Mixed-Signal POp/J Computing with Nonvolatile Memories

M. Reza Mahmoodi and Dmitri B. Strukov
ECE Department, UC Santa Barbara
Santa Barbara, CA 93106-5630 USA
{mrmahmoodi,dimastrukov}@ucsb.edu

ABSTRACT

The present-day revolution in deep learning was triggered not by any significant algorithm breakthrough, but by the use of more powerful GPU hardware [1]. Though this revolution has stimulated the development of even more powerful dedicated digital systems [2, 3], their speed and energy efficiency are still insufficient for ultrafast pattern classification and more ambitious cognitive tasks. The main reason is that the use of digital operations for the implementation of neuromorphic networks, with their high redundancy and noise/variability tolerance, is inherently unnatural. On the other hand, the network performance may be dramatically improved using mixed-signal integrated circuits, where the key inference-stage operation, the vector-by-matrix multiplication, is implemented on the physical level by utilization of the fundamental Ohm and Kirchhoff laws [4-6].

In our talk, we will discuss the recent progress of such analog and mixed-signal neuromorphic networks based on floating-gate memories and metal-oxide memristors. In our earlier work we have shown that a minor modification of a highly optimized embedded NOR flash memory [7-9] enabled a successful demonstration of the first medium-scale network for pattern classification [10, 11]. Remarkably for such a first attempt, still using the older, 180-nm technology, the experimentally measured time delay and energy dissipation (per one pattern classification) were below, respectively, 1 μs and 20 nJ [10], i.e. at least three orders of magnitude better than those reported for the best digital implementation of the same task, with a similar fidelity, using the 28-nm IBM's TrueNorth chip [12]. Experimental results for the chip-to-chip statistics, long-term drift, and temperature sensitivity showed no evident showstoppers on the way toward practical deep neuromorphic networks with unprecedented performance [11].

Another way to further scaling down the mixed-signal neuromorphic networks is provided by novel nonvolatile two-terminal devices - "memristors" [13], whose conductance G may be continuously adjusted by the application of short voltage pulses of higher (~1 V) amplitude. These devices have a very low chip footprint, which is determined only by the overlap area of the metallic electrodes, and may be scaled down below 10 nm without sacrificing their endurance, retention, and tuning accuracy. Our group has developed [13, 14] and then improved [15] a new technology of fabrication of these devices (so far, with the ~200 x 200 nm^2 area), sufficiently reproducible to demonstrate the first simple neuromorphic network providing pattern classification based on the most prospective, passive (0T1R) crossbar circuits. The passive memristive technology is also naturally suitable for the 3D integration, e.g. for monolithical back-end ntegration with CMOS circuits, and we have already made the first steps toward such 3D circuits [17, 18].

Our more recent work [19, 20] shows that the performance for vector-matrix multiplier based on nonvolatile memories may be further improved, potentially exceeding the Pop/J computing regime, using a better peripheral circuitry design and a more advanced memory technology. This, in turn, enables more than 100x advantage in speed and an almost 10,000x advantage in energy efficiency over the state-of-the-art purely digital circuits for classification of large, complex patterns (Table 1).

KEYWORDS

Mixed-signal circuits, nonvolatile memories, floating-gate memories, artificial neural networks, metal-oxide memristor, vector-by-matrix multiplier

REFERENCES

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

[2] Y. H. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks", in: *Proc. ISSCC'16,* San Francisco, CA, Jan. 2016, pp. 262-263.

[3] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "Envision: A 0.26-to-10 TOps/W subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28 nm FDSOI", in: *Proc. ISSCC'17,*

AlexNet [1] single pattern classification:	Digital [2, 3]			Analog [10, 11, 16]					Visual cortex (estimates)
	GPU 28 nm (16 bit)	ASIC 65 nm (16 bit)	ASIC 28 nm (4 bit)	NOR ESF-1 180 nm (~6 bit)	NOR ESF-3 55 nm (~6 bit)	2D (1T1R) 55/200 nm memristor (~5 bit)	2D memristor 200 nm (~5 bit)	3D memristor 10 nm (~5 bit)	
time (s)	1.5×10^{-2}	2.9×10^{-2}	$\sim0.6\times10^{-2}$	$\sim1\times10^{-4}$	$\sim6\times10^{-5}$	$\sim3\times10^{-5}$	$\sim5\times10^{-6}$	$\sim10^{-6}$	$\sim3\times10^{-2}$
energy (J)	1.5×10^{-1}	0.8×10^{-2}	1×10^{-3}	$\sim3\times10^{-7}$	$\sim2\times10^{-7}$	$\sim2\times10^{-7}$	$\sim2\times10^{-8}$	$\sim10^{-8}$	$\sim5\times10^{-8}$

Table I. Time delay and energy consumption of the signal propagation through the convolutional (dominating) part of a large deep network [1], with ~0.65x10^6 neurons, at its various 2D implementations. The mixed-signal network estimates are based on the 55×55 = 3,025-step time-division multiplexing (TDM), natural for this particular network, the measured performance of our image classifier prototype [10], and the experimentally measured parameters of the ESF1 and ESF3 cells [7-9] and metal-oxide memristors [16].

San Francisco, CA, Feb. 2017, pp. 246-247.

[4] C. A. Mead. *Analog VLSI and Neural Systems*. Addison Wesley, 1989.

[5] F. Merrikh Bayat *et al*., "Memory technologies for neural networks", in: *Proc. IMW'15*, Monterey, CA, May 2015, pp. 1-4.

[6] L. Ceze *et al*., "Nanoelectronic neurocomputing: Status and prospects", in: *Proc. DRC'16*, Newark, DE, June 2016, pp. 1-2.

[7] F. Merrikh Bayat *et al.*, "Redesigning commercial floating-gate memory for analog computing applications", in: *Proc. ISCAS'15*, Lisbon, Portugal, May 2015, pp. 1921-1924.

[8] F. Merrikh Bayat *et al*., "Model-based high-precision tuning of NOR flash memory cells for analog computing applications", in: *Proc. DRC'16*, Newark, DE, June 2016, pp. 1-2.

[9] X. Guo *et al*., "Temperature-insensitive analog vector-by-matrix multiplier based on 55-nm NOR flash memory cells" in: *Proc. CICC'17*, Austin, TX, May 2017, pp. 1-4.

[10] F. Merrikh Bayat et al., "High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cells", *IEEE Trans. Neural Networks & Learning Systems*, 2018 (early access).

[11] X. Guo *et al*., "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology", in: *Proc. IEDM'17*, San Francisco, CA, Dec. 2017, pp. 6.5.1-6.5.4.

[12] S. K. Esser, R. Appuswamy, P. Merolla, J. V. Arthur, and D. S. Modha, 2015. Backpropagation for energy-efficient neuromorphic computing. in: *Proc. NIPS'15*, Montreal, Canada, Dec. 2015, pp. 1117-1125.

[13] M. Prezioso *et al*., "Training and operation of an integrated neuromorphic network based on metal-oxide memristors", *Nature*, vol. 521, pp.61-64, 2015.

[14] M. Prezioso *et al*., "Modeling and implementation of firing rate neuromorphic-network classifiers with bilayer Pt/Al_2O_3/TiO_{2-x}/Pt memristors", in: *Proc. IEDM'15*, Washington, DC, Dec. 2015, pp. 17.4.1-17.4.4.

[15] F. Merrikh Bayat, M. Prezioso, B. Chakrabarti, I. Kataeva, and D. Strukov, "Memristor-based perceptron classifier: Increasing complexity and coping with imperfect hardware", in: *Proc. ICCAD'17*, Irvine, CA, Nov. 2017, pp. 549-554.

[16] F. Merrikh Bayat, M. Prezioso, B. Chakrabarti, I. Kataeva, and D. B. Strukov, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits", ArXiv:1611.04465, submitted to *Nature Comm.*, Nov. 2017.

[17] G. C. Adam *et al*., "3-D memristor crossbars for analog and neuromorphic computing applications", *IEEE TED*, vol. 64, pp. 312-318, 2017.

[18] B. Chakrabarti *et al*., "A multiply-add engine with monolithically integrated 3D memristor crossbar/CMOS hybrid circuit", *Nature Scientific Reports*, vol. 7, art. 42429, 2017.

[19] M.R. Mahmoodi and D.B. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier circuit based on NOR flash memory technology", in: *Proc. DAC'18*, San Francisco, CA, June 2018 (accepted).

[20] M. Bavandpour, M. R. Mahmoodi, and D. B. Strukov, "Energy-efficient time-domain vector-by-matrix multiplier for neurocomputing and beyond", *ArXiv:1711.10673,* Nov. 2017.