CMOS-integrated nanoscale memristive crossbars for CNN and optimization acceleration

Can Li, Jim Ignowski, Xia Sheng, Rob Wessel, Bill Jaffe, Jacqui Ingemi, Cat Graves, John Paul Strachan
Hewlett Packard Labs and Silicon Design Lab, Hewlett Packard Enterprise, Palo Alto, CA, USA
john-paul.strachan@hpe.com

(Invited Paper)

***Abstract*—While memristive crossbars have been reported to offer substantial performance efficiency benefits orders of magnitude above digital processors, there remain high risks in analog computing platforms using emerging non-volatile memory technologies, primarily due to device performance, variability, yield, and interactions with peripheral circuits. We directly integrated CMOS and nanoscale (down to 25 nm) memristors for fully on-chip reading/programming/computing demonstrations. We operate in a low power regime, program with fine control, showing high yield and low variability across our memristive arrays. With the integrated chip, we successfully demonstrated a multi-layer convolutional neural network with MNIST classification accuracy of above 95.3%, demonstrating several concepts in proposed architectures for hybrid analog-digital computing. The ability to tackle NP-hard optimization problems is also experimentally demonstrated with this platform. This work de-risks many of the chief concerns for an accelerator based on analog rather than purely digital computing circuits, as well as validating the core elements of a future in-memory computing architecture.**

***Index Terms*—memristor, non-volatile memory, in-memory computing, analog computing, neural networks**

I. Introduction

Computational workloads in the exciting areas of Artificial Intelligence (AI) and Big Data Analytics are growing faster than Moore's law, particularly with the sizes of the neural networks (NN), and the number of operations needed to train them. Current digital hardware is not well-suited to these workloads due to their data intensive nature and the von Neumann bottle-neck in the architecture. New and emerging architectures, such as Google TPU's systolic array, do not fully address these challenges, utilizing distributed and near-memory approaches. A fully in-memory computing approach, while more promising, requires a method to co-locate computing and data-storage. This is achievable with memristor/ReRAM crossbar arrays, which allow for non-volatile storage of NN weights with tunable resistances and parallel analog computation using Ohm's and Kirchhoff's laws.

Recent studies have shown promising results to accelerate various computing tasks [1]–[6]. Our first prototype chip used large 2 μm memristors and off-chip driving and sensing circuitry, but showed the feasibility of the approach [1], [2]. In parallel, we have proposed scalable architectures [7]–[9] targeting deep learning workloads with increased power efficiency forecasted. Nonetheless, the technology has remained high risk, with challenges in coupling highly dynamical and sensitive analog memory cells with yield and variability challenges with digital sensing/driving circuits.

Here we describe recent progress developing an integrated platform combining nanoscale memristive devices with CMOS driving and sensing circuits. The memristors exhibit a wide dynamic range and high yield, with promising potential to operate in a low power regime. The chip integration de-risks many circuit design concerns and enables the demonstration of several concepts proposed in our architectures, while reducing the unfavorable parasitics of off-chip driving circuits.

II. Nano-scale memristors with multi-bit capability

The memristive devices are integrated with foundry CMOS in a back-end-of-the-line (BEOL) process. The layers of the memristor materials stack (Ta/TaO$_x$/Pt) are deposited with room temperature sputtering, without damage to underlying CMOS elements. Electron-beam lithography was used to pattern the electrodes for sub-100 nm features. Fig. 1 shows a cross-sectional transmission electron microscope (TEM) image and a top view scanning electron microscopy (SEM) image of an integrated Ta/TaO$_x$/Pt memristor (50 nm×50 nm and 25 nm×25 nm) [10].

The integrated nanoscale memristive devices exhibit a promising conductance range up to 10^6 with a high yield, particularly for the smallest device sizes (See Fig. 2(a) and Fig. 2(b)). After conductance programming, read stability was gauged using a highly sensitive semiconductor parameter analyzer, with the standard deviation (σ_G) and coefficient of variation (σ_G/G) shown in Fig. 2(c) and Fig. 2(d). From the data one finds that the σ_G is generally smaller than 1 μS, and more importantly, the absolute σ_G is smaller when the device is programmed to lower conductances. We have previously

This research was based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2017-17013000002, and by the Army Research Office under Grant Number W911NF-19-1-0494. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ODNI, IARPA, Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

This manuscript was presented at IEEE International Memory Workshop 2020, download at: http://dx.doi.org/10.1109/IMW48823.2020.9108112

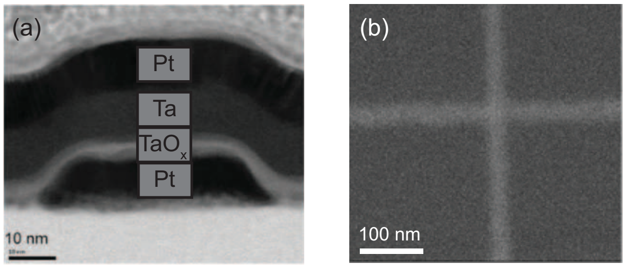

Fig. 1. (a) Cross-sectional TEM of an integrated Ta/TaO$_x$/Pt memristor. (b) Top view SEM of a cross-point memristor of dimensions 25 nm×25 nm

demonstrated that the device can be programmed to more than 256 states, or 8-bit, with $\pm 2\sigma$ as a discrete state [10].

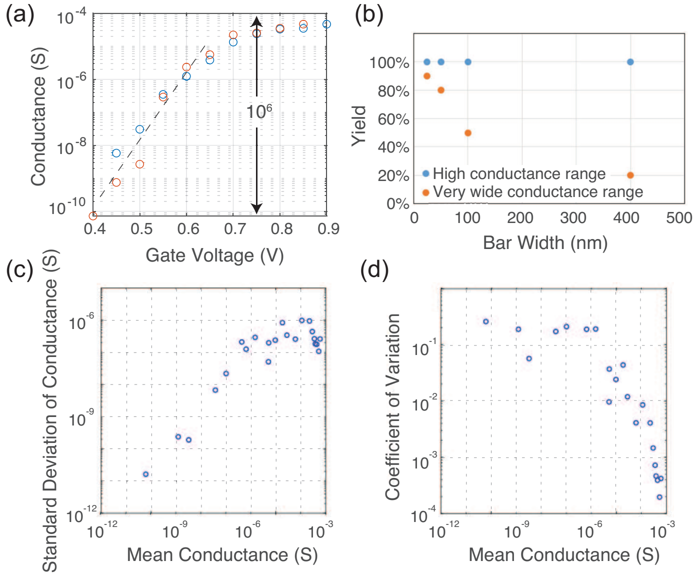

Fig. 2. (a) The relationship between the programmed conductance and the applied gate voltage during programming, showing a tunable conductance range from 0.1 nS to 100 μS. (b) Smaller devices show higher yield in operating at the wider conductance range. (c,d) The standard deviation and the percent variation of the repeated conductance reads with respect to the conductance state.

III. Analog Circuit Designs for Integrated 180 nm CMOS and Memristive Devices

Our new integrated chip platform includes both driving and sensing analog circuits, which are designed and taped out with TSMC's 180 nm technology node. Each test chip (called "SuperT") consists of three 64×64 memristive crossbar arrays, along with digital control and analog sensing circuits for performing in-memory analog computations. A die photo is shown in Fig. 3(a). Fig. 3(b) shows a representative diagram of a crossbar and its analog circuits. Each array utilizes digital to analog converters (DACs) to drive analog voltages (inputs) to the rows of the array, circuits to rapidly sense ($< 1\mu s$) the output currents, and analog to digital converters (ADCs) to provide final results.

After integrating the memristors with the silicon circuits, the chips are wire-bonded in a package and inserted into a custom printed circuit board (PCB) control system, which provides a digital interface to the test chip and supplies a number of DC analog reference signals (see Fig. 3(c)).

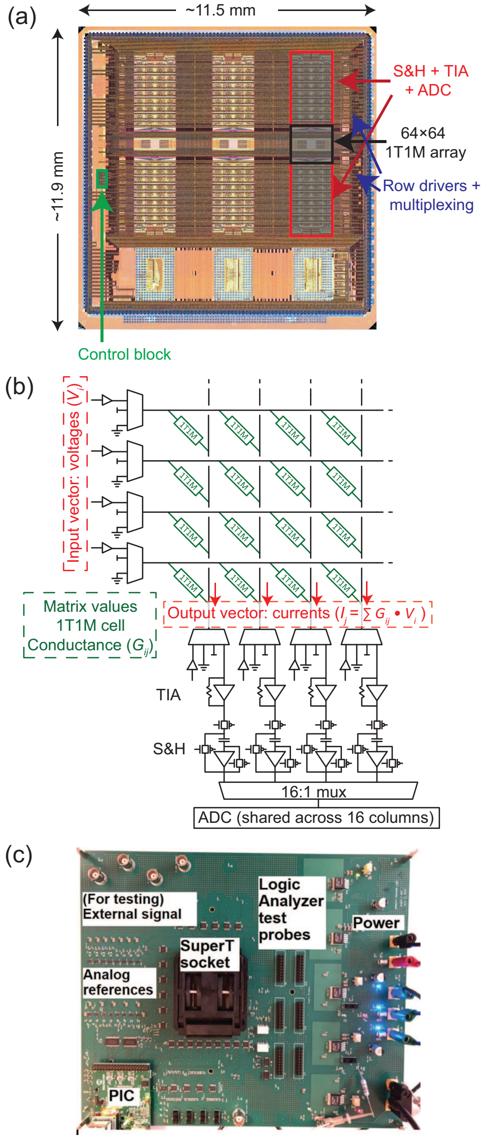
(a)
~11.5 mm
~11.9 mm
S&H + TIA + ADC
64×64 1T1M array
Row drivers + multiplexing
Control block
(b)
Input vector: voltages (V_i)
1T1M
Matrix values 1T1M cell Conductance (G_{ij})
Output vector: currents ($I_j = \sum G_{ij} \bullet V_i$)
TIA
S&H
16:1 mux
ADC (shared across 16 columns)
(c)
(For testing) External signal
Logic Analyzer test probes
Power
SuperT socket
Analog references
PIC

Fig. 3. (a) The floorplan of our integrated analog design and memristive crossbar arrays. (b) A simplified circuit diagram showing just four rows and columns for clarity, while the actual memristive crossbar arrays are 64×64. (c) Testboard interfacing our chip with a micro-controller. The board provides various DC analog references and debugging capabilities.

Utilizing nanoscale memristors, we have shown higher resistances [10] than previously reported in micrometer scale devices [2], allowing for lower power operation. The memristor arrays were successfully programmed to target conductances below 200 μS as shown in Fig. 4(a). The integrated platform significantly reduced circuit parasitics, enabling finer memristor conductance programming in this higher resistance range. Once programmed, different analog conductance levels in the memristors are maintained for a prolonged period of time, with Fig. 4(a) showing minimal conductance drift over 20 hours at

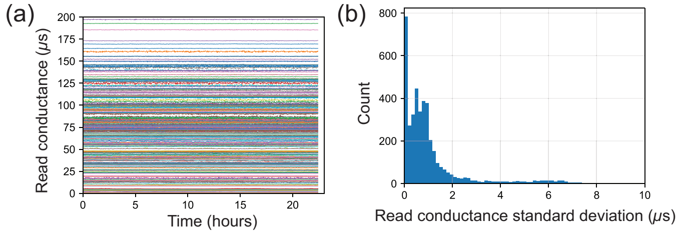

Fig. 4. (a) The rention performance of all devices in a 64×64 array that measured with the integrated peripheral circuit. Each datapoint is averaged from 50 repeated reads. (b) Multilevel retention performance of our integrated TaO_x device. The data was generated from conductance reads with 0.2 V read voltage from all devices in a 64×64 array for 10,000 times.

room temperature. Each datapoint is averaged from 50 reads to improve individual read accuracy. The effective number of bits supported by these analog nanoscale devices is determined by the stability of the conductance states. We observe the majority of the memristors have a standard deviation of 1 μS or smaller (4(b)), consistent with individual device measurements in Fig. 2, and supporting the multi-bit programmability. Over 90% of the devices can be formed/programmed with voltages smaller than 2.4 V and pulse widths of 20 ns, although a minority required higher voltages (the 180 nm transistors supplies the voltages up to 5 V). These are encouraging results, particularly in the favorable scaling of driving voltages and lower conductances with reduced device size.

IV. Applications

A. Convolutional neural networks

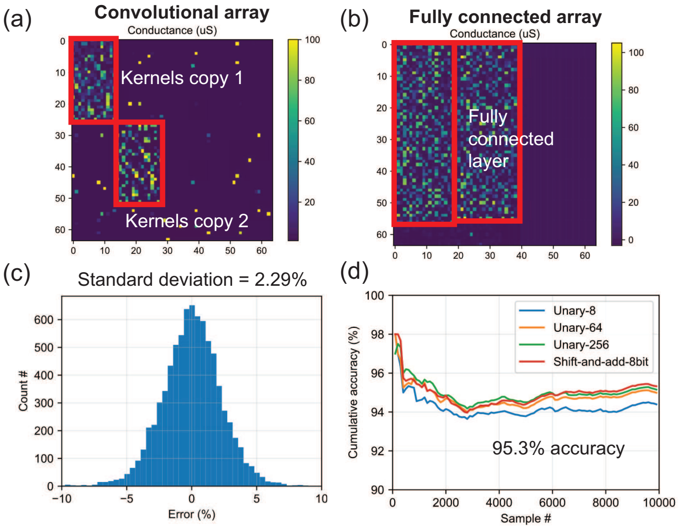

Fig. 5. (a, b) Two different 4k (64×64) memristor crossbars in a same chip were programmed for the convolutional and fully-connected layer weights with conductances of each device within 100 μS (i.e. larger than 10 kΩ in resistance). (c) The analog computing error follows a normal distribution with a standard deviation of about 2.29%; (d) The shift-and-add digitization approach achieves highest 95.3% accuracy with the most power efficiency.

Integration of the nanoscale memristors and peripheral driving/sensing circuits enables improved computing and programming performance, demonstrated by building a multilayer convolutional neural network (CNN) to classify the handwritten digits in the MNIST (Modified National Institute of Standards and Technology) dataset. To allow for negative NN weights, we use the conductance difference between two memristors (a differential pair) to represent one weight value. Therefore, 7 × (5×5) convolutional kernels plus 7 biases are mapped to a 26 (=5×5+1) × 14 (=7×2) array, and a 112×10 fully connected layer plus 10 biases maps to a 113×10 memristor array. These are programmed in two crossbar arrays within the same chip (see Fig. 5(a) and Fig. 5(b)).

This work represents the first direct experimental implementation of several concepts proposed in our ISAAC and PUMA architectures [7], [8]: reproducing convolutional kernels for higher parallelism and a novel shift-and-add technique for streaming inputs that minimizes the analog-digital converter (ADC) precision requirements, which otherwise bottlenecks overall system performance. ISAAC and PUMA [7], [8] forecast a 10-1,000x improvement over GPUs. 5(a) shows the duplicated kernels for higher parallelism, with input vectors first normalized between 0 and 1 (image input and intermediate layer input after ReLu are both non-negative). The massively parallel multiplications are performed by applying voltages to the rows, sensing the current differences from the columns. The results are fed to the next crossbar array bit-by-bit using a fixed input amplitude, while bit-shifting and adding the resulting output stream [7]. This novel technique saves overhead in chip area and energy consumption compared to using analog inputs, and enables the use of devices with very low conductance, further saving power.

Densely connected NN layers are costly in GPU and other digital architectures, but naturally supported here. We measured our analog multiplication errors (difference between experimental outputs and expected ideal outputs, divided by the output) to be around 2.29% (Fig. 5(c)). Using the 8-bit shift-and-add input encoding approach, we easily achieved a 95.33% experimental classification accuracy from our chip through all 10,000 samples of the MNIST test-set (Fig. 5(d)). The software baseline is 98.15%. While current work is promising, we expect higher NN accuracy may be achieved with finer programming of the memristor conductances.

B. Accelerator for optimization problems

Optimization problems, where a cost function (or Energy) is minimized across a high-dimensional set of variables, are of great importance with applications including improving airline scheduling, resource allocation, and wiring in VLSI. Such problems are extremely time-consuming to solve exactly, but can be tackled with heuristic and approximate methods more rapidly. Ising machines and Hopfield Neural Networks (HNN) implement the latter, where candidate solutions are iteratively improved based on local gradient calculations and stochastic exploration of the cost function landscape. In the physical implementation of such algorithms, the most power-intensive tasks are vector-matrix multiplications and the injection of controlled noise. These can both be implemented using

This manuscript was presented at IEEE International Memory Workshop 2020, download at: http://dx.doi.org/10.1109/IMW48823.2020.9108112

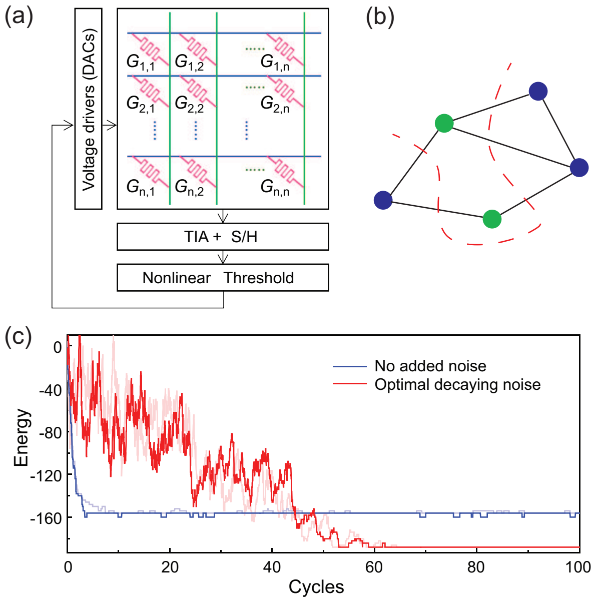

Fig. 6. (a) A crossbar-based system for implementing a Hopfield network, a type of recurrent neural network. (b) Illustration of a simple Max-Cut graph problem. (c) Experimental data for a 60 × 60 Max-Cut problem solved with our memristor crossbar chip. With the proper control of analog noise, the globally optimal solution can be reached with high probability. Transparent lines show the simulation data while solid lines are experiment data.

crossbars of memristors computing in the analog rather than digital domain, as illustrated in Fig. 6(a).

Target optimization problems are programmed into the memristor array values, representing the set of objectives and/or constraints. The output of the array is sampled and thresholded, with the result directly fed to I/O buffers to update the binary status of the neurons, which are used as inputs for the next cycle (more details in [6]). An example problem is shown in Fig. 6(b) for what is known as the ”Max-Cut” optimization problem. Given a graph with nodes and edges, the objective is to determine a partitioning into two sets of nodes such that the edges between them is maximized. This is an NP-hard optimization problem without any known efficient algorithm. Fig. 6(c) shows experimental data solving a 60×60 Max-Cut problem in our analog memristor crossbar chip. Each cycle represents an update of all nodes, and a temporal noise profile is added to inject advantageous randomness early on, while tuning this down to zero in the final calculations (similar to the simulated annealing algorithm). We have forecasted that such a memristor-based optimization solver has over four orders of magnitude higher solution throughput per power compared to fully digital approaches and present-day quantum and optical accelerators [6].

V. Related work

The compelling advantages of analog in-memory based computing over conventional computing have attracted many related approaches to that described here. Ambrogio et al. [11] explored the use of phase change memory, a more mature non-volatile technology. One challenge addressed in this work was the resistance state drift of such devices, and the additional circuit elements needed to mitigate these effects. Cai et al. [6] recently reported a chip with on-chip integrated peripheral circuits, while using passive WO_x memristive crossbars which limited demonstrations to small portions of an array. Chen et al. [4] built a functional integrated chip successfully demonstrating experimentally MNIST classification. While using foundry ReRAM, the work was limited to binary states, reducing some of the performance potential. Yao et al. [3] has also shown integrated chips for convolutional neural networks, with 128×16 sub-micron (0.5 μm) memristive crossbars on each chip. The present work showcased fully analog computations across multiple cascaded memristive crossbar arrays, with on-chip integrated peripherals with sub-100 nm (25 nm) devices, supporting a wide range of applications and demonstrations.

References

[1] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, “Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine,” *Advanced Materials*, vol. 1705914, pp. 1–10, 2018.

[2] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves *et al.*, “Analogue signal and image processing with large memristor crossbars,” *Nature Electronics*, vol. 1, no. 1, p. 52, 2018.

[3] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, and H. Qian, “Fully hardware-implemented memristor convolutional neural network,” *Nature*, vol. 577, no. 7792, pp. 641–646, 2020.

[4] W. H. Chen, C. Dou, K. X. Li, W. Y. Lin, P. Y. Li, J. H. Huang, J. H. Wang, W. C. Wei, C. X. Xue, Y. C. Chiu, Y. C. King, C. J. Lin, R. S. Liu, C. C. Hsieh, K. T. Tang, J. J. Yang, M. S. Ho, and M. F. Chang, “CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors,” *Nature Electronics*, vol. 2, no. 9, pp. 420–428, 2019.

[5] S. Yin, X. Sun, S. Yu, and J.-s. Seo, “High-Throughput In-Memory Computing for Binary Deep Neural Networks with Monolithically Integrated RRAM and 90nm CMOS,” *arXiv*, 2019. [Online]. Available: http://arxiv.org/abs/1909.07514

[6] F. Cai, S. Kumar, T. Van Vaerenbergh, R. Liu, C. Li, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. Lu *et al.*, “Harnessing intrinsic noise in memristor hopfield neural networks for combinatorial optimization,” *arXiv preprint arXiv:1903.11194*, 2019.

[7] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, “Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars,” in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2016, pp. 14–26.

[8] A. Ankit, I. E. Hajj, S. R. Chalamalasetti, G. Ndu, M. Foltin, R. S. Williams, P. Faraboschi, W.-m. W. Hwu, J. P. Strachan, K. Roy *et al.*, “Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference,” in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2019, pp. 715–731.

[9] A. Ankit, I. E. Hajj, S. R. Chalamalasetti, S. Agarwal, M. Marinella, M. Foltin, J. P. Strachan, D. Milojicic, W.-m. Hwu, and K. Roy, “Panther: A programmable architecture for neural network training harnessing energy-efficient reram,” *arXiv preprint arXiv:1912.11516*, 2019.

[10] X. Sheng, C. E. Graves, S. Kumar, X. Li, B. Buchanan, L. Zheng, S. Lam, C. Li, and J. P. Strachan, “Low-conductance and multilevel cmos-integrated nanoscale oxide memristors,” *Advanced Electronic Materials*, p. 1800876, 2019.

[11] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. W. Burr, “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature*, vol. 558, no. 7708, pp. 60–67, 2018.