Analog and Neuromorphic Computing with Memristor Arrays

C. Li[1], M. Hu[2], Y. Li[1], D. Belkin[1,5], H. Jiang[1], N. Ge[3], E. Montgomery[2], J. Zhang[2], W. Song[1], N. Dávila[2], C. Graves[2], Z. Li[2], J. P. Strachan[2], P. Lin[1], Z. Wang[1], M. Barnell[4], Q. Wu[4], R. S. Williams[2], J. J. Yang[1], Q. Xia[1]

[1]Department of ECE, University of Massachusetts, Amherst, MA 01003
[2]Hewlett Packard Labs, Hewlett Packard Enterprise, Palo Alto, CA 94304
[3]HP Labs, HP Inc., Palo Alto, CA 94304
[4]Information Directorate, Air Force Research Laboratory, Rome, NY 13441
[5]Swarthmore College, Swarthmore, Pennsylvania 19081, USA.

Email: can@umass.edu

The surge of artificial intelligence (AI) development calls for more powerful computing hardware. Conventional von Neumann architecture separates computing and memory units, and the communication between them limits the performance[1]. Memristors with tunable resistance perform computation in the place where data is stored, which promises an ideal solution for future AI hardware. However, its large-scale demonstration is yet to be reported mostly due to the large stochastic variation. Here we report our progress on this issue from device engineering, integration process, and peripheral electronics development perspectives. We experimentally demonstrated image compression and image convolutional filtering with 128×64 memristor crossbar arrays, the largest of its kind to date.

The memristor crossbar performs vector-matrix multiplication (VMM) in one step. The analog voltage inputs are weighted by the corresponding memristor conductance (Ohm's Law) and summed up in the column wires (Kirkoff's current law). We monolithically integrate the custom developed Ta/HfO_2[2] memristors on a foundry-made chip with transistors arrays. The fabrication processes and results are shown in Fig. 1. Custom measurement system is built to supply 128 different voltages amplitude and to measure 64 channels of current in parallel[3].With this system, we programmed the 8,192 memristors in the crossbar into different conductance levels, with 6-bit / 64-levels precision of and 99.8% yield. We also achieved linear current-voltage relation and stable multilevel states[4], all contribute to accurate computation.

The 128×64 memristor crossbar is firstly configured to perform discrete cosine transform (DCT)[4], which is widely used for image/videos compression and signal processing. Each matrix element is represented by the conductance difference of two memristors on the same columns, and the subtraction is performed in the crossbar by applying voltages of different polarity but same amplitude on the memristor pair. The image (Fig. 2a) is compressed by converting it to a spectrum with DCT, keeping only the low frequency part. The experimental result after keeping only 15% spectrum information is shown in Fig 2b, which is well reconstructed as compared with the result from a pure software approach. We then configure the arrays to perform image filtering, which is equivalent to the convolutional layers in ConvNets. The input image with intentionally added noise is shown in Fig. 2c, and two representative outputs are shown in Fig. 2d, in which the memristor filters have successfully reduced the image noises or have detected edges. We further implement a two-layer fully connected perceptron to recognize MNIST handwritten digits (Fig. 3a, 3b). The result shows that, after *in-situ* training on 80,000 handwritten digits, we experimentally achieved 91.71% accuracy on a separate test dataset of 10,000 images (Fig. 3c).

The memristor arrays perform VMM in one step with low latency. The in-memory computing paradigm breaks von-Neumann bottleneck. The arrays could also handle analog signals without digital conversion, promising fast and low-power implementation of neural networks for future AI.

[1] Y. Lecun, Y. Bengio & G. Hinton, *Nature*, 521, 436 (2015)
[2] H. Jiang et al, *Sci. Rep*. 6, 28525 (2016); H. Jiang, L.L. Han, H. L. Xin & Q. Xia, *EIPBN* 2016
[3] M. Hu et al, *Adv. Mater.*,30, 1705914 (2018)
[4] C. Li et al, *Nat. Electron.*, 1, 52 (2018)
[5] C. Li et al, under review *by Nat. Commun.*

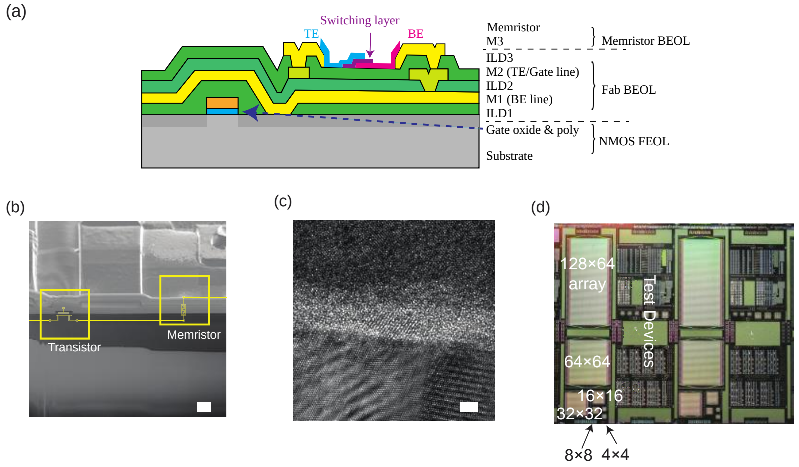

Figure 1: (a) The cross-sectional schematic of integrating memristors with transistor. (b) The fabrication result of an integrated one-transistor one-memristor cell in a scanning electron microscope (SEM) image. Scale bar, 2μm. (c) The transmission electron microscope (TEM) image of a Ta/HfO2 memristor. Scale bar, 2 nm. (d) The chip photo in one-die with integrated memristors.

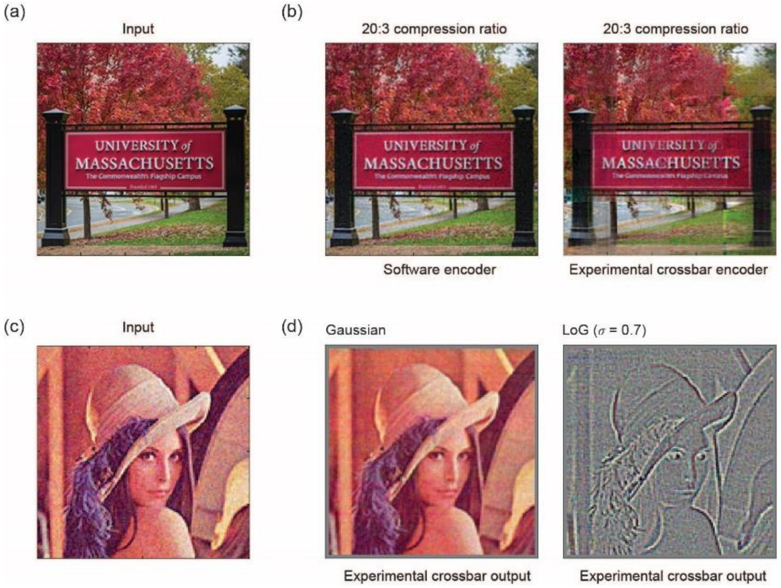

Figure 2: (a) The original image for compression experiment. (b) The reconstructed image with the compressed data from a software encoder and crossbar DCT experimental encoder respectively. (c) The original Lena image with artificial Gaussian noise for convolution experiment. (d) Two (out of ten) representative filtering output for noise reduction and edge detection respectively.

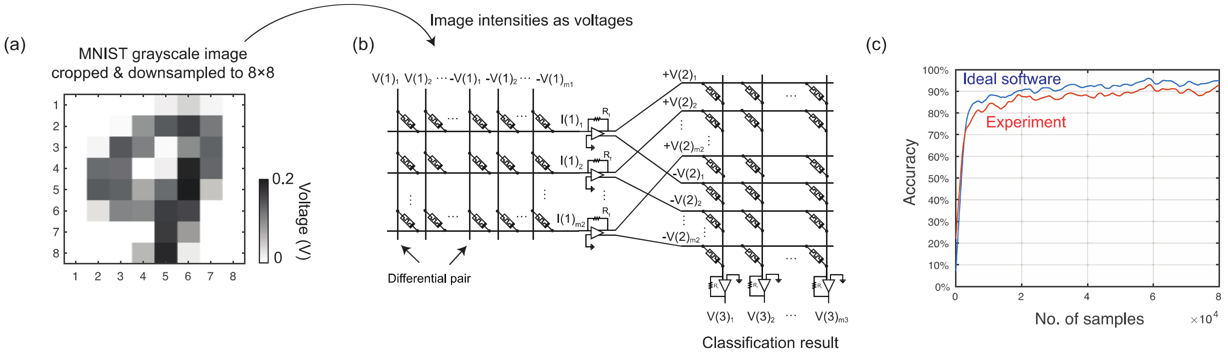

Figure 3: (a) A typical MNSIT handwritten digit for neural network recognition. The original 28×28 images are cropped to 20×20, and then down-sampled to 8×*8 to fit the array size. (b) The two-layer perceptron is implemented by the 128×64 crossbar. Part of the hidden layer neuron functionality is implemented in software. (c) The memristor neural network classifies 91.71% of all 10,000 digits correctly after training on 80,000 handwritten digit images. The gap between experiment and simulation based on ideal devices is 2%-4%.*