High-Performance Mixed-Signal Neurocomputing With Nanoscale Floating-Gate Memory Cell Arrays

Farnood Merrikh-Bayat, Xinjie Guo, Michael Klachko, Mirko Prezioso, Konstantin K. Likharev, *Fellow, IEEE*, and Dmitri B. Strukov, *Senior Member, IEEE*

***Abstract*— Potential advantages of analog- and mixed-signal nanoelectronic circuits, based on floating-gate devices with adjustable conductance, for neuromorphic computing had been realized long time ago. However, practical realizations of this approach suffered from using rudimentary floating-gate cells of relatively large area. Here, we report a prototype 28 × 28 binary-input, ten-output, three-layer neuromorphic network based on arrays of highly optimized embedded nonvolatile floating-gate cells, redesigned from a commercial 180-nm nor flash memory. All active blocks of the circuit, including 101 780 floating-gate cells, have a total area below 1 mm^2. The network has shown a 94.7% classification fidelity on the common Modified National Institute of Standards and Technology benchmark, close to the 96.2% obtained in simulation. The classification of one pattern takes a sub-1-μs time and a sub-20-nJ energy—both numbers much better than in the best reported digital implementations of the same task. Estimates show that a straightforward optimization of the hardware and its transfer to the already available 55-nm technology may increase this advantage to more than $10^2\times$ in speed and $10^4\times$ in energy efficiency.**

***Index Terms*— Deep learning, floating-gate memory cells, multilayer perceptron, neuromorphic networks, pattern classification.**

I. Introduction

THE concept of using nonvolatile memories in analog- and mixed-signal neuromorphic networks, far superior to digital circuits of the same functionality in speed and energy efficiency, is at least 30 years old [1]. Recent work [2]–[5] has shown that such circuits, utilizing nanoscale devices, may increase the neuromorphic network performance dramatically, leaving far behind their digital and biological counterparts, and approaching the energy efficiency of the human brain. The background of these advantages is the fact that in analog circuits, the vector-by-matrix multiplication, i.e., the key operation performed at signal propagation through any neuromorphic network, is implemented on the physical level, in a resistive crossbar circuit, using the fundamental Ohm and Kirchhoff laws (Fig. 1). On the other hand, the basic handicap of analog circuits, their finite precision, is typically not crucial in neuromorphic networks, due to the inherently high tolerance of their operation to synaptic weight variations [6].

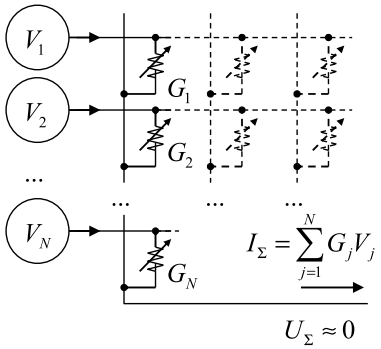


Fig. 1. Analog vector-by-matrix multiplication in a crossbar with adjustable crosspoint devices. For clarity, the output signal is shown for just one column of the array.

The key component of such mixed-signal neuromorphic networks is a device with adjustable (tunable) conductance—essentially an analog nonvolatile memory cell, mimicking the biological synapse. Up until recently, such devices were implemented mostly as floating-gate "synaptic transistors" [4], [7], which may be fabricated using the standard complimentary metal–oxide–semiconductor (CMOS) technology. Recently, some rather sophisticated neuromorphic systems were demonstrated [8], [9] using this approach. However, synaptic transistors have relatively large areas ($\sim 10^3 F^2$, where F is the minimum feature size), leading to larger time delays and energy consumption [4].

There have been significant recent advances in the development of alternative nanoscale nonvolatile memory devices, such as phase change, ferroelectric, and magnetic memories, and memristors—for a review see [10]–[14]. In particular, these emerging devices have already been used to demonstrate small neuromorphic networks [15]–[19]. However, their fabrication technology is still in much need for improvement and not ready yet for the large-scale integration, which is necessary for practically valuable neuromorphic networks.

In this paper, we describe a network prototype based on other alternative devices—the highly optimized, nanoscale, and nonvolatile floating-gate memory cells that are used in the recently developed embedded NOR flash memories [20].

Manuscript received February 9, 2017; revised July 21, 2017; accepted November 24, 2017. Date of publication December 22, 2017; date of current version September 17, 2018. This work was supported by the DARPA's UPSIDE Program under Contract HR0011-13-C-0051UPSIDE via BAE Systems, Inc. *(Farnood Merrikh Bayat and Xinjie Guo contributed equally to this work.) (Corresponding author: Dmitri B. Strukov.)*

F. Merrikh-Bayat, X. Guo, M. Klachko, M. Prezioso, and D. B. Strukov are with the Electrical Engineering Department, University of California, Santa Barbara, CA 93106-9560 USA (e-mail: strukov@ece.ucsb.edu).

K. K. Likharev is with the Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794-3800 USA (e-mail: Konstantin.Likharev@stonybrook.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2778940

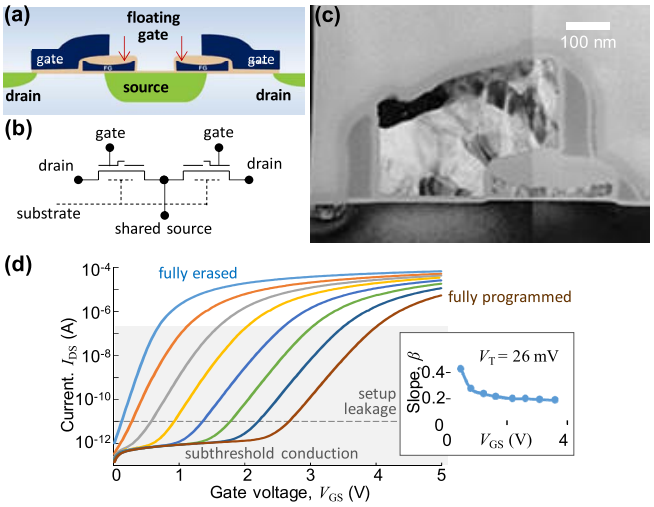


Fig. 2. ESF1 NOR flash memory cells. (a) Cross section of the two-cell "supercell" (schematically) and (b) its equivalent circuit. (c) TEM cross-sectional image of one memory cell, fabricated in a 180-nm process. (d) Drain current of the cell as a function of the gate voltage, at $V_{DS} = 1$ V, for several memory states. (d) Gray-shaded region shows the subthreshold conduction region; the currents below $I_{DS} = 10$ pA (the level shown with the dashed line) are significantly contributed by leakages in the experimental setup used for the measurements. Inset: extracted slope of this semilog plot, measured at $I_{DS} = 10$ nA, as a function of the memory state (characterized by the corresponding gate voltage).

These cells are quite suitable to serve as adjustable synapses in neuromorphic networks, provided that the memory arrays are redesigned to allow for individual, precise adjustment of the memory state of each device. Recently, such modification was performed [21], [22] using the 180-nm ESF1 embedded commercial NOR flash memory technology of SST Inc. [20] (Fig. 2), and, more recently, the 55-nm ESF3 technology of the same company [23], with good prospects for its scaling down to at least $F = 28$ nm. Though such modification nearly triples the cell area, it is still at least an order of magnitude smaller, in terms of F^2, than that of synaptic transistors [4].

The main result reported in this paper is the first successful use of this approach for the experimental implementation of a relatively simple mixed-signal neuromorphic network, which could perform a high-fidelity classification of patterns of the standard Modified National Institute of Standards and Technology (MNIST) benchmark, with record-breaking speed and energy efficiency.

II. Memory Array Characterization

Our network design uses the energy-saving gate coupling [4], [21], [23], [24] of the peripheral and array cells, which works well in the subthreshold mode, with a nearly exponential dependence of the drain current I_{DS} of the memory cell on the gate voltage V_{GS} [Fig. 2(d)]

$$I_{DS} \approx I_0 \exp\left\{\beta \frac{V_{GS} - V_t}{V_T}\right\} \quad (1)$$

where V_t is a threshold voltage depending on the memory state of the cell (physically, the electric charge of its floating gate), $V_T \equiv k_B T/e$ is the voltage scale of the thermal excitations, equal to ~26 mV at room temperature, while $\beta < 1$ is the dimensionless subthreshold slope $d(\ln I_{DS})/dV_{GS}$, measured in the units of V_T, and characterizing the efficiency of the gate-to-channel coupling. As the inset in Fig. 2(d) shows, in the ESF1 cells this slope stays relatively constant in a broad range of memory states—a feature enabling the gate-coupled circuit operation. [For lower V_t, the slope becomes higher, apparently due to the specific cell design shown in Fig. 2(a).]

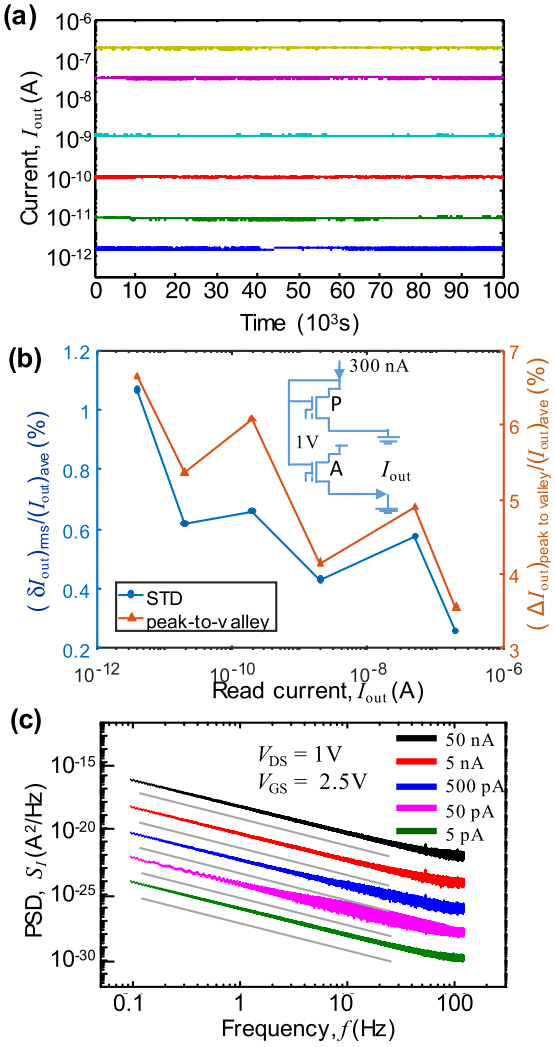


Fig. 3. (a) Results of analog retention measurements for several memory states, performed in the gate-coupled array configuration. There are 1000 points for each state, each point representing an average over 65 samples taken within a 130-ms period. (b) Relative rms variation and the full (peak-to-valley) swing of the currents during the same time interval. Inset: equivalent circuit of the used gate coupling. (c) Spectral density of cell current's noise measured at room temperature; the gray lines are just guides for the eye, corresponding to $S_I \propto 1/f^{1.6}$.

With the requirement to keep the relative current fluctuations [Fig. 3(b)] below 1%, the dynamic range of the subthreshold operation is about five orders of magnitude, from ~10 pA to ~300 nA, corresponding to the gate voltage swing of ~1.5 V.

The ESF1 flash technology guarantees a 10-year digital-mode retention at temperatures up to 125 °C [20]. Our experiments have shown that these cells also feature at least a-few-days analog-level retention, with very low fluctuations of the output current [see Fig. 3(a)]. (A more extensive testing of the analog-level retention [23], performed for the 55-nm

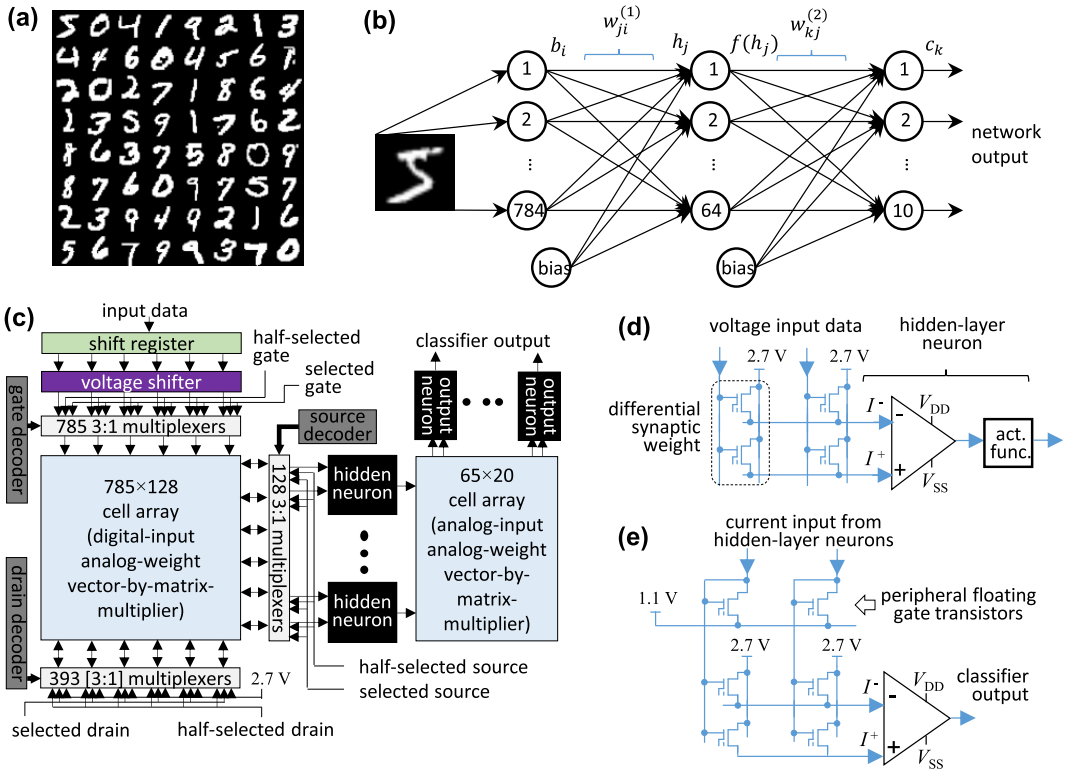


Fig. 4. Network architecture. (a) Typical examples of B/W hand-written digits of the MNIST benchmark set. (b) Graph representation of our three-layer perceptron network. Each synapse is implemented using a differential pair of floating-gate memory cells. (c) High-level architecture, with the weight tuning circuitry for the second array (similar to that of the first one) not shown for clarity. (d) 2 × 2 cell fragment of the first crossbar array shown together with a hidden-layer neuron, consisting of a differential summing operational amplifier pair and an activation-function circuit. (e) 2 × 2 cell fragment of the second crossbar array with an output-layer neuron; these neurons do not implement an activation function. The voltage shifter, shown in (c), enables using voltage inputs of both polarities over a 1.65-V bias, and is also used to initiate the classification process by increasing the input background from 1.8 to 4.2 V.

ESF3 NOR cells fabricated using a similar technology, has shown no substantial drift in memory states for almost 1 day even at an elevated temperature of 85 °C.)

Other features of the used ESF1 cell arrays, including the details of their modification, switching dynamics and statistics, and a demonstration of fast weight tuning with a ∼0.3% accuracy, were reported earlier [21], [22]

III. Network Design

For the first, proof-of-concept demonstration of this new hardware technology, we have selected the simplest possible neuromorphic network architecture suitable for classification of the most common MNIST benchmark set, with a reasonable fidelity. The binary inputs of this benchmark simplify the design of the first synaptic array. The implemented network (Fig. 4) was a three-layer (one hidden layer) perceptron with 784 binary inputs b_i, which may represent, for example, 28 × 28 black-and-white pixels of an input image [such as the MNIST data set images illustrated in Fig. 4(a)], 64 hidden layer neurons with the rectified-tanh activation function, and ten output neurons [Fig. 4(b)]. The goal of the network is to perform the pattern inference by the following sequential transformation of the input signals.

Here, h_j and f_j (with $j = 1, 2,\ldots, 64$) are, respectively, the input and output signals of the hidden-layer neurons, c_k (with $k = 1, 2,\ldots, 10$) are the output signals, providing the class of the input pattern, while $w^{(1)}$ and $w^{(2)}$ are two matrices of tunable synaptic weights, characterizing the coupling of the adjacent network layers. In our network, these weights are provided by floating-gate cells of two crossbar arrays of the floating-gate memory cells with tunable weights [Fig. 4(c)]. Each neuron also gets an additional input from a bias node, with a tunable weight based on a similar cell [Fig. 4(b)]. With the differential-pair implementation of each synapse (see below), the total number of utilized floating-gate memory cells is $2 \times [(28 \times 28 + 1) \times 64 + (64 + 1) \times 10] = 101\ 780$.

The mixed-signal vector-by-matrix multiplication in the first crossbar array is implemented by applying input voltages (4.2 V for black pixels or 0 V for white ones) directly to the gates of the array cell transistors, with fixed voltages on

$$h_j = \sum_{i=1}^{784} w_{ji}^{(1)} b_i + w_{j,785}^{(1)}, \quad c_k = \sum_{j=1}^{64} w_{kj}^{(2)} f(h_j) + w_{k,65}^{(2)} f_{\max}, \quad f(h) \equiv f_{\max} \times \begin{cases} \tanh(h), & \text{for } h \geq 0, \\ 0, & \text{for } h < 0 \end{cases} \tag{2}$$

their sources (1.65 V) and drains (2.7 V) [see Fig. 4(d)]. As a result, the transistor source-to-drain current of the cell located at the crosspoint of the ith column and the jth row of the array does not depend on the state of any other cells, and is equal to the product of the binary input voltage b_i by the analog weight $w_{ji}^{(1)}$ prerecorded in the memory cell. The sources of the transistors of each row are connected to a single wire (with an externally fixed voltage on it), so that the jth output current of the array is just the sum of products $w_{ji}^{(1)} b_i$ over all columns i, thus implementing the vector-by-matrix multiplication described by the first expression of (2), as shown at the bottom of the previous page.

In order to reduce random drifts, and also to work with zero-centered signals h_j, we used a differential scheme, in which each synaptic weight is recorded in two adjacent cells of each column, and the output currents [Fig. 4(d), I_j^+ and I_j^-] of two adjacent cell rows are subtracted in an operational amplifier, with its output, $h_j \propto I_j^+ - I_j^-$, passed to the activation function circuit performing the function $f(h)$. The used sharing of the weight $w_{ji}^{(1)}$ between the two cells of the differential pair is very simple: one of the cells (depending on the sign of the desirable weight) is completely turned OFF, giving virtually no contribution to the output current. This arrangement keeps half of the cells virtually idle, but simplifies the design and speeds up the weight tuning process.

The analog vector-by-matrix calculation in the second array was performed using the gate-coupled approach [Fig. 4(e)]. In this approach [24], the synaptic gate array is complemented by the additional row of "peripheral" cells, which are physically similar to the array cells, and hence having the same subthreshold slope β. The gate electrode of the peripheral cell of each column is connected to those of all cells of this column, so that their voltages V_{GS} are also equal. Applying (1) to the current of the cell located at the crosspoint of the kth row and the jth column of the array (I_{kj}), and that of the peripheral cell of this column (I_j), and dividing the results, we get

$$w_{kj}^{(2)} \equiv \frac{I_{kj}}{I_j} = \exp\left\{\beta \frac{(V_t)_j - (V_t)_{kj}}{V_T}\right\}. \tag{3}$$

The resulting currents I_{kj} are summed up exactly as those in the first array (with the similar differential scheme for drift reduction), so that if the array is fed by the output currents of the activation function circuits, $I_j \propto f(h_j)$, it performs the vector-by-matrix multiplication described by the second expression of (2), with the synaptic weights given by (3), which depend on the preset memory states of the corresponding cells, but are independent of the input currents. To minimize the error due to the dependence of β on the memory state [see the inset in Fig. 2(d)], in the second array, we used a higher gate voltage range (1.1–2.7 V), with the upper bound due to the technology restrictions.

Fig. 5(a) shows the circuit used to subtract the currents I^+ and I^- of the differential-scheme rows, based on two operational amplifiers [Fig. 5(c)]. Assuming that the resistances R_F are equal and that the outputs of both opamps do not saturate (which was ensured by the following relation for the maximum value of currents $I^{\pm}$: $I_{max} R_F < 1$ V for the chosen

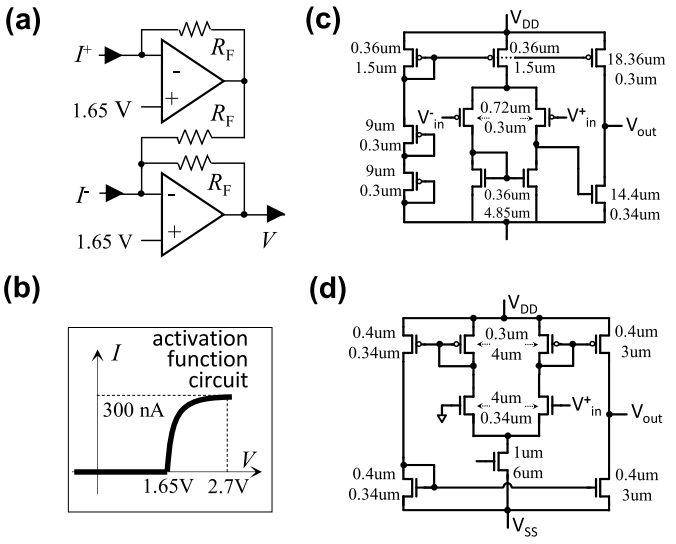


Fig. 5. (a) Circuit-level diagram of a differential summing amplifier used in the hidden-layer and output-layer neurons; $R_F = 16$ kΩ for hidden neurons and $R_F = 128$ kΩ for output neurons. (b) Implemented activation function. Transistor-level schematics of (c) operational amplifier and (d) activation function, $V_{SS} = 0$ V, $V_{DD} = 2.7$ V.

value, $R_F = 16$ kΩ in the first layer and $R_F = 128$ kΩ in the second one), the output voltage of the scheme is

$$V = R_F(I^+ - I_-) + \text{const} \tag{4}$$

Fig. 5(b) shows the rectified-tanh activation function $f(h)$ used in the hidden-layer neurons [see (2)], with h [V] $= 10R_F[\Omega]$ $(I^+ - I^-)$ [A] and $f_{max} = 300$ nA, while Fig. 5(d) shows the CMOS circuit used for the implementation of this function.

The desirable synaptic weights, calculated in an external computer running a similar "precursor" software-implemented network, using the standard error backpropagation algorithm, were imported into the network by analog tuning of the memory state of each floating-gate cell, with peripheral analog demultiplexer circuitry [Fig. 3(c)]. In order to simplify this first, prototype design, the weights were tuned one by one, by applying proper bias voltage sequences to select and half-select lines [21], [22]. (In principle, this process may be significantly parallelized.) The large voltages required for the weight import are decoupled from the basic, low-voltage circuitry, using high-voltage pass transistors. The input pattern bits are shifted serially into a 785-b register before each classification; to start it, the bits are read out into the network in parallel.

The digital encoders and shift register circuits and their layouts were synthesized from Verilog in a standard 1.8-V digital CMOS process. All other circuits were designed manually for the embedded 180-nm process of Silterra Corp. (Such an approach was practicable due to the modular, repetitive design of the circuit.) All active components of the circuit have a total area of 0.78 mm^2 (Fig. 6), with the two synaptic arrays occupying less than a quarter of this area, while the total chip area, including very sparse routing (which was not yet optimized for this design), is about 5 × 5 mm^2.

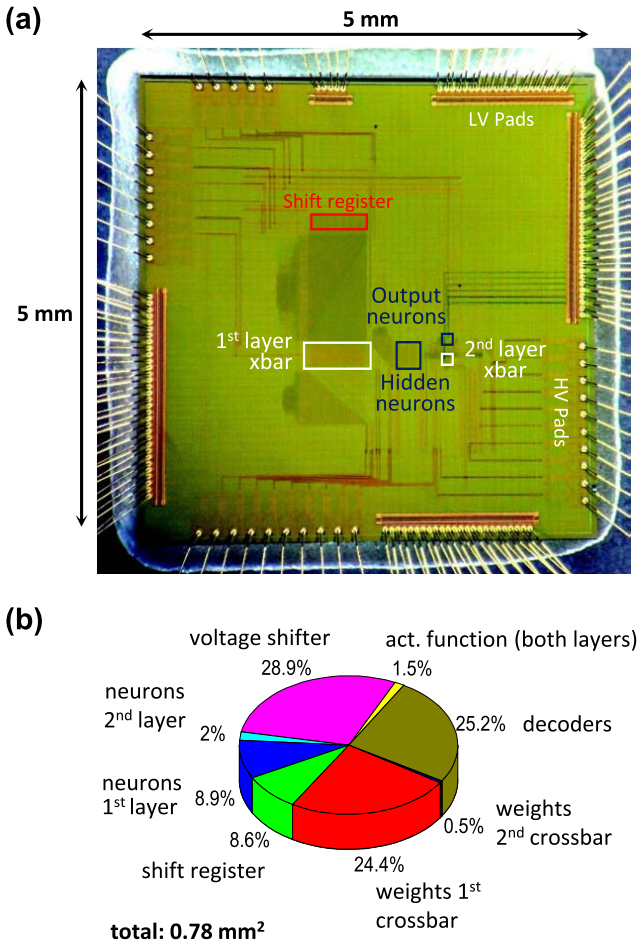


Fig. 6. (a) Micrograph of the chip and (b) area breakdown of its active components (excluding wiring between the blocks, which was not optimized at this stage).

IV. Network Testing

Because of the digital (fixed voltage) input of the first synaptic array, the subthreshold conduction was not enforced there, so that the output currents of some cells exceeded 300 nA [Fig. 7(a)]. To reduce the computation error due to the potential slope mismatch between peripheral and array cells, all peripheral floating gate transistors in the second array were tuned to provide output currents of 300 nA at $V_G = 2.7$ V, i.e., at the largest voltage that could be supplied by the hidden layer neuron in our design. With such a scheme, the error is conveniently smallest for the largest weight $w_{ki} = 1$, corresponding to the array cell tuned to run a current of 300 nA at $V_{GS} = 1.6$ V. The target current values for all cells in the second array (excluding bias ones) were ensured to be between 0 and 300 nA by clipping the weights during training of the precursor network.

To decrease the weight import time, only one cell of each pair, corresponding to a particular sign of the weight value, was tuned, while its counterpart was kept at a very small, virtually zero, initial conductance. Additionally, all nonbias cells in the first array, for which the target conductances were below 30 nA, were also not tuned, because of their negligible impact on the classification fidelity, confirmed by modeling. As a result, only about ∼30% of the cells were fine tuned. Because of the sequential character of the tuning process, it took several hours to complete it, with the chosen accuracy, for the whole chip. (In the future, the tuning may be greatly sped up by adjusting multiple weights at a time via integrated on-chip tuning circuitry [9], and using the better tuning algorithms we have developed [22].)

Moreover, also to speed up the import process, the weight tuning accuracy for a single-cell tuning was set to a relatively high value of 5%. As Fig. 7(b) indicates, some of the already tuned cells were disturbed beyond the target accuracy during the subsequent weight import. In this first experiment, these cells were not retuned, in part because even for such rather crude weight import the experimentally tested classification fidelity (94.65%) on MNIST benchmark test patterns (Fig. 8) is already remarkably close to the simulated value (96.2%) for the same network (Fig. 9). Both these numbers are also not too far from the maximum fidelity (97.7%) of the similar perceptron of this size, optimized without hardware constraints, with ∼0.5% in fidelity recovered by taking into account small weights and ∼1% by not clipping the nonbias weights.

Excitingly, such classification fidelity in our network, with large optimization reserves (see below), is achieved at an ultralow (sub-20-nJ) energy consumption per average classified pattern [Fig. 10(a)], and the average classification time below 1 μs [Fig. 10(b)]. The upper bound of the energy is calculated as a product of the measured average power, 5.6 mA × 2.7 V + 2.9 mA × 1.05 V ≈ 20 mW, consumed by the network, by the upper bound, 1 μs, of the average signal propagation delay. A more accurate measurement of the time delay, and hence the energy, requires a redesign of the signal input circuitry, currently rather slow [see Fig. 10(b)].

V. Discussion

The achieved speed and energy efficiency are much better than those demonstrated, for the same task, at any digital network we are aware of. For example, the best results for the same MNIST benchmark classification were reported for IBM's TrueNorth chip [25]. For the comparable 95% fidelity, that chip can classify 1000 images per second while consuming 4 μJ of energy per image [26], i.e., it is at least three orders of magnitude slower and less energy-efficient than our, still unoptimized analog circuit. This difference is rather impressive, taking into account the advanced 28-nm CMOS process used for the TrueNorth chip implementation.

In a less direct comparison, in terms of energy per a multiply-and-accumulate (MAC) operation, our network also outperforms the best reported digital systems. Indeed, the measured upper bound of the energy efficiency of our circuit is 0.2 pJ per MAC. This is a factor of 60 smaller than the 12 pJ per MAC reported for 65-nm Eyeriss chip [27], which is highly optimized for machine learning applications. (It performs 16-b operations and, like the TrueNorth chip, was implemented using an advanced fabrication technology.) Note that both the TrueNorth and Eyeriss chips, in turn, far outperform the modern graphical processing units (GPUs) for neuromorphic-network applications. Our result is also much

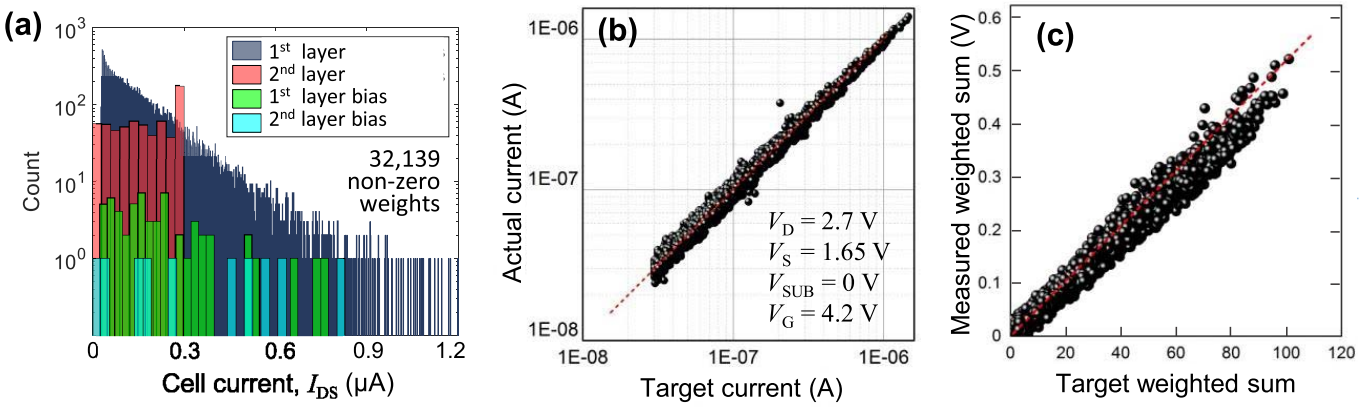


Fig. 7. Weight export statistics. (a) Histogram showing the imported cell current values (weights), measured at $V_D = 2.7$ V, and $V_S = 1.65$ V and $V_G = 4.2$ V in the first synaptic array, and $V_S = 1.1$ V and $V_G = 2.7$ V for the second one, which were used in the experiment. (b) Comparison between the target synaptic cell currents (computed at the external network training) and the actual cell currents measured after their import, i.e., cell tuning. (c) Similar comparison for the positive fraction of hidden neuron output computed for all test patterns. (The negative outputs are not shown, because they are discarded by the used activation function.) Red-dashed lines are guides for the eye, corresponding to the perfect weight import.

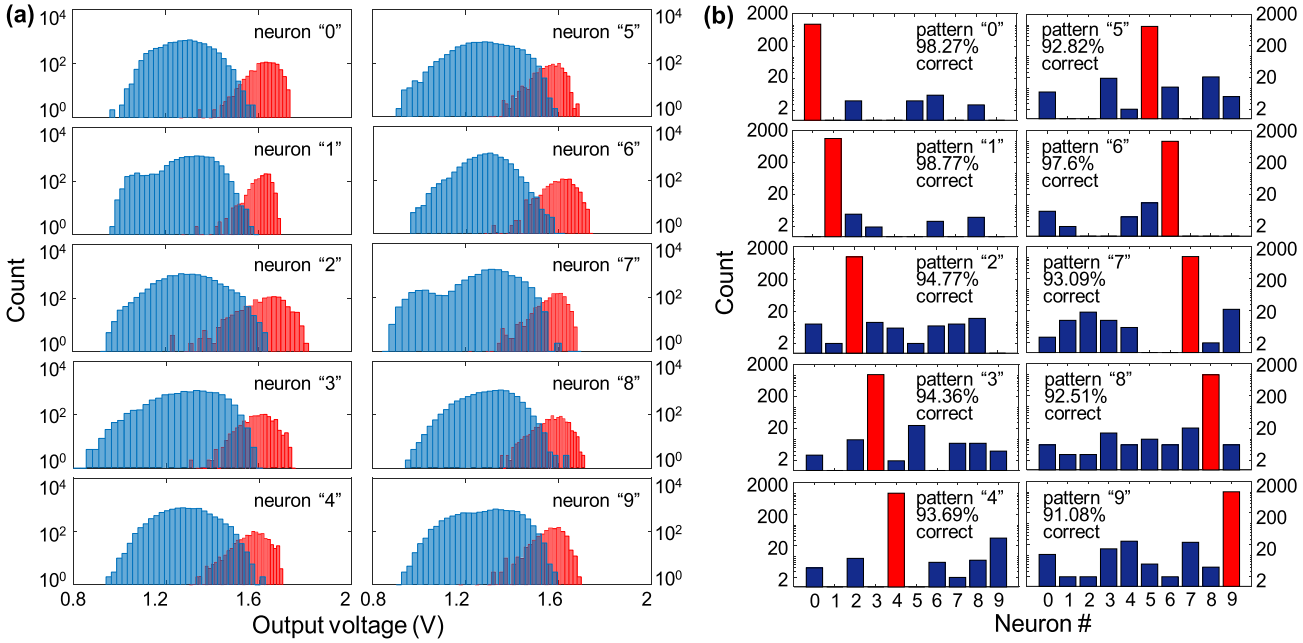


Fig. 8. Experimental results for the classification of all 10 000 MNIST test set patterns. (a) Histograms of voltages delivered by each output neuron. Red bars correspond to the patterns whose class belongs to this particular output, while the blue ones are for all remaining patterns. (b) Histograms of the largest output voltages (among all output neurons) for all test patterns of each class, showing that the correct outputs (red bars) always dominate. Note the logarithmic vertical scales.

better than the ~1 pJ per analog operation, recently reported for a small 130-nm mixed-signal neural networks based on synaptic transistors [8]. It is also comparable with the best results obtained using the switched-capacitor approach [28], for example, the recent ~0.1 pJ per operation achieved in a much smaller circuit, with only $8 \times 8 \times 3$ discrete (3-b) synaptic weights, using a 40-nm process [29]. (Note that this approach does not allow analog tuning of synaptic weights, and its extension to larger circuits may be problematic because of the relatively large capacitor size.)

It should be also noted that the energy-per-MAC metric is generally less objective, because it does not account for the operation precision and the complexity and functionality of the implemented system (e.g., general-purpose systems like a typical GPU versus application-specific ones like the Eyeriss chip).

There are still several unused reserves in our design. The most straightforward improvement is to use for neurons the current-mirror design similar to the gate-coupled circuits shown in Fig. 4(e), but implemented with the floating-gate-free transistors, and hence with the signal transfer weight $w = 1$. (In our current design, neurons give dominant contributions to the network latency and energy dissipation [see Fig. 10(a)]). The second direct path forward is to use the more advanced 55-nm memory technology ESF3 of the same company [20]. (Our preliminary testing [23] of its similar redesign has not

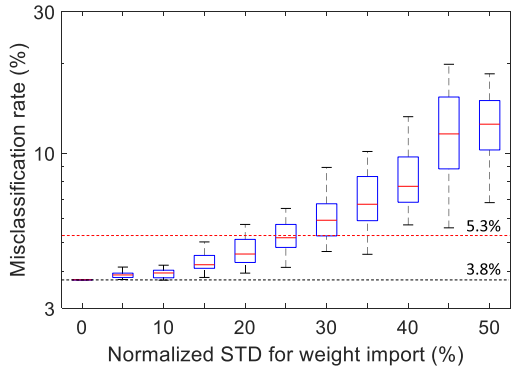


Fig. 9. Simulated classification fidelity, computed with the 32-b floating-point precision, as a function of weight precision import for the implemented network, with the particular set of weights used in the experiment. The weight error was modeled by adding, to its optimized value, a normally distributed noise with the shown standard deviation. The red, blue (rectangles), and black (segment) markers denote, respectively, the 25%–75% percentile, and the minimum and maximum values for 30 simulation runs. The black and red horizontal dashed lines show, respectively, the calculated misclassification rate for perfect (no noise) weights, and the rate obtained in the experiment.

found any evident showstoppers on that path.) The time delay and energy dissipation of the network with current-mirror neurons will be dominated by the synaptic arrays, and may be readily estimated using the experimentally measured values of the subthreshold current slope β for 180-nm ESF1 cells and 55-nm ESF3 cells. For example, our modeling of a large-scale network deep-learning convolutional networks, suitable for classification of large, complex patterns [30] (i.e., the same network which was implemented by Eyeriss chip [27]), using these two improvements, showed at least a $\sim 100\times$ advantage in the operation speed, and an enormous, $>10^4\times$ advantage in energy efficiency, over the state-of-the-art purely digital (GPU and custom) circuits [see Table I]. (In this table, the estimates for the floating-gate networks take into account the $55 \times 55 = 3025$-step time-division multiplexing, natural for this particular network. The crude estimate of the human visual cortex operation is based on the ~25-W power consumption of $\sim 10^{11}$ neurons of the whole brain, and a 30-ms delay of the visual cortex, and assumes the uniform distribution of the power over the neurons, and the same number of neurons participating in a single-pattern classification process.) Moreover, the energy efficiency of the floating-gate networks would closely approach that of the human visual cortex, at much higher speed [see the last two columns in Table I].

It should be also noted that the cell area of sub-100-nm embedded NOR floating-gate memories is only slightly larger than that of the "1T1R" variety of many emerging nonvolatile memory technologies [10]–[15]. (Here, "T" stands for a dedicated select transistor and "R" for the adjustable resistive memory element.) On the other hand, our crude estimates show [5] that the density and performance may be significantly improved using truly passive ("0T1R") memristor circuits and especially their 3-D versions [31].

Note also that the recent progress [32], [33] in the development of machine learning algorithms using binary weights implies that our approach may be also extended to novel

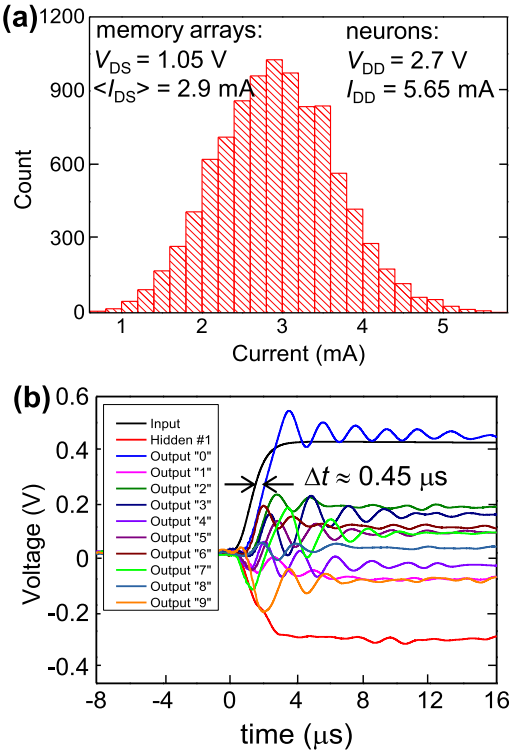


Fig. 10. Physical performance. (a) Histogram of the experimentally measured total currents flowing into the circuit, characterizing the static power consumption of the both memory cell arrays, for all patterns of the MNIST test set. Inset: list of the pattern-independent static current of the neurons. (b) Typical signal dynamics after an abrupt turn ON of the voltage shifter power supply, measured simultaneously at the network input, the output of a sample hidden-layer neuron, and all network's outputs. (The actual input voltage is 10× larger.) The oscillatory behavior of the outputs is a result of a suboptimal phase stability design of the operational amplifiers. Before it has been improved, and the input circuit is sped up, we can only claim a sub-1-μs average time delay of the network, though it is probably closer to 0.5 μs.

TABLE I

SPEED AND ENERGY CONSUMPTION OF THE SIGNAL PROPAGATION THROUGH THE CONVOLUTIONAL (DOMINATING) PART OF A LARGE DEEP NETWORK [30]

AlexNet [30] single pattern classification:	Digital circuits [27]		Mixed-signal floating-gate circuits (estimates)		Visual cortex (crude estimates)
	GPU 28 nm	ASIC 65 nm	ESF1 180 nm	ESF3 55 nm	
time (s)	1.5×10^{-2}	2.9×10^{-2}	$\sim1\times10^{-4}$	$\sim6\times10^{-5}$	$\sim3\times10^{-2}$
energy (J)	1.5×10^{-1}	0.8×10^{-2}	$\sim3\times10^{-7}$	$\sim2\times10^{-7}$	$\sim5\times10^{-8}$

3-D NAND flash technologies. Such memories may ensure much higher areal densities of the floating-gate cells, but their redesign to analog weights may be more problematic. Also, the results of [33] show a significant drop in classification performance that results from using binary weights in convolutional layers of large-scale neuromorphic networks. The performance of such networks may be improved by increasing the network size; however, its speed and energy efficiency

may suffer. So, the tradeoff between the density and weight precision effects in 3-D memories is far from certain yet, and requires further study.

To summarize, we believe that the reported results give an important proof-of-concept demonstration of the exciting possibilities opened for neuromorphic networks by mixed-signal circuits based on industrial-grade floating-gate memory cells.

ACKNOWLEDGMENT

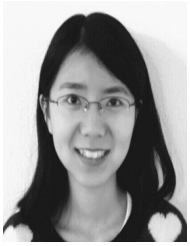
The authors would like to thank P.-A. Auroux, M. Bavandpour, N. Do, J. Edwards, M. Graziano, and M. R. Mahmoodi for their useful discussions and technical support.

REFERENCES

[1] C. Mead, *Analog VLSI and Neural Systems*. Reading, MA, USA: Addison-Wesley, 1989.
[2] G. Indiveri *et al.*, "Neuromorphic silicon neuron circuits," *Frontiers Neurosci.*, vol. 5, no. 73, pp. 1–23, 2011.
[3] K. Likharev, "CrossNets: Neuromorphic hybrid CMOS/nanoelectronic networks," *Sci. Adv. Mater.*, vol. 3, pp. 322–331, Jun. 2011.
[4] J. Hasler and H. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers Neurosci.*, vol. 7, Sep. 2013, Art. no. 118.
[5] L. Ceze *et al.*, "Nanoelectronic neurocomputing: Status and prospects," in *Proc. DRC*, Newark, DE, USA, Jun. 2016, pp. 1–2.
[6] E. Säckinger, "Measurement of finite-precision effects in handwriting- and speech-recognition algorithms," in *Artificial Neural Networks—ICANN* (Lecture Notes in Computer Science), vol. 1327. Springer, 1997, pp. 1223–1228.
[7] C. Diorio, P. Hasler, A. Minch, and C. A. Mead, "A single-transistor silicon synapse," *IEEE Trans. Electron Devices*, vol. 43, no. 11, pp. 1972–1980, Nov. 1996.
[8] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015.
[9] S. George *et al.*, "A programmable and configurable mixed-mode FPAA SoC," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 6, pp. 2253–2261, Jun. 2016.
[10] D. B. Strukov and H. Kohlstedt, "Resistive switching phenomena in thin films: Materials, devices, and applications," *MRS Bull.*, vol. 37, no. 2, pp. 108–114, Feb. 2012.
[11] S. Raoux, D. Ielmini, M. Wuttig, and I. Karpov, "Phase change materials," *Mater. Res. Soc. Bull.*, vol. 37, no. 2, pp. 118–123, 2012.
[12] W. Lu, D. S. Jeong, M. Kozicki, and R. Waser, "Electrochemical metallization cells—Blending nanoionics into nanoelectronics?" *MRS Bull.*, vol. 37, no. 2, pp. 124–130, 2012.
[13] J. J. Yang, I. H. Inoue, T. Mikolajick, and C. S. Hwang, "Metal oxide memories based on thermochemical and valence change mechanisms," *MRS Bull.*, vol. 37, no. 2, pp. 131–137, 2012.
[14] E. Y. Tsymbal, A. Gruverman, V. Garcia, M. Bibes, and A. Barthélémy, "Ferroelectric and multiferroic tunnel junctions," *MRS Bull.*, vol. 37, no. 2, pp. 138–143, 2012.
[15] S. Park *et al.*, "RRAM-based synapse for neuromorphic system with pattern recognition function," in *IEDM Tech. Dig.*, Dec. 2012, pp. 10.2.1–10.2.4.
[16] Y. Nishitani, Y. Kaneko, and M. Ueda, "Supervised learning using spike-timing-dependent plasticity of memristive synapses," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 2999–3008, Dec. 2015.
[17] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, pp. 61–64, May 2015.
[18] S. Kim *et al.*, "NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning," in *IEDM Tech. Dig.*, Dec. 2015, pp. 443–446.
[19] F. Merrikh-Bayat, M. Prezioso, B. Chakrabarti, I. Kataeva, and D. B. Strukov. (Nov. 2016). "Advancing memristive analog neuromorphic networks: Increasing complexity, and coping with imperfect hardware components." [Online]. Available: https://arxiv.org/abs/1611.04465
[20] SST Inc. *Superflash Technology Overview*. Accessed: Feb. 7, 2017. [Online]. Available: www.sst.com/technology/sst-superflash-technology
[21] F. Merrikh-Bayat, X. Guo, H. A. Om'mani, N. Do, K. K. Likharev, and D. B. Strukov, "Redesigning commercial floating-gate memory for analog computing applications," in *Proc. ISCAS*, Lisbon, Portugal, May 2015, pp. 1921–1924.
[22] F. Merrikh-Bayat, X. Guo, M. Klachko, N. Do, K. Likharev, and D. Strukov, "Model-based high-precision tuning of NOR flash memory cells for analog computing applications," in *Proc. DRC*, Newark, DE, USA, Jun. 2016, pp. 1–2.
[23] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in *Proc. CICC*, Austin, TX, USA, Apr./May 2017, pp. 1–4.
[24] C. R. Schlottmann and P. E. Hasler, "A highly dense, low power, programmable analog vector-matrix multiplier: The FPAA implementation," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 1, no. 3, pp. 403–411, Sep. 2011.
[25] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.
[26] S. K. Esser, R. Appuswamy, P. Merolla, J. V. Arthur, and D. S. Modha, "Backpropagation for energy-efficient neuromorphic computing," in *Proc. NIPS*, Montreal, QC, Canada, Dec. 2015, pp. 1117–1125.
[27] Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Jan./Feb. 2016, pp. 262–263.
[28] D. Bankman and B. Murmann, "Passive charge redistribution digital-to-analogue multiplier," *Electron. Lett.*, vol. 51, no. 5, pp. 386–388, 2015.
[29] E. H. Lee and S. S. Wong, "A 2.5 GHz 7.7 TOPS/W switched-capacitor matrix multiplier with co-designed local memory in 40 nm," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Jan./Feb. 2016, pp. 418–420, 2016.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, Lake Tahoe, CA, USA, Dec. 2012, pp. 1097–1105.
[31] G. C. Adam *et al.*, "3-D memristor crossbars for analog and neuromorphic computing applications," *IEEE Trans. Electron Devices*, vol. 64, no. 1, pp. 312–318, Jan. 2017.
[32] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training deep neural networks with binary weights during propagations," in *Proc. NIPS*, Montreal, QC, Canada, Dec. 2015, pp. 3105–3113.
[33] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. (Sep. 2016). "Quantized neural networks: Training neural networks with low precision weights and activations." [Online]. Available: https://arxiv.org/abs/1609.07061

Farnood Merrikh-Bayat received the M.Sc. and Ph.D. degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2008, and 2012, respectively, and the Ph.D. degree in computer engineering from the University of California, Santa Barbara, CA, USA, in 2015.

He was a Post-Doctoral Researcher in computer engineering with the University of California until 2017. His current research interests include abusing emerging device technologies to efficiently implement hardware accelerators for deep neural networks and machine learning algorithms, high-performance computing with beyond complimentary metal–oxide–semiconductor, nonconventional architectures and designs, and smart systems and vehicles.



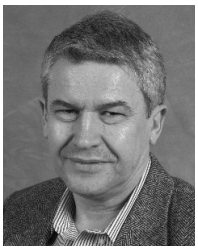
Xinjie Guo recievied the B.S. degree in electronics engineering and computer science from Peking University, Beijing, China, in 2011, the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, CA, USA, in 2013 and 2016, respectively.

Michael Klachko received the M.S. degree in computer engineering from the University of California, Santa Barbara, CA, USA, in 2016, where he is currently pursuing the Ph.D. degree, as a part of Prof. Strukov's Research Group.

His current research interests include novel deep learning architectures and algorithms, and low-precision neural computation.

Mirko Prezioso received the master's degree in condensed matter physics and the Ph.D. degree in advanced materials from the University of Parma, Parma, Italy, in 2008.

He was a Post-Doctoral Researcher with the National Research Council, Institute for the Study of Nanostructured Materials, Bologna, Italy, where he focused on organic spintronics and memristive systems. In 2013, he joined the Electrical and Computing Engineering Department, University of California, Santa Barbara, CA, USA, as a Research Assistant under the supervision of Prof. Strukov, where he was involved in the fabrication, characterization, and application of oxide-based RRAM devices. His current research interests include the application for neuromorphic computing.



Konstantin K. Likharev (M'91–SM'06–F'08) received the Ph.D. degree in physics from the Lomonosov Moscow State University, Russia, in 1969, and a habilitation degree of doctor of sciences from USSR's High Attestation Committee in 1979.

From 1969 to 1988, he was a Staff Scientist with Moscow State University, where he was the Head of the Laboratory for Cryoelectronics, from 1989 to 1991. In 1991, he assumed a Professorship at Stony Brook University, Stony Brook, NY, USA, where he has been a Distinguished Professor since 2002. He was involved in the fields of nonlinear classical and dissipative quantum dynamics, and solid-state physics and electronics, notably including superconductor electronics and nanoelectronics. His current research interests include the architecture and nanoelectronic implementation of high-performance neuromorphic networks. He has authored more than 350 original publications, more than 75 review papers and book chapters, and two monographs, and holds several patents.

Prof. Likharev is an APS Fellow. In 1979, he joined the USSR Highest Attestation Committee.

Dmitri B. Strukov (M'02–SM'16) received the M.S. degree in applied physics and mathematics from the Moscow Institute of Physics and Technology, Moscow, Russia, in 1999, and the Ph.D. degree in electrical engineering from Stony Brook University, Stony Brook, NY, USA, in 2006.

From 2006 to 2007, he was a Post-Doctoral Associate with Stony Brook University. From 2007 to 2009, he was with Hewlett Packard Laboratories, Palo Alto, CA, USA, where he was involved in various aspects of nanoelectronic systems. He is currently a Professor of electrical and computer engineering with the University of California, Santa Barbara, CA, USA. His current research interests include different aspects of computation, in particular addressing questions on how to efficiently perform computation on various levels of abstraction and hardware implementations of artificial neural networks with emerging memory devices.