180-nm NOR-Flash Mixed-Signal Neuromorphic Image Classifier: Chip-to-Chip Statistics, Retention, and Temperature Sensitivity

X. Guo, F. Merrikh-Bayat, M. Prezioso, and D. B. Strukov
Department of Electrical and Computer Engineering
UC Santa Barbara
Santa Barbara, CA, U.S.A.
strukov@ece.ucsb.edu

***Abstract* - Mixed-signal neuromorphic circuits based on analog nonvolatile memory devices may far surpass their digital counterparts in performance and energy efficiency. Recently, we have successfully implemented one such circuit – a medium-scale multilayer perceptron image classifier based on floating-gate memory cell matrices, redesigned from a commercial-grade 180-nm NOR flash memory, and have experimentally confirmed its superior performance and energy efficiency. In this paper, we report results of extended experimental testing of this circuit, focused on the chip-to-chip statistics, long-term drift, and temperature sensitivity. The results are very encouraging, showing no evident showstoppers on the way toward practical deep neuromorphic networks based on this technology.**

Keywords - neuromorphic computing; embedded NOR fash memory; reliability; temperature-insensitive design

I. INTRODUCTION

Although the main principles of analog neuromorphic circuits have been proposed three decades ago [1], their real-world applications were rather limited, in part due to the lack of efficient implementations of adjustable synaptic weights - a critical component of almost any neural network. In the most sophisticated demonstrated systems of this type [2, 3], the weights were implemented using "synaptic transistors" [2, 3], floating-gate memory devices fabricated using a standard CMOS logic process. Unfortunately, the chip footprint of such devices is rather large (~1,000 F^2 per cell, where F is a process feature size), resulting in lower speed and energy efficiency.

Recently, we have successfully designed, taped-out, and tested a 28×28-binary-input, 10-output, three-layer perceptron classifier [4] (Fig. 1), based on two floating-gate cell matrices, redesigned from a commercial embedded 180-nm NOR flash memory, to allow individual, precise analog tuning of cell conductances [5, 6]. The testing showed a 94.7% fidelity of classification of the MNIST benchmark, within 2% of the best value obtained in computer simulations of this network. Both the inference time delay (below 1 μs) and the energy consumption (below 20 nJ per pattern) are about $\times 10^3$ lower than those of the IBM TrueNorth chip, fabricated using a 28-nm technology, for the same task, at similar fidelity [4]. In this paper, we describe results of additional testing of chip-to-chip variations, weight retention, and temperature sensitivity, and discuss their impact on the system's performance.

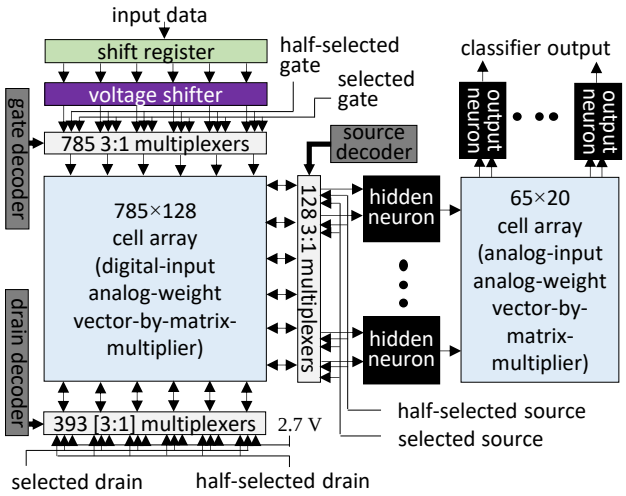

Fig. 1. High-level architecture of the demonstrated multilayer perceptron network (with the 2nd array's tuning circuitry not shown for clarity).

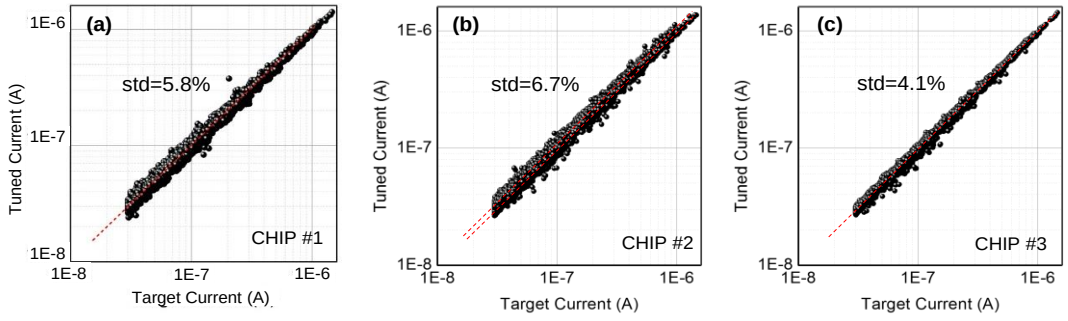

Fig. 2. Synaptic weight import (i.e. cell tuning) statistics: the measured cell currents, at the input voltage of V_G=2.5V, V_D=1V, vs. the target currents (computed at the external network training). The dashed red lines correspond to perfect tuning.

This work was supported by DARPA's UPSIDE program under contract HR0011-13-C-0051UPSIDE via BAE Systems, Inc and Google Faculty award.

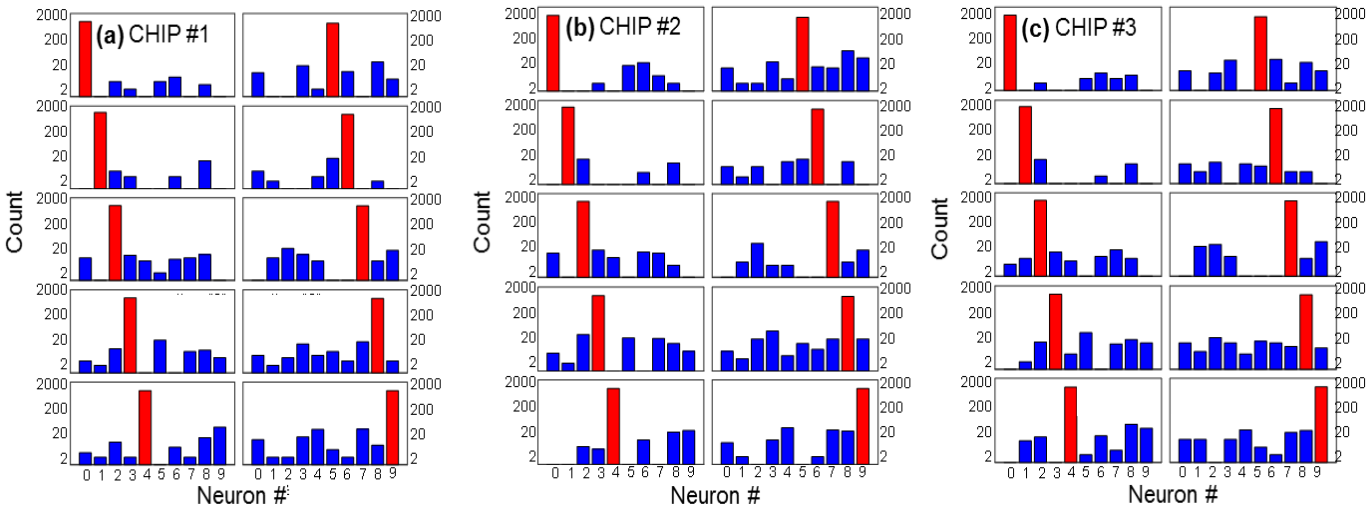

Fig. 3. Classification statistics: Histograms of the 10 output voltages for 1,000 MNIST test patterns of each class, for three different chips showing that the correct outputs (red bars) always dominate. Note the logarithmic vertical scales.

II. Variability, Time Stability, and Temperature Sensitivity

To evaluate chip-to-chip variations, we tested two more chips with the same classifier network. Their testing results are labeled by #2 and #3 below, while the data for chip #1 were taken from Ref. [4]. In all three cases, the memory cells' conductances were tuned with ~5% precision to the same target values, using an automated "write-verify" algorithm. As the data in Fig. 2 shows, the average actual tuning accuracy for these chips was, respectively, 4.4%, 5.6%, and 3.6%. Similarly to the original work (Fig. 2a), the currents of some cells, measured after the import of all weights, were outside of 5% tuning specifications, because of the half-select disturb effect and noise. The measured classification fidelity for these chips was 94.7%, 94.1%, and 94.2% (Fig. 3). Interestingly, the best average tuning accuracy in chip #3 did not result in the best classification performance - likely due to variations in other circuit parameters, such as opamp offsets.

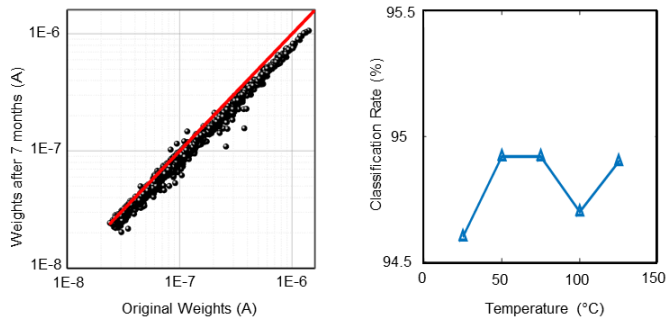

Fig. 4. Weight retention and temperature sensitivity of chip #1: (a) The original tuned weights vs. those measured 7 months later; (b) MNIST test set classification fidelity as a function of the ambient temperature.

Though a 10-year-long retention at temperatures up to 125˚C is guaranteed for digital NOR flash memory, analog circuits are naturally more prone to parameter drifts. As Fig. 4a shows, the cell conductances have slightly decreased (by 14% on the average) over a 7-month period. The conductance drift had a minor though noticeable impact on the output voltages of the circuit (Fig. 5); however, the classification fidelity remained unchanged at 94.7%.

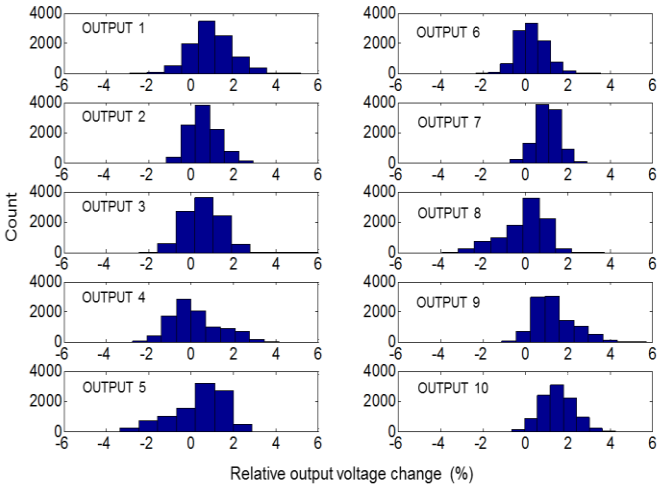

Fig. 5. Retention results: Histograms of the relative changes (drifts) of the output voltages for all 10,000 MNIST test set patterns, shown as a normalized difference between the originally measured values and those measured 7 months later.

Finally, we have measured the impact of the ambient temperature on the classification fidelity. Encouragingly, despite the exponential temperature dependence of subthreshold currents, used for the network operation [4], and no special efforts to make the circuit temperature-insensitive, the classification performance does not suffer (and actually slightly improves) at elevated temperatures (Fig. 4b).

These very encouraging results may be in part explained by the specific cost function used in training, whose goal was to

maximize the voltage difference between the correct and the second largest network output [4]. An additional contribution was probably given by the differential style of our gate-coupled design, and the fact that the changes in the cell conductances due to the drift in time and the temperature change are always in the same direction for all cells.

In summary, our results show no evident obstacles for the development of much more complex neuromorphic networks, based on commercial-grade floating-gate memory cells.

ACKNOWLEDGMENT

The authors are grateful to P.-A. Auroux, M. Bavandpour, N. Do, J. Edwards, M. Graziano, K. K. Likharev, and M. R. Mahmoodi for useful discussions and technical support.

REFERENCES

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

[2] S. Chakrabartty and G. Cauwenberghs, "Sub-microwatt analog VLSI trainable pattern classifier", *IEEE JSSC*, vol. 42, pp. 1169-1179, 2007.

[3] C. R. Schlottmann, and P. E. Hasler, "A highly dense, low power, programmable analog vector-matrix multiplier: The FPAA implementation," *IEEE JETCAS*, vol. 1 (3), 403-411, 2011.

[4] F. Merrikh Bayat *et al.*, "Sub-1-us, sub-20-nJ pattern classification in a mixed-signal circuit based on embedded 180-nm floating-gate memory cell arrays", ArXive 2016, available at https://arxiv.org/abs/1610.02091

[5] 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.

[6] 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.

[7] 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.