Fast, Energy-Efficient, Robust, and Reproducible Mixed-Signal Neuromorphic Classifier Based on Embedded NOR Flash Memory Technology

X. Guo[1*], F. Merrikh Bayat[1*], M. Bavandpour[1], M. Klachko[1], M. R. Mahmoodi[1], M. Prezioso[1], K. K. Likharev[2#], and D. B. Strukov[1&]

[1]UC Santa Barbara, Santa Barbara, CA 93106-9560, U.S.A., [2]Stony Brook University, Stony Brook, NY 11794-3800, U.S.A.

*these authors contributed equally to this work, #Konstantin.Likharev@stonybrook.edu, &strukov@ece.ucsb.edu

***Abstract*---**We have designed, fabricated, and tested a prototype mixed-signal, 28×28-binary-input, 10-ouput, 3-layer neuromorphic network based on embedded nonvolatile floating-gate cell arrays redesigned from a commercial 180-nm NOR flash memory. Each array performs a very fast and energy-efficient analog vector-by-matrix multiplication, which is the bottleneck for signal propagation in neuromorphic networks. All functional components of the prototype circuit, including 2 synaptic arrays with 101,780 floating-gate synaptic cells, 74 analog neurons, and the peripheral circuitry for weight adjustment and I/O operations, have a total area below 1 mm^2. Its testing on the MNIST benchmark set has shown a classification fidelity of 94.65%, close to the 96.2% obtained in simulation. The classification of one pattern takes <1 μs time and ~20 nJ energy – both numbers $>10^3\times$ better than those of the 28-nm IBM TrueNorth digital chip for the same task at a similar fidelity. Estimates show that this performance may be further improved using a better neuron design and a more advanced memory technology, leading to a $>10^2\times$ advantage in speed and a $>10^4\times$ advantage in energy efficiency over the state-of-the-art purely digital circuits for classification of large, complex patterns. Experimental results for the chip-to-chip statistics, long-term drift, and temperature sensitivity show no evident showstoppers on the way toward practical deep neuromorphic networks with unprecedented performance.

I. Introduction

The idea of using nonvolatile floating-gate memory cells in analog neuromorphic networks has been around for almost 30 years [1]. Fig. 1a shows the concept of fast and energy-efficient analog vector-by-matrix multiplication (VMM) - the bottleneck for signal propagation in neuromorphic networks. Up until recently, implementations of such circuits were based on "synaptic transistors" [2-4], which can be fabricated using the standard CMOS technology. While sophisticated, energy-efficient systems have been demonstrated [3, 4] using this approach, synaptic transistors have relatively large areas ($\sim 10^3 F^2$, where F is the minimum feature size [4]), leading to larger time delays and energy consumption.

Fortunately, by now the nonvolatile floating-gate memory cells have been highly optimized and scaled down all the way to $F \sim 20$ nm, and may be embedded into CMOS integrated circuits [5]. 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 for the 180-nm ESF1 [6, 7] (Fig. 2) and the 55-nm ESF3 [8] embedded commercial NOR flash memory technology of SST Inc. [7], 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 10× smaller, in terms of F^2, than that of synaptic transistors [4]. In these terms, the modified cell area is also comparable to that of 1T1R-cells of emerging nonvolatile (e.g., memristive) technologies [9]. Moreover, due to the internal gain of their cells, floating-gate arrays for the analog VMM (Fig. 1a) have an additional advantage over memristive arrays (Fig. 1b), in terms of the necessary gain and energy consumption of the peripheral circuitry.

The main result reported in this paper is an experimental demonstration of a reproducible, stable, and robust neuromorphic network, based on redesigned floating-gate memory arrays, which can perform high-fidelity classification of images of the standard MNIST benchmark, with record-breaking speed and energy efficiency.

II. Network Design

Our design uses the energy-saving gate coupling [1, 4] 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. 3). The sub-1% current fluctuations across a $\sim 10^5\times$ dynamic range of the subthreshold operation (Figs. 3, 4a,c) and low variations of the program and erase voltages (Fig. 5) enable precise VMM operation of such gate-coupled arrays (Fig. 4b). (Other features of the modified ESF1/3 cell arrays, including their long-term analog retention and fast weight tuning with a ~0.3% accuracy, were reported in Refs. [6-8].)

The implemented neuromorphic network is a 3-layer (one-hidden-layer) MLP perceptron with 784 inputs, representing 28×28 black-and-white pixels of the input patterns, 64 hidden layer neurons with the rectify-tanh activation function, and 10 output neurons (Fig. 6). The differential synaptic coupling between the neuron layers is provided by two crossbar arrays with the total of $2\times[(28\times28+1)\times64 + (64+1)\times10] = 101{,}780$ floating-gate memory cells.

The proper synaptic weights were calculated in an external computer using the standard error backpropagation algorithm with a specific cost function, maximizing the voltage difference between the correct and the second-largest network output. Such training, and a differential gate-coupled design, enable very low sensitivity to weight tuning errors, time drift, and temperature changes, which was confirmed by both simulations (Fig. 7) and testing (Figs. 10-13). The weights were "imported" into the circuit, i.e. used to tune the memory state of each array cell to the proper analog value, by utilizing an on-chip peripheral circuitry. In the first experiments reported here, only ~30% of the cells were tuned (Fig. 8) and the weight import accuracy for a single cell tuning was limited to 5%, to decrease the import time.

The input pattern bits are shifted serially into a 785-bit register before each classification; to start it, the bits are read out in parallel into the network. The mixed-signal VMM in the first crossbar array is implemented by applying input voltages (either 4.2 V or 0 V) directly to the gates of the array cell transistors (Fig. 6b). The resulting output currents are passed, through differential amplifiers, to the activation function circuit. The fully analog VMM in the second array is implemented using the gate-coupled approach [1, 4] (Fig. 6c). To minimize the error due to the subthreshold slope's dependence on the memory state (Fig. 2), we have used higher gate voltages (1.1 V to 2.7 V), limited only by technology restrictions. The 10 voltage outputs of the network have been measured externally.

III. Experimental Results

We tested three chips with the same classifier network. As Fig. 9 shows, the average tuning accuracy for the tested chips was 4.4%, 5.6%, and 3.6%. Some of the already tuned cells were disturbed beyond the target accuracy during the subsequent weight import, because of the half-select disturb effect and noise. At this preliminary testing stage, these cells were not re-tuned, because even for such rather crude weight import, the experimentally tested classification fidelity (94.7%, 94.1%, and 94.2%, respectively, for the three tested chips) on the MNIST benchmark test set is already remarkably close to the simulated value (96.2%) for the same network (Fig. 10), and not too far from the maximum fidelity (97.7%) of a perceptron of this size. (We have also obtained preliminary high-fidelity results for the CIFAR-10 set, using our chip, in the TDM mode, in the first layer of the convolutional network described in Ref. [10] – see Fig. 15).

As Fig. 11 shows, the cell conductances slightly decreased (by 13% on the average) over a 7-month storage period. Thought this drift had a noticeable impact on the output voltages of the circuit (Fig. 12), the classification fidelity remained unchanged at 94.7%. Very encouragingly, the classification performance does not suffer at a temperature elevation to 130°C, even without additional efforts to make the circuit temperature-insensitive (Fig. 13).

However, the most important experimental result is that the already mentioned high fidelity of the MNIST set classification has been achieved at an ultralow (sub-20-nJ) energy consumption per average classified pattern (Fig. 14a), and the average classification time below 1 μs (Fig. 14b).

IV. Discussion and Summary

The achieved speed and energy efficiency are much better than those demonstrated, for the same task, with any digital network we are aware of. In particular, both numbers are $>10^3\times$ better than those obtained with the 28-nm IBM TrueNorth digital chip [11]. There are still many ready reserves in our design. For example, current mirrors in neuron circuits may strongly decrease their (currently dominant) contribution to latency and energy. Our estimates (Fig. 16) show that using such mirrors, and a more advanced 55-nm memory technology ESF3 of the same company [7, 10] may enable an at least ~100× advantage in the operation speed, and an enormous, $>10^4\times$ advantage in the energy efficiency, over the state-of-the-art purely digital (GPU and custom) circuits [12], at the classification of large, complex patterns using deep-learning networks – see, e.g., Ref. [13].

Acknowledgment

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

References

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

[2] C. Diorio, P. Hasler, A. Minch, and C. A. Mead, "A single-transistor silicon synapse", *IEEE Trans. Elec. Dev.*, vol. 43, pp. 1972-1980, 1996.

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

[4] J. Hasler and H. Marr, "Fin ing a roadmap to achieve large neuro-morphic hardware systems", *Front. Neurosci.*, vol. 7, art. 118, 2013.

[5] "Superflash Technology Overview", SST, Inc,; available online at www.sst.com/technology/sst-superflash-technology .

[6] 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.

[7] 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.

[8] 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, Apr.-May 2017.

[9] D. B. Strukov and H. Kohlstedt, "Resistive switching phenomena in thin films: Materials, devices, and applications", *MRS Bulletin*, vol. 37, pp. 108-114, 2012.

[10] 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: *Proc. ISSCC'16*, San Francisco, CA, Feb. 2016, pp. 418-420.

[11] S. K. Esser *et al*., "Backpropagation for energy-efficient neuromorphic computing", in: *Proc. NIPS'15*, Montreal, Canada, Dec. 2015, pp. 1117-1125.

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

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

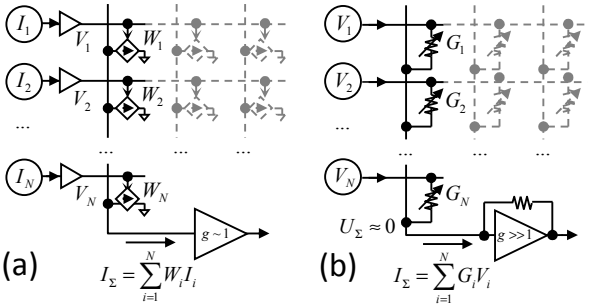


Fig. 1. Analog VMM circuits with (a) floating-gate and (b) memristive cells. Note the difference in opamp gain requirements.

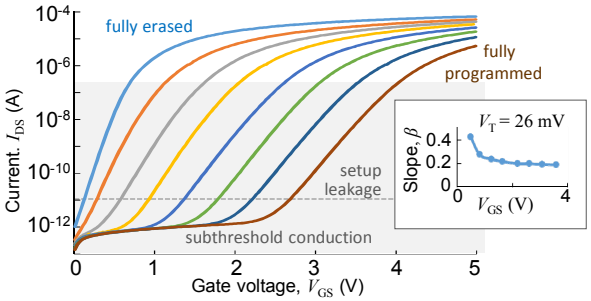


Fig. 3. Drain current of an ESF1 cell, at V_{DS} = 1 V, as a function of the gate voltage, for several memory states. Inset: the log slope β, measured at I_{DS} = 10 nA, as a function of the memory state (shown as the corresponding gate voltage).

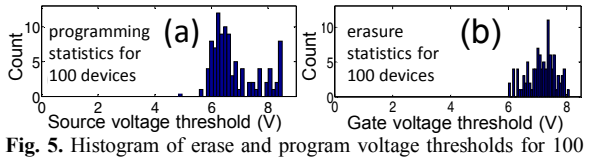


Fig. 5. Histogram of erase and program voltage thresholds for 100 ESF1 memory cells.

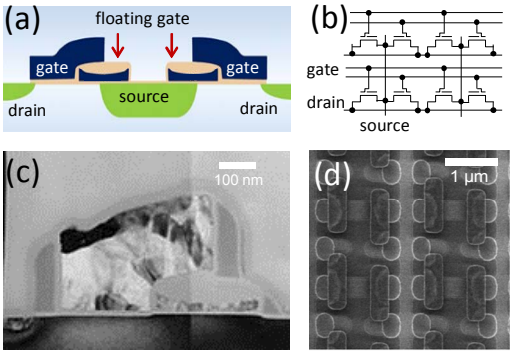
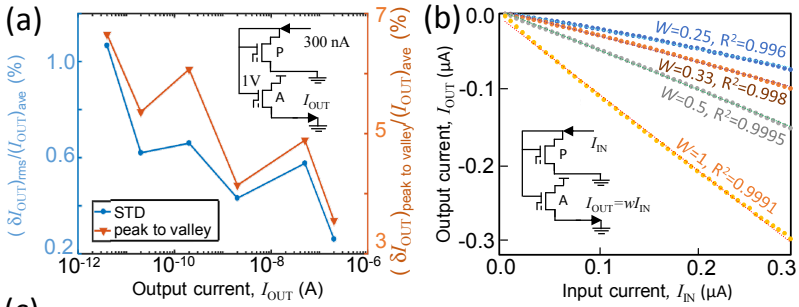


Fig. 2. SST's 180-nm ESF1 NOR flash memory: (a) Cross-section of the two-cell "supercell" (schematically); (b) schematics of 4-supercell fragment of the redesigned array [6, 7], with the gate lines routed vertically to the source lines, to enable single-cell erase; (c) TEM cross-section image of one memory cell; (d) top-view SEM image of four memory supercells.



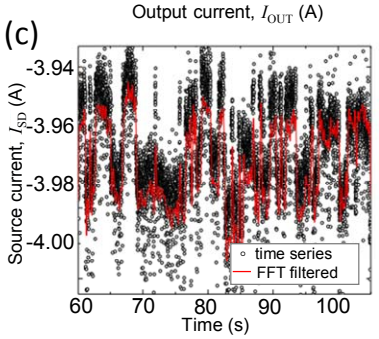


Fig. 4. (a) The relative r.m.s. variation and the full (peak-to-valley) swing of the current for memory cells within a 10×10 array, tuned to different states, measured for the gate-coupled configuration (inset) during a 27+ hour interval. (b) Transfer function of a gate-coupled VMM cell (inset), tuned to four different states. (c) Random telegraph noise of a memory cell at a current of 4 nA.

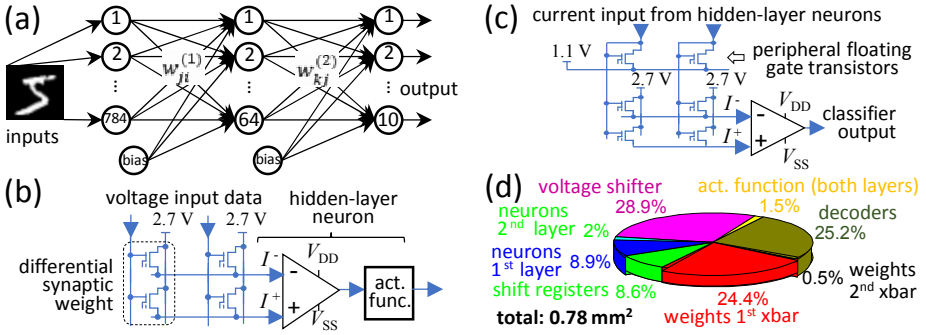
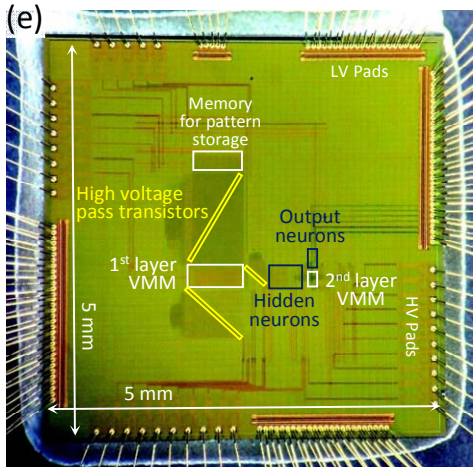


Fig.6. Implementation of a multilayer perceptron network: (a) Graph representation of the 3-layer network; (b, c) circuit diagrams of 2×2 slices of the (b) 1st and (c) 2nd layers. The hidden-layer neuron consists of a differential summing amplifier and an activation function (rectified-tanh) circuit, while the output layer neurons do not implement an activation function. (d) The breakdown of the network's the active component areas, and (e) a micrograph of the fabricated multilayer perceptron chip.



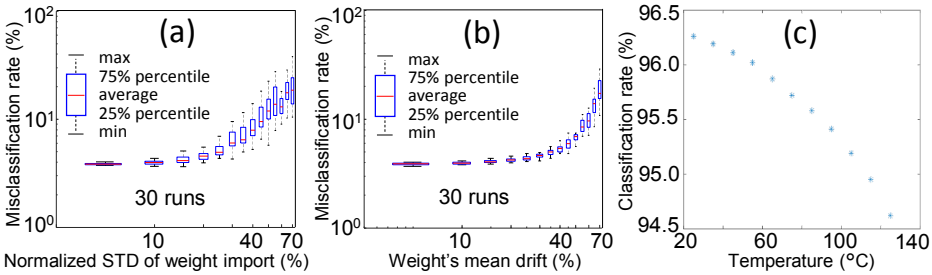


Fig. 7. The simulated classification fidelity of the implemented network as a function of (a) the weight import precision, (b) the weight drift, and (c) temperature, all for the set of weights used in the experiment. The weight variations were modeled by adding, to its optimized value, a normally distributed fluctuations with: on panel (a), zero mean and the shown standard deviation; on panel (b), the shown mean and the 5% standard deviation. The temperature dependence shown on panel (c) was simulated using the experimental results reported in Ref. [8].

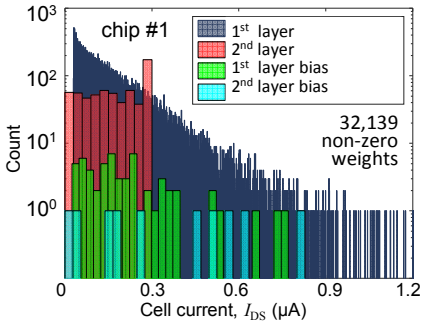


Fig. 8. Histogram of the synaptic weights (measured as the cell currents at V_D = 2.7 / 2.7 V, V_S = 1.65 / 1.1 V, and V_G = 4.2 / 2.7 V in the 1st / 2nd array) used in the experiment.

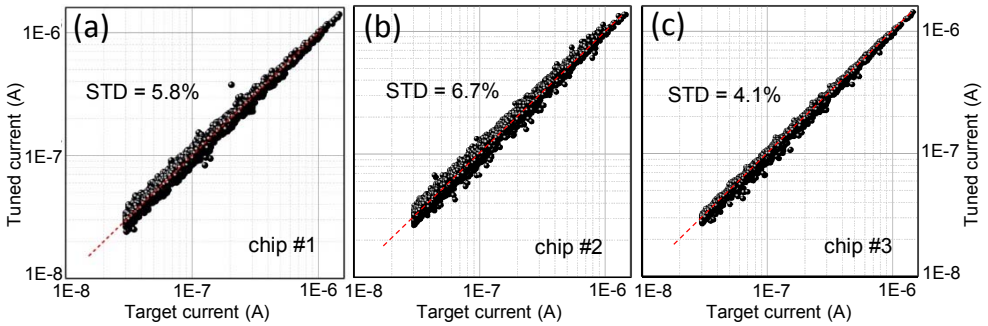


Fig. 9. Synaptic weight import (i.e. cell tuning) statistics: The measured cell currents, at the input voltage of V_G=2.5V, V_{DS}=1V, vs. the target currents (computed at the external network training). The dashed red lines correspond to perfect tuning.

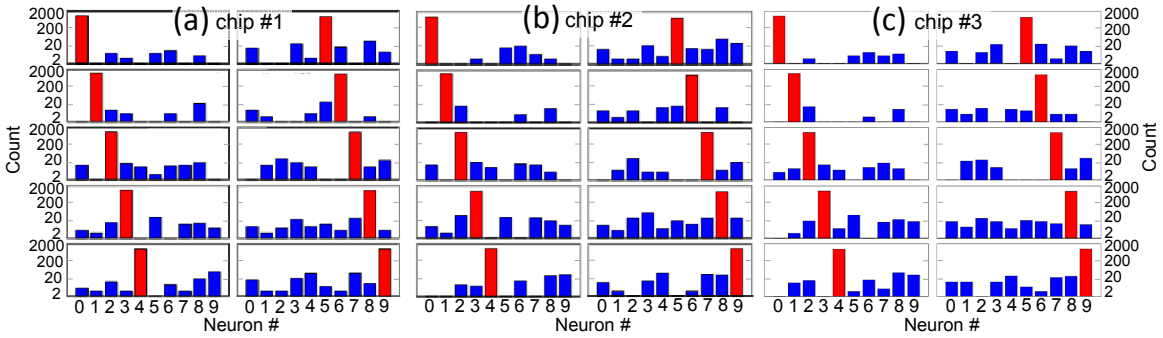


Fig. 10. Classification fidelity statistics: Histograms of the largest output voltages of each class, measured for all 10,000 MNIST test patterns, for three different chips. The correct outputs (red bars) always dominate. Note the logarithmic vertical scale.

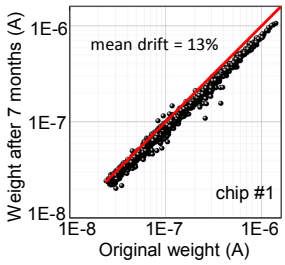


Fig. 11. The freshly tuned weights (cell currents) vs. those measured 7 months later, without re-tuning.

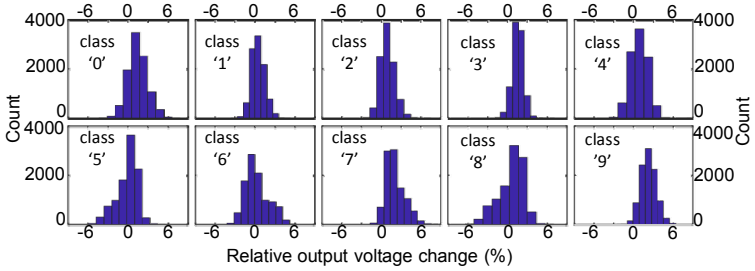


Fig. 12. Histograms of the relative changes of the output voltages for all 10,000 MNIST test set patterns, shown as a normalized difference between the originally measured values and those measured 7 months later, for chip #1.

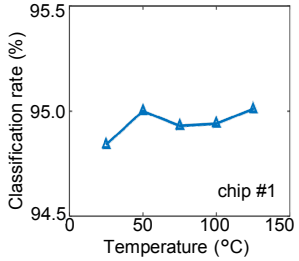
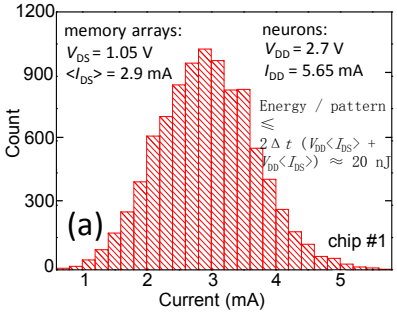


Fig. 13. Measured classification fidelity as a function of the ambient temperature.



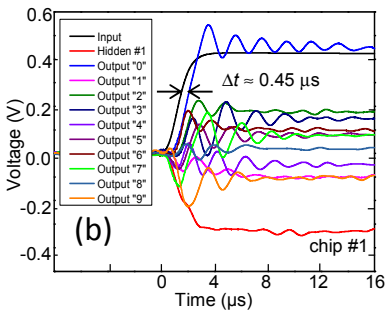


Fig.14. Physical performance: (a) Histogram of the experimentally measured total currents flowing into the circuit, characterizing the static power consumption of the both memory cell arrays, for all patterns of the MNIST test set. The inset lists the pattern-independent static current of the neurons. (b) The typical signal dynamics after an abrupt turn-on of the voltage shifter's power supply, measured concurrently at the network input, at the output of a sample hidden-layer neuron, and at 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. Until it has been improved, and the input circuit sped up, only a sub-1-μs average time delay of the network may be claimed, though it is probably closer to 0.5 μs.

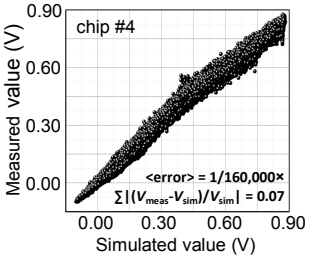


Fig. 15. Experimentally measured pre-activation inputs for 10,000 CIFAR-10 test images, for a particular feature map in the first layer of the network described in Ref. [10], corresponding to the 84% top-3% classification fidelity (vs. the 84.8% simulated for the 6-bit weight precision, and the 85.7% for the full precision).

AlexNet [13] single pattern classification:	Digital circuits [12]		Mixed-signal floating-gate circuits (estimates)	
	GPU 28 nm	ASIC 65 nm	ESF1 180 nm	ESF3 55 nm
time (s)	1.5×10^{-2}	2.9×10^{-2}	$\sim1\times10^{-4}$	$\sim6\times10^{-5}$
energy (J)	1.5×10^{-1}	0.8×10^{-2}	$\sim3\times10^{-7}$	$\sim2\times10^{-7}$

Fig. 16. Speed and energy consumption of signal propagation through the convolutional (dominating) part of a large deep network [13], with ~0.65×10^6 neurons, at its various implementations. The estimates for floating-gate networks take into account the 55×55 = 3,025-step time-division multiplexing (natural for this particular network), and the experimental values of the subthreshold current slope of the cells (see the inset in Fig. 3).