. a very high computational parallelism. For example, consider deep-learning convolutional neural network classifiers [3, 4] trained by the backpropagation algorithm [5]. (Such networks have been recently advanced to outperform other computer vision approaches in classification fidelity.) A Toronto group's network of this type has as many as ~5×10^5 neurons and ~ 5×10^8 synapses [6]. The complexity of neuromorphic networks required for performing more advanced cognitive tasks should be even higher, of the order of that of the human cerebral cortex, i.e. about 10^{11} neurons and 10^{15} synapses [7].

In principle, such complexity can be achieved with digital supercomputers or specialized graphics processing unit clusters – see, e.g., [8]. However, in addition to high cost, these approaches also suffer from mediocre energy efficiency as compared to biological networks, which consume much less energy primarily because they perform low-precision analog computation. Indeed, ~5-bit precision has been shown to be sufficient for at least most important cognitive tasks [4, 9]. The purely-CMOS analog circuits [10] are insufficient for meeting this challenge, mostly because CMOS-implemented synapses [11] are way too bulky.

II. CrossNets based on Memristive Crossbars

These challenges may be met using the so-called CrossNets [1, 2], based on the hybrid circuits combining CMOS technology with crossbar-integrated memory devices [12] (Fig. 1), for example memristors. In its simplest incarnation, memristor is a passive two-terminal device consisting a metal-oxide layer(s) sandwiched between two metallic electrodes. Layer's conductance may be increased (set) or decreased (reset), reversibly and continuously, applying relatively high positive or negative voltages. Thus, a typical response to symmetric high-voltage sweep is a pinched-hysteresis *I-V* loop (Fig. 1a). Because of highly nonlinear switching kinetics, device conductance may be sensed (read out) at lower voltages without disturbing the memory state. Due to the ionic nature of the memory, the compatibility of the manufacturing process with the monolithic 3D integration, and a very small footprint (determined only by the overlap area of the metallic electrodes), memristors are one of the best candidates for super-dense, nonvolatile, multibit memory cells [12-15]. This is why the (so-far, small) neural network hardware community may piggyback on the much larger drive of the electronic industry toward advanced nonvolatile memories.

To sustain high density, memristors are usually integrated into crossbar structures consisting of two layers of parallel electrodes (wires) perpendicular to each other (Fig. 1b). Because of the limited functionality of memristors, crossbar structures need to be combined with CMOS circuits, which may be more sparse. For serving relatively small memristor crossbars, such circuits may be located on their periphery, but prospective large-scale (say, $10^3\times10^3$ or larger) crossbars require area-distributed interfaces with CMOS subsystems (Figs. 1c,g). Such hybrid CMOS/memristor circuits are especially suited for artificial neural networks (Figs. 1d-f), in which memristors implement density-critical adjustable synapses, while the CMOS circuitry is used to mimic neural-body ("somatic") functions. Indeed, if somatic outputs are represented by voltages V_j applied to wires of one crossbar layer, and the CMOS circuits of the recipient cells sustain the virtual-ground condition on the wires of the counterpart layer, the current through each memristor equals $I_{ij} = G_{ij}V_j$. These partial currents are naturally summed up in each output wire *i*: $I_i = \Sigma_j G_{ij}V_j$, thus computing a vector-by-matrix product $y_i =$

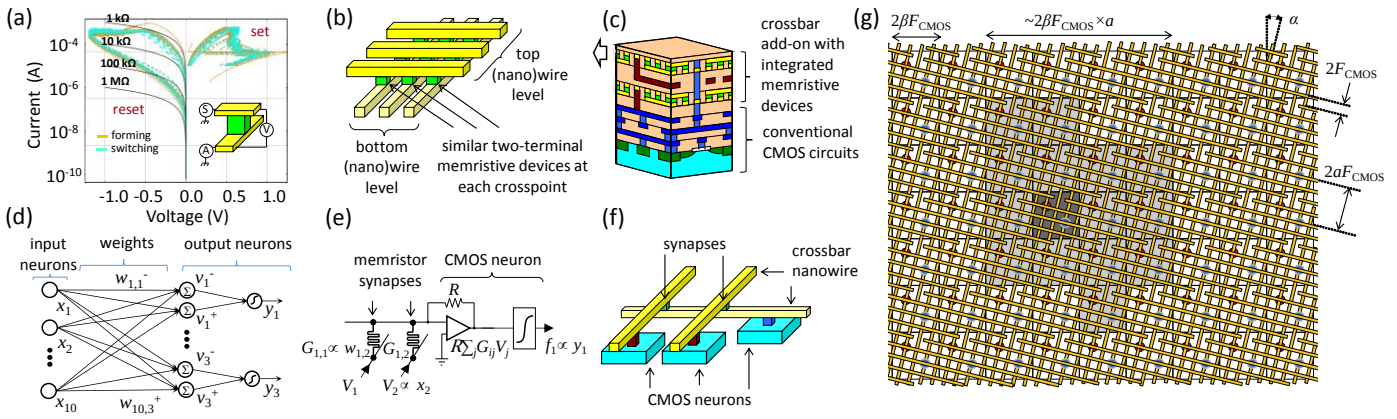

Fig. 1. Implementing CrossNets, artificial neural networks based on CMOS/memristive crossbar circuits [1, 2, 15]: (a) Typical experimental *I-V* curves of bipolar Pt/Al_2O_3/TiO_{2-x}/Pt memristors, (b) passive (transistor-free) memristor crossbar, (c) 3D hybrid CMOS/memristive crossbar circuit, (d) graph representation of a single-layer feedforward neural network, (e) analog implementation of the vector-by-matrix product, (f) its mapping on a hybrid circuit, and (g) top view of a hybrid circuit showing the set of neural cells (light gray) connected to a given cell (dark gray) via a wire / memristive device / wire link. For clarity, panel (e) shows only a small portion of a circuit, while panel (g) shows an example with small cell connectivity; in practical hybrid circuits the connectivity domain should be much larger.

$\Sigma_j w_{ij} x_j$, i.e. performing the operation that is the bottleneck in artificial neural networks, in a compact and energy-efficient way [10].

An essential feature of CrossNets is the area-distributed interface between the memristive crossbar and the CMOS circuitry, which ensures high bandwidth and low area overhead. It may be implemented, for example, using the so-called CMOL topology [12, 16], enabled by a crossbar array rotation by angle $\alpha = \arcsin(1/\beta) = \arctan(1/a)$ with respect to the mesh of CMOS-controlled vias, and a double decoding scheme, which provides a unique access to each crosspoint memristor (Fig. 1g). More specifically, two types of vias, one (shown with red dots) connecting to the lower and the other one (blue dots) connecting to the upper wire level of the crossbar, are arranged into a square array of cells with sides $2\beta F_{CMOS}$, where F_{CMOS} is a CMOS half-pitch and $\beta = (a^2 + 1)^{1/2} > 1$ is a dimensionless number that depends on the cell size (i.e., complexity) in the CMOS subsystem. With each CMOS cell implementing a soma, the interface allows each neuron to be connected to a domain of other neurons surrounding it via the crossbar's memristor synapses. Such connectivity domain may be quite large, especially for the 3D variety of the CMOL interface [16], limited mostly by the requirement of keeping crossbar wire resistances relatively small.

III. Single Layer Perceptron Demo with Memristive Crossbar Circuits

The practical implementation of memristive CrossNets is still challenging, mainly due to the immature memristor fabrication technology. The most critical requirement to the technology is to ensure a relatively low (within one octave) distribution of device forming and switching threshold voltages. (This condition enables individual forming, and then fine-tuning of every memristor of the crossbar, without disturbing already formed devices.) We have recently developed bilayer Al_2O_3/TiO_{2-x} memristors which feature such low variability [17] (Fig. 1a). The optimized technology was then used for the fabrication of integrated 12×12 crossbars (Figs. 2a, b). The crossbars featured a high uniformity of virgin (pre-formed) crossbar-integrated devices (Figs. 2c, d).

The fabricated memristive crossbar was used to implement a simple neural network (a single-layer perceptron) with 10 inputs and 3 outputs, fully connected with 30 synaptic weights (Fig. 1d). Such network is sufficient for performing, for example, the classification of 3×3-pixel black-and-white images with 9 network inputs (V_1,…,V_9) corresponding to pixel values, into 3 classes. (One more input, V_{10}, was used for the source of 3 adjustable biases of nonlinear activation functions.) We have tested the network on a set of 30 patterns including 3 stylized letters ("z", "v", and "n") and 3 sets of 9 noisy versions of each letter, formed by flipping one of the pixels of the original image - see the inset in Fig. 2g. Because of the limited set size, it was used for both training and testing.

Physically, each input signal was represented by voltage V_j equal to either +0.1 V or -0.1 V, corresponding, respectively, to the black or the white pixel. (The bias input V_{10} was -0.1V.) Each synaptic weight was implemented with a pair of memristors, so that $w_{ij} = G_{ij}^+ - G_{ij}^-$, enabling negative weights values [18]. The effective conductances $G_{ij}^{\pm}$ were in the range from 10 to 100 μS, so that the output currents I_i were of the order of a few μA. The network was trained "in situ", i.e. without using its external computer model, with the so-called Manhattan Update Rule, which is essentially a coarse-grain, batch-mode variation of the usual Delta Rule of supervised training [5]. At each iteration ("epoch") of the procedure, the training set patterns were applied, one by one, to network's input, and its outputs $f_i(n)$, where n is pattern's number, were used to calculate the weight increments. Once all patterns of the training set had been applied, and all due increments ΔG were calculated, and the synaptic weights modified.

In our system, the weights were modified in parallel for each half-column of the crossbar (corresponding to a certain value of index i in the above formulas), using two sequential voltage pulses. Namely, first a "set" pulse with amplitude V_{w+} = 1.3 V was applied to increase conductances of the synapses whose ΔG had been positive; then a "reset" pulse V_{W-} = -1.3V was

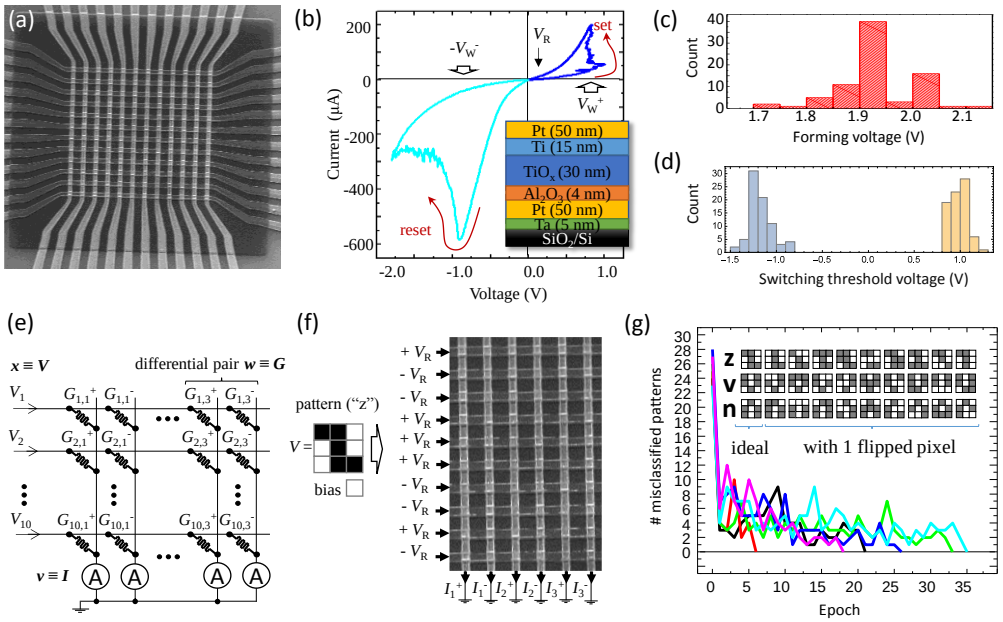

Fig. 2. Perceptron classifier demonstration [17]: (a) Integrated 12×12 crossbar with an Al_2O_3/TiO_{2-x} memristor at each crosspoint; (b) a typical *I-V* curve of a formed memristor, (c-d) histograms of forming voltages (c) and effective switching thresholds voltages (d) for set and reset transitions; (e) perceptron implementation using a 10×6 fragment of the memristive crossbar; (f) example of the classification operation for a specific input pattern; and (g) the convergence of network outputs, in the process of training, to the perfect (zero-error) set, for 6 different initial states. (The classification was considered successful when the output signal corresponding to the correct class of the applied pattern was larger than all other outputs.) The insets in panels (b) and (g) show device's cross-section and the used input pattern set, correspondingly. On panel (d), the positive / negative switching threshold voltages were defined as the smallest amplitudes of 500-μs voltage pulses that caused resistance change by more than 2 kΩ in memristors pre-set to their high / low resistive states.

applied to the remaining synapses of that half-column. This fixed-amplitude pulse procedure followed the Manhattan Update Rule only approximately, because the actual increment of conductance *G* depends on its initial value.

Due to this specific (though quite representative [13-15]) switching dynamics, the best classification performance was achieved when the memristors had been initialized somewhere in the middle of their conductance range, around 35 μS. At such initialization, the perfect classification was always reached - on the average, after 23 training epochs – see Fig. 2g.

IV. Redesigning NOR Flash Memory for Neural Network Applications

The results presented in the previous section are very encouraging; however, the development of VLSI memristor circuits may still require the introduction of an advanced memristor technology by at least one major chipmaker. On the other hand, flash memories, while less dense, are broadly used in VLSI circuits. After customization for analog state tuning of each cell, this technology may be used for artificial neural network applications. For example, Fig. 3a shows a commercial NOR flash memory array architecture from Silicon Storage Technology, Inc. (SST) [19], in which cells of the same row share transistor source and control gate ("word") lines, while transistor drains of all cells of the same column are connected to the same "bit" line. The array design is optimized for digital applications, and permits individual programming of the selected cells, e.g., by applying 1.6 V and 7.6 V to the selected gate and source lines, correspondingly, and grounding the selected drain line, while avoiding disturbance of half selected cells by applying > 2V on unselected drain lines and grounding the remaining lines. However, the cells cannot be erased individually because the process responsible for erasure (the Fowler-Nordheim tunneling of electrons from the floating gate to the control gate) is only weakly affected by the drain voltage – the only voltage which may be different for two adjacent cells.

To resolve this problem, we have modified the array structure (without changing the highly optimized cell fabrication technology) as shown in the Fig. 3b, i.e. by re-routing the gate lines in the "vertical" direction, i.e. perpendicular to the source lines [20]. The new cell arrays have been designed, fabricated (so far in the 180-nm technology of SilTerra, Inc.) and successfully tested. Fig. 3c shows the layout of the new array. Its cell area is 2.3 times larger than the original one (Fig. 3c) due to the additional real estate needed to accommodate two gate lines for each cell column, but is still much smaller than prior CMOS-implemented synapses [11].

To verify that the new array architecture enables a full inhibition of half-select disturb effects, we have performed a series of experiments, tuning all 8 cells in a 2×2 supercell array, one by one, to pre-selected goal values with a ~1% precision (Fig. 3e), using a simple, fully automated feedback procedure that had been originally developed for tuning memristors [21]. We have used the high-precision individual tuning of cells in the modified array for a preliminary demonstration of a small-scale four-quadrant gate-coupled vector-by-matrix multiplication [22], in which peripheral floating-gate transistors had been implemented with the same SST memory technology and integrated on the same chip (Fig. 4f). The results (Fig. 4g) show an excellent linearity (derivative variations below 1%) of circuit's transfer characteristics over a wide range of input currents.

V. Conclusion

We believe our group has made two significant steps toward high-performance, hybrid neuromorphic networks. First, an integrated crossbar with individually tunable memristors has been fabricated, and used for a successful demonstration of a small neural network classifier of B/W

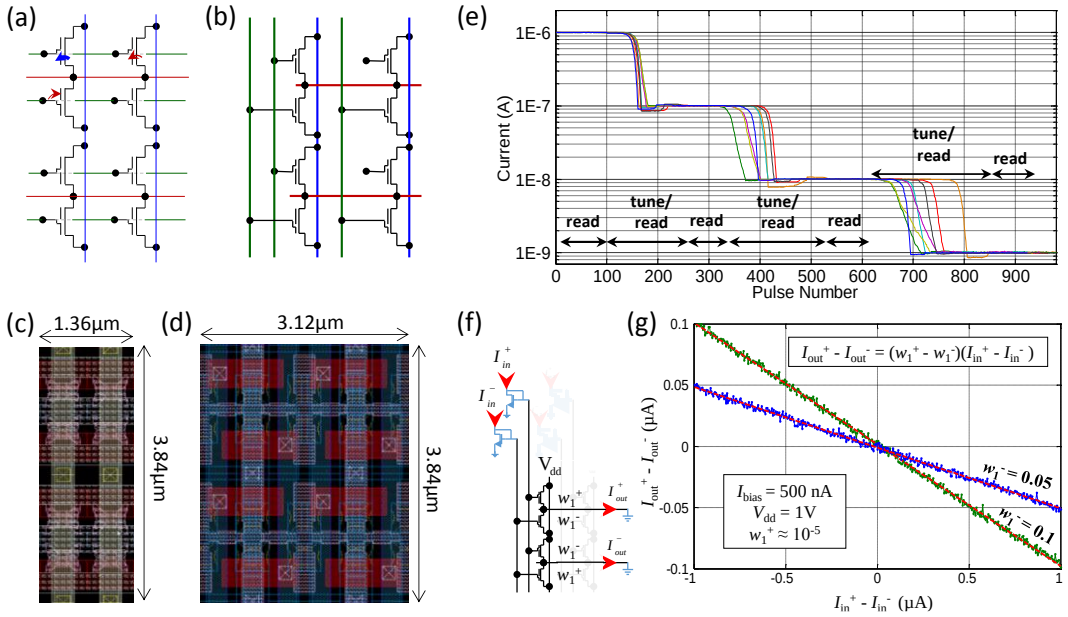

Fig. 3. Redesigning commercial NOR flash technology for analog applications [20]: (a) original and (b) modified 2×2 supercell array fragment, and (c, d) their corresponding layouts; (e) high-precision, sequential tuning of all cells of the modified array to 4 target values (1 μA, 100 nA, 10 nA, and 1 nA) of the readout current (as measured at V_G = 2.5 V, V_D = 1 V, V_S = 0 V); (f) circuit schematics and (g) preliminary experimental results for a gate-coupled vector-by-matrix multiplier, showing the measured transfer characteristics for two sets of matrix elements w_1^-. On panel (a), blue and red arrows show, respectively, the useful and undesirable processes of floating gate recharging.

images. Second, the high-density, industrial-grade NOR flash memory cells have been redesigned for analog applications.

Acknowledgments

Useful discussions with G. Adam, N. Do, M. Graziano, D. Hammerstrom, I. Kataeva, O. Kavehei, and L. Sengupta are gratefully acknowledged. This work was supported by the AFOSR under the MURI grant FA9550-12-1-0038, DARPA under Contract No. HR0011-13-C-0051UPSIDE via BAE Systems, and DENSO Corporation, Japan.

References

[1] K. K. Likharev, "CrossNets: Neuromorphic hybrid CMOS/nanoelectronic networks", Sci. Adv. Mater., vol. 3, pp. 322-331, Jun. 2011.

[2] D. B. Strukov, "Nanotechnology: Smart connections", Nature, vol. 476, pp. 403–405, Aug. 2011.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", Proc. IEEE, vol. 86, pp. 1–46, Nov. 1998.

[4] C. Farabet, B. Martini, B. Codra, P. Akselrod, E. Culurciello, and Y. LeCun, "NeuFlow: A runtime reconfigurable dataflow processor for vision", in: Proc. CVPRW'11, Colorado Springs, CO, June 2011, pp. 109-116.

[5] J. Hertz, A. Krogh, and R. G. Palmer, Introduction to the Theory of Neural Computation, Perseus: Cambridge, MA, 1991.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", in: Proc. NIPS'12, Lake Tahoe, NE, Dec. 2012, pp. 1097-1105.

[7] V. B. Mountcastle, The Cerebral Cortex, Harvard U. Press: Cambridge, MA, 1998.

[8] R. Ananthanarayanan, S. K. Esser, H. D. Simon, and D. S. Modha, "The cat is out of the bag: cortical simulations with 10^9 neurons and 10^{13} synapses", in: Proc. Supercomputing'09, Portland, OR, Nov. 2009, pp. 14-20.

[9] E. Säckinger, "Measurement of finite-precision effects in handwriting- and speech- recognition algorithms", Lecture Notes in Computer Science, vol. 1327, pp. 1223-1228, 1997.

[10] C. Mead, Analog VLSI and Neural Systems, Addison-Wesley: Boston, 1989.

[11] J. Hasler and B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems", Frontiers in Neuroscience, vol.7, art. 119, 2013.

[12] K. K. Likharev, "Hybrid CMOS/nanoelectronic circuits", J. Nanoelectron. & Optoelectron., vol. 3, pp. 203-230, Dec. 2008.

[13] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox - based resistive switching memories", Adv. Materials, vol. 21, pp. 2632-2663, Jul. 2009.

[14] H. S. P. Wong et al., "Metal - oxide RRAM", Proc. IEEE, vol. 100, pp. 1951 - 1970, May 2012.

[15] J. J. Yang, D. B. Strukov and D. R. Stewart, "Memristive devices for computing", Nature Nanotechnology, vol. 8, pp. 13-24, Jan. 2013.

[16] D. B. Strukov and R. S. Williams, "Four-dimensional address topology for circuits with stacked multilayer crossbar arrays", Proc. Nat. Acad. Sci., vol. 106, pp. 20155-20158, Dec. 2009.

[17] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. K., Likharev, and D. B. Strukov, "Training and operation of integrated neuromorphic network based on metal-oxide memristors", available online at http://arxiv.org/abs/1412.0611.

[18] F. Alibart, E. Zamanidoost, and D. B. Strukov, "Pattern classification by memristive crossbar circuits with ex-situ and in-situ training", Nature Commun., vol. 4, art. 2072, Jun. 2013.

[19] http://www.sst.com/technology/superflash-technology.aspx

[20] F. Merrikh-Bayat, X. Guo, H. A. Ommani, N. Do, K. K. Likharev, and D. B. Strukov, "Redesigning commercial floating-gate memory for analog computing applications", accepted for presentation at ISCAS'15, Lisbon, Portugal, June 2015; available online at http://arxiv.org/abs/1410.4781.

[21] F. Alibart, L. Gao, B. Hoskins, and D. B. Strukov, "High-precision tuning of state for memristive devices by adaptable variation-tolerant algorithm", Nanotechnology, vol. 23, art. 075201, Jan. 2012.

[22] C. R. Schlottmann and P. E. Hasler, "A highly dense, low power, programmable analog vector-matrix multiplier: The FPAA implementation", IEEE JETCAS, vol. 1, pp. 403-411, May 2011.