DEVICE TECHNOLOGY

Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing

Elliot J. Fuller[1], Scott T. Keene[2]*, Armantas Melianas[2]*, Zhongrui Wang[3], Sapan Agarwal[1], Yiyang Li[1], Yaakov Tuchman[2], Conrad D. James[4], Matthew J. Marinella[4], J. Joshua Yang[3], Alberto Salleo[2]†, A. Alec Talin[1]†

Neuromorphic computers could overcome efficiency bottlenecks inherent to conventional computing through parallel programming and readout of artificial neural network weights in a crossbar memory array. However, selective and linear weight updates and <10-nanoampere read currents are required for learning that surpasses conventional computing efficiency. We introduce an ionic floating-gate memory array based on a polymer redox transistor connected to a conductive-bridge memory (CBM). Selective and linear programming of a redox transistor array is executed in parallel by overcoming the bridging threshold voltage of the CBMs. Synaptic weight readout with currents <10 nanoamperes is achieved by diluting the conductive polymer with an insulator to decrease the conductance. The redox transistors endure >1 billion write-read operations and support >1-megahertz write-read frequencies.

The increasing collection and availability of data over the past decade have led to a surge of interest in developing artificial neural networks (ANNs) to tackle problems across diverse fields, such as image and natural language processing, autonomous driving, finance, and bioinformatics (*1*). However, because of the decline of complementary metal oxide semiconductor (CMOS) scaling, innovative approaches to computing are needed to address the rapidly growing computation density and efficiency requirements (*2*). One approach is to use large crossbar arrays of synaptic memory elements to execute ANN algorithms. The conductance, or state, of each synaptic element is first tuned during learning with a training data set and then used to process new information during inference (e.g., recognition and classification). By executing learning and inference within memory, neuromorphic computers circumvent energy-costly data movement between processor and memory, thus reducing both power consumption and computation time. However, to realize clear advantages over digital approaches (*3*, *4*), neuromorphic computers must execute a large number of analog operations—for example, summations, multiplications, and so on—in parallel and without data movement (*5*). For efficient and parallel execution of learning and inference, crossbars should be large (>1000 synapses by 1000 synapses), and devices should have linear and symmetric conductance tuning, low-current write-read operations, fast switching speeds, and high endurance (*5*).

Despite compelling demonstrations of inference using crossbars based on a variety of synaptic devices, efficient learning remains a challenge owing to the nonideal electrical characteristics of synaptic devices. For example, two-terminal devices, such as memristors based on phase-change memory (PCM) (*6*) or filament forming metal oxides (FFMO) (*7*), typically exhibit a superexponential dependence of the current on the applied voltage during writes. This is because the device must be driven out of thermal equilibrium to overcome a large energy barrier to change the electronic conductance state (i.e., melting or oxygen vacancy motion). The resulting nonlinear and asymmetric conductance tuning rapidly degrades ANN accuracy (*8*). Improved linearity and symmetry can be realized by programming with current and using state resetting procedures, but these schemes cannot be used for parallel programming to accelerate learning: Instead, memristor arrays require time- and energy-costly element-by-element programming (Fig. 1A) (*9*).

Furthermore, crossbar approaches to inference (e.g., classification) hinge upon summing "trickle currents" from each synaptic element along crossbar columns in order to execute enough "summations" (through an analog dot product) to beat digital approaches in latency and energy. However, two-terminal memristors draw more than a trickle, with currents greater than a microampere during reads and writes that notably reduce accuracy in large arrays. The problem arises because the currents flowing through individual synapses sum along interconnecting wires, and the resulting total current grows beyond the wire capacity as the array is scaled to many elements. For example, currents flowing through interconnects lead to large voltage drops across the parasitic crossbar wire resistance that increasingly reduce write-read accuracy as these wires are scaled to <100 nm in width (*9*, *10*). For these reasons, existing neuromorphic computing demonstrations with two-terminal devices have been limited to relatively small arrays (<100 by 100) and large wires (>1 μm width), and dot product efficiencies below conventional approaches (e.g., Google TPU and NVIDIA Xavier) (*5*).

Recently developed redox-transistor memory is a promising approach to circumvent existing memristor technology limitations. The three-terminal redox transistor decouples the write and read operations using a "gate" electrode to tune the conductance state through electrochemical reactions involving Li^+ or H^+ ion injection into the channel electrode through a solid electrolyte (see Fig. 1, C and D). The insertion of cations into the bulk of the channel acts to dope the material through a gradual composition modulation that leads up to thousands of finely spaced conductance levels with near-ideal analog behavior. For example, redox transistors based on inorganic and organic materials have been recently demonstrated with conductance tuning occurring at potentials of just a few millivolts and symmetrically programmable conductance states that enable near-ideal accuracy in neural network simulations (*11*–*15*). However, redox-transistor memory has not been demonstrated in an array, which requires some way of addressing each element in parallel during write operations as well as ensuring state retention after write operations. Without a way to address these memory devices in parallel and preserve their conductance state they cannot execute efficient learning. Furthermore, high read currents inhibit their use in any neuromorphic computing implementation.

Here we enable parallel programming and state retention by integrating a polymer-based redox transistor (*12*) and a volatile conductive-bridge memory (CBM) (*16*) to produce a nonvolatile, addressable synaptic memory we call ionic floating-gate memory (IFG). The three-terminal design allows the channel to be engineered for ultralow-current read operations without sacrificing analog performance through diluting the conductive polymer in a polymeric insulator, which we discuss in detail later. Figure 1, C and D, is an illustration of the IFG device concept with the gate terminal of a redox transistor electrically connected to the CBM. The redox transistor is composed of a semiconducting polymer blend poly(3,4-ethylenedioxythiophene): poly(styrene sulfonate) (PEDOT:PSS) (dark blue) separated by a solid electrolyte (light blue). The synaptic weight is stored as the transistor source-drain conductance *G*. The CBM mediates electronic coupling to the redox transistor gate during and after programming. During programming, electron injection (extraction) through the CBM into the top PEDOT:PSS gate results in the reversible electrochemical oxidation (reduction) of

[1]Sandia National Laboratories, Livermore, CA, USA. [2]Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA. [3]Department of Computer Science and Electrical Engineering, University of Massachusetts Amherst, Amherst, MA, USA. [4]Sandia National Laboratories, Albuquerque, NM, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: aatalin@sandia.gov (A.A.T.); asalleo@stanford.edu (A.S.)

the bottom PEDOT:PSS channel, thereby increasing (decreasing) G.

Figure 1E shows G modulation during IFG programming with voltage pulses sent to the CBM terminal. A programming voltage of V^W = ±650 mV results in analog tuning through 50 conductance states, whereas programming with V^W = ±1 V achieves the same number of states but with higher (2×) modulation (light blue). Similarly, hundreds of conductance states can be achieved through adjusting the write pulse duration and/or amplitude (*12*, *17*). Although the ON/OFF ratio is smaller compared with that of memristors, this is a desirable feature for neuromorphic computing. Unlike for digital logic, a smaller contrast between ON and OFF prevents large currents from saturating neurons (e.g., stuck ON pixels for PCM), although at lower ON/OFF ratios, reducing noise becomes increasingly important to network accuracy. To fully account for noise, nonlinearity, and asymmetry in programming, we acquire a statistical distribution of IFG conductance levels from more than 5000 switching events (see fig. S1). The resulting write linearity, symmetry, and signal-to-noise ratio of $\Delta G^2/\sigma^2$ = 91, where σ is the standard deviation of the conductance update, meet the requirements for high ANN training accuracy when executing updates (*8*).

To enable state retention, we designed our redox transistors to operate as concentration cells to eliminate large built-in voltages observed in previously reported devices due to the differences in chemical potential between the gate and channel [i.e., Si and Li_xCoO_2 in (*11*)]. Here, the gate and channel have been symmetrically doped with polyethyleneimine (*18*) to lower the built-in voltage below the CBM ON threshold V^{th} and allow the CBM device to transition to the OFF state after write operations. After write operations, when $|V^W| \ll |V^{th}|$, the conductive bridge dissolves and the device switches to the OFF state, in which case charge neutrality and the absence of an electronic pathway between the redox-transistor gate and channel preserve the redox-transistor conductance state (Fig. 1D). The CBM device is based on a Pt/Ag/SiO_xN_y/Ag/Pt stack (purple) that is similar to previously reported devices (*16*). As with flash memory, the CBM allows current injection only above a threshold with $|V^W| > V^{th} = \pm 400$ mV and not during the OFF state (Fig. 1C). Therefore, IFG operates as a floating-gate memory similar to flash, but with more than an order of magnitude lower voltage operation and with near-perfect linear characteristics.

To demonstrate parallel addressing with IFG, we programmed two cells with conductance G_{11} and G_{12} electrically connected in a crossbar array as schematically illustrated in Fig. 1F. In this scheme, a weight update is encoded by voltages applied along the rows V_i^W and columns V_j^W. The elements G_{11} or G_{12} program when the total cell voltage is greater than the CBM threshold $|V_i^W - V_j^W| > V^{th}$. This scheme can be used to execute parallel writes to memory during supervised learning (e.g., backpropagation) through encoding the outer-product vector values V_i and V_j as ANN weight updates (*19*). Figure 1G demonstrates selective addressing by subjecting one of the crossbar elements to 50 weight updates without disturbing the other element. The highly selective and parallel addressing is enabled by the nearly 1-mV/decade slope and large ON/OFF ratio (10^{10}) of the CBM (*20*). Although a silicon field-effect transistor could support sequential, nonvolatile programming of a redox transistor, its gradual ≥60-mV/decade subthreshold slope and limited ON/OFF ratio of 10^6 make it inferior for parallel selectivity and state retention compared with the CBM.

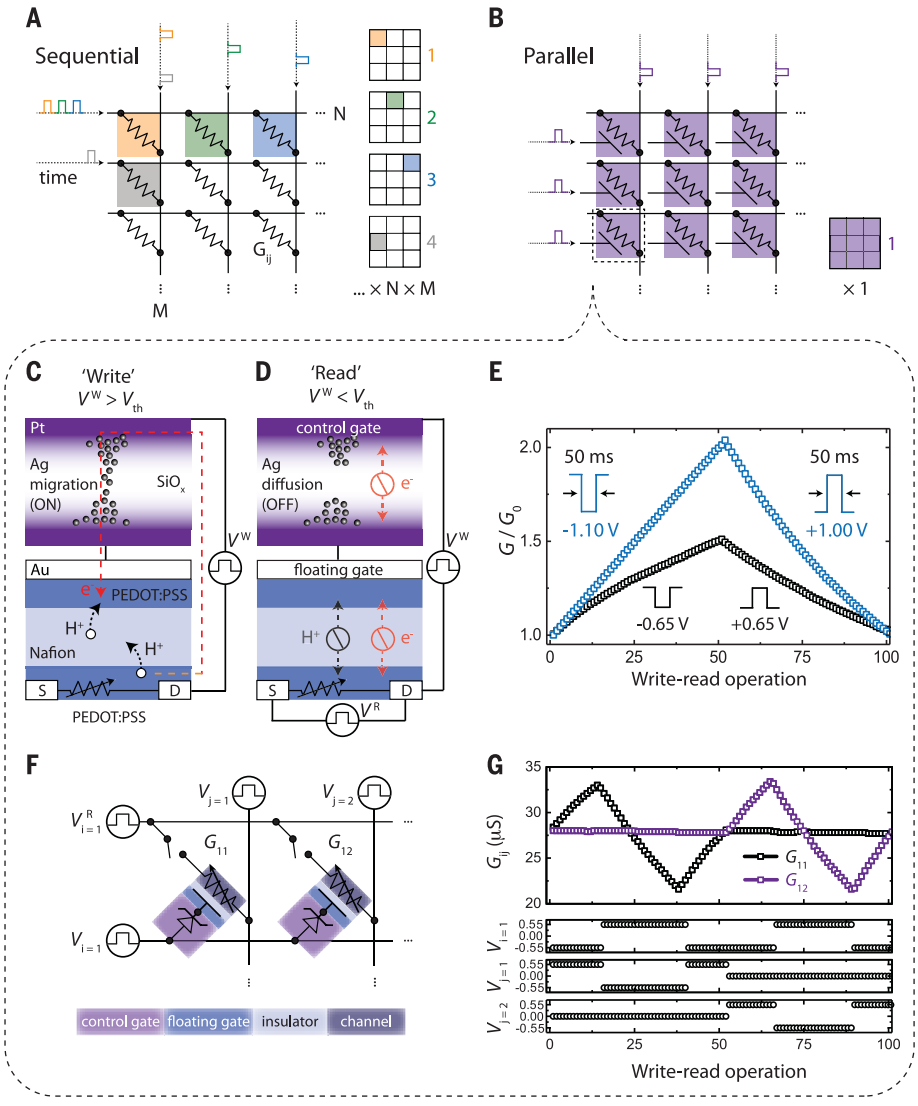

Fig. 1. IFG memory device characteristics. (**A**) Sequential programming of an N-by-M element resistive memory crossbar array has a latency of O(N × M) write cycles, where N and M are the number of array elements along a row and column, respectively, and O denotes the number of operations required to program the array. (**B**) Using parallel programming, the array is updated within a single cycle, resulting in an O(N × M) advantage. (**C**) An IFG cell consists of a CBM that mediates electron injection in and out of a redox transistor. Selective programming occurs when V^W is greater than the ON threshold V^{th} for Ag filament formation. During electron injection (extraction) into the top PEDOT:PSS gate, the PEDOT:PSS channel is oxidized (reduced), increasing (decreasing) the electronic conductance between the source (S)–drain (D) electrodes G_{SD}. (**D**) Below the programming threshold, that is, for $V^W << V^{th}$, the Ag filament ruptures, resulting in a high-resistance OFF state that prevents further electronic coupling to the redox transistor and enables nonvolatility. The state of the redox transistor is read out by applying a read voltage V^R across the source-drain electrodes. (**E**) Programming of an IFG cell at two different write voltages, V^W = ±0.65 and ±1.00 V, with 1.5× and 2× modulation of G_{SD}, respectively. (**F**) Schematic of a 1-by-2 IFG resistive memory array. Programming is carried out by applying voltages V_i^W and V_j^W along the rows and columns, respectively. (**G**) Selective addressing by subjecting G_{11} or G_{12} to 50 weight updates without disturbing the adjacent element.

As arrays scale to more than 1000 elements by 1000 elements, read currents must be <10 nA to reduce voltage losses below 1% across scaled wires at 10-nm half-pitch. The IFG three-terminal cell and polymer-based weight storage medium provide an opportunity to lower read currents without the loss of linear or symmetric programmability or the introduction of write noise. To lower the read current, we adjusted the PEDOT:PSS formulation to 1:41 monomer ratio (PEDOT to PSS), lowering the average channel conductance to <100 nS (i.e., read current <10 nA at 100-mV read voltage) while maintaining a high signal-to-noise ratio during nearly linear and symmetric programming (Fig. 2A). The combined low noise, low current, and near-linear conductance tuning observed in Fig. 2A has not been achieved by other resistive memories. Although some devices have been engineered to operate at <50 nA (*21*), they either are binary or suffer from write noise that severely reduces ANN accuracy (*22*).

Previously reported polymer-based redox transistors required at least a 6-ms write pulse to change conductance state, which is too slow for practical neuromorphic computing (*12*). However, on the basis of the measured proton mobility in PEDOT:PSS (*23*), substantially faster operation is expected for downscaled devices (see fig. S2). To investigate the limits of PEDOT:PSS devices, we have scaled both the channel and gate dimensions down to 45 μm by 125 μm, which enables <1-μs write operations. Figure 2B shows complementary data for the time-resolved conductance response to 200-ns programming voltage pulses and subsequent 500-ns readout voltage pulses, resulting in a total write-read duration of <1 μs. Volatile CBMs, similar to devices in this study, have demonstrated ON and OFF transition times of 7.5 and 20 ns, respectively (*24*), that should support IFG write times of <1 μs. Although the OFF transition time may contribute to memory loss owing to discharge, programming is possible so long as the write times exceed the OFF transition time leading to a net charge injection.

The programming current required to modulate the channel conductance scales with decreasing channel area, and, using a previously reported model (*17*), we project that further downscaled devices will enable <100-ns switching (Fig. 2C). The write speed is estimated to further improve by replacing the lateral geometry used here with a vertical stacking of redox-transistor materials (fig. S2). However, aggressive scaling also introduces an increased sensitivity to environmental effects and leakage through the CBM, which can affect state retention, as was found in a previous study (*18*). For relevance to applications beyond learning acceleration (i.e., embedded dot-product engines), improvements to device encapsulation processes and the CBM OFF resistance will be required.

Endurance is another important criterion for neuromorphic computing technology. In batteries, parasitic reactions associated with solid electrolyte interface formation and cathode dissolution limit endurance to ~1000 charge-discharge cycles (*25*). By contrast, redox transistors demonstrated here operate near 0 V with considerably improved endurance. We observed that a redox transistor exhibited no degradation after 10^9 binary write-read operations (fig. S3) and 10^8 write-read operations sampling the entire device range (Fig. 2D). It is noteworthy that the endurance of our organic device is several orders of magnitude greater than that reported for flash memory, which is subject to oxide stresses (*26*).

To further demonstrate the feasibility of our proposed approach, we executed parallel inference and weight update operations using a 3-by-3 prototype array illustrated in Fig. 3A. The IFG cell array was programmed in parallel and without feedback circuits through the outer-product vector values V_i and V_j, similar to the two-device case (Fig. 1F). Four examples of programming patterns are shown in Fig. 3B, with the relative change in conductance $\Delta G_{ij} = G_{ij} - G_0$ indicated explicitly for every cell in the array and the corresponding outer-product vector values indicated at row and column edges. The average update in the cells is 53 μS with a standard deviation of 11 μS, whereas the unselected cells remain mostly undisturbed during programming.

Until now, accurate and scalable parallel weight updates have remained elusive. For example, previous demonstrations have implemented parallel programming of elements along a crossbar column to accelerate learning (*27*). However, the nonlinear dependence of the weight update ΔG_{ij} on the previous conductance state G_0 rendered the approach less generalizable to many classification problems, whereas programming currents greater than a microampere prevent scaling to large arrays. Compared with the two-device case (Fig. 1F), we attribute the greater update variance in the IFG array in Fig. 3B to not-yet-perfected channel and electrolyte processing. Reducing device-to-device variability is critical to realizing working neuromorphic computers, and the update variance should be reduced to <1% of the average update for high accuracy. It should be noted that for redox transistors that rely upon bulk ion insertion, device

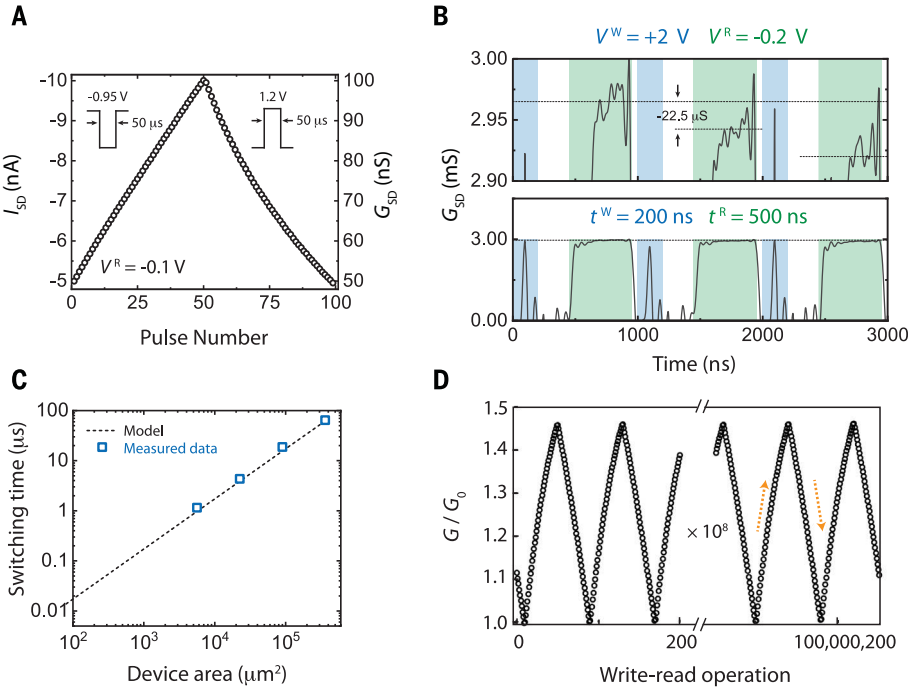

Fig. 2. Redox-transistor low-current operation, speed, scaling, and write-read endurance (without CBM). (**A**) Programming at a small channel current (I_{SD}) of 5 to 10 nA (50 to 100 nS). The high-impedance redox transistor is enabled by the engineered channel formulation of PEDOT:PSS (1:41 monomer ratio). (**B**) Time-resolved source-drain conductance of a PEDOT:PSS redox-transistor channel (125 μm by 45 μm) programmed with 200-ns write pulses. The bottom panel is the entire source-drain conductance trace, and the top panel is a magnified section of the bottom panel highlighting the conductance changes between write-read operations. Blue shading indicates the write duration, and green shading indicates the read duration. (**C**) Estimated (dashed line) and measured (open squares) redox-transistor switching speed scaling with channel area. Switching times correspond to the required write pulse duration to span the total device conductance range (0.5 mS) using 100 pulses (V_{pulse} = 2 V), that is, each write pulse corresponds to 1% device conductance change (ΔG = 5 μS). The equivalent circuit model from (*17*) (dashed trace) describes the redox transistor as a charging capacitor under pulsed bias. (**D**) Demonstration of >10^8 write-read operations (cycling between the low- and high-conductance state) without deterioration of device properties. The red arrows indicate the conductance change direction.

variability represents a distinctly different challenge than controlling, for example, nanoscale filament formation in FFMO devices. These uniformity challenges can be overcome by appropriate process control, as demonstrated by the organic light-emitting diode (OLED) industry (*28*).

In addition to accelerating the weight update step, we demonstrate inference after mapping the "exclusive or" (XOR) function to a two-layer neural network using a 3-by-3 IFG array (Fig. 3C). The ability to classify the XOR function is a basic demonstration of nonlinear function mapping, which is required for general purpose ANN computation. For this demonstration, neurons are chosen to fire on the basis of summing currents I_j along the columns using a transimpedance amplifier circuit followed by passing the resulting sum to the sigmoidal activation function (implemented in software). To visualize network performance, the state of each element is read out during classification and plotted in Fig. 3D. Here, an input example X = [1, 0] (blue) is fed to the first layer of the network (orange), whereas the output Y^T = [1, 1], where T denotes matrix transpose, is sent to the last layer (green). The final output of the network correctly classifies X = [1, 0] as Z^T = [1] according to the XOR truth table in Fig. 3C. The network is found to classify all XOR inputs with 100% accuracy. Our crossbar is used to execute analog dot products during inference (Fig. 3D) and analog outer-product updates during write operations (Fig. 3B); however, all other calculations are executed entirely in CMOS following the design of a hybrid analog-digital accelerator that was reported previously (*5*).

Finally, we simulate the performance of a 1024-by-1024 IFG array implemented in a hybrid IFG-digital accelerator (see "Architectural Simulations" section in the supplementary text). We evaluate the network performance in classification of the Modified National Institute of Standards and Technology (MNIST) handwritten digit dataset based on the experimentally measured statistical distribution of IFG noise, nonlinearity, and asymmetry (fig. S1). The resulting network achieves ideal accuracy (Fig. 3E), far better than similar simulations of PCM (*6*) or FFMO crossbars (*29*). Recently, two-terminal devices based on Ag migration in etched dislocations were found to have similar accuracy in network simulations, owing to improved linearity and symmetry in programming (*30*). However, there has been no demonstration of both low currents and low noise with these devices.

We modeled the overall energy, latency, and area of a 1024-by-1024 hybrid IFG-digital accelerator by modifying the architectural level analysis in (*5*). Our model takes into account devices (area = 300 nm by 300 nm; thickness = 200 nm) with read and write currents <10 nA based on our polymer dilution process and previous experimental projections of write current scaling (*12*). With reduced currents, the energy of a learning accelerator becomes dominated by the circuit overheads owing to analog-to-digital conversion of the analog integrator output. Taking these overheads into account, an IFG accelerator is projected to provide energy, latency, and area advantages of 476×, 16×, and 9.5×, respectively, when compared directly with an optimized eight-bit static random-access memory (SRAM) accelerator using a 14- to 16-nm technology node (Fig. 3, F to H). Although these advantages are attractive for learning accelerators, notable challenges remain in integrating the organic polymer and Ag-based devices with Si CMOS, owing to the strict temperature and contamination requirements. However, alternative CMOS integration pathways (i.e., chip-to-wafer bonding) and new materials (e.g., Cu-based CBM) are being explored to improve such compatibility.

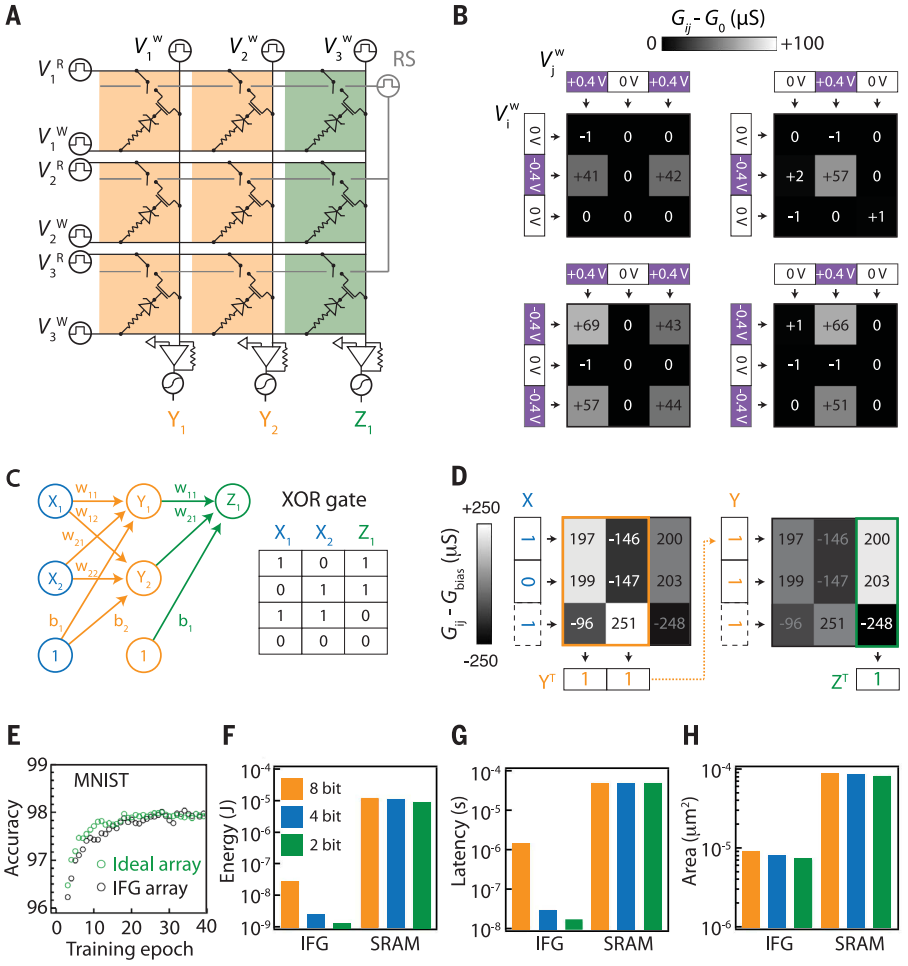

Fig. 3. Parallel programming, XOR mapping, and simulations of an IFG neuromorphic array. (**A**) Schematic of a 3-by-3 prototype IFG array divided into a two-layer neural network, as indicated by orange and green. Network inputs V_i^R are applied across the source-drain rows, while programming inputs V_i^W and V_j^W are applied along the gate row and drain column, respectively. An amplifier is used to read the currents along the crossbar columns. The amplifier output is read using an analog-to-digital converter and passed to a sigmoidal activation function to dictate whether the neurons Y_1, Y_2, and Z_1 will fire. A read-select transistor (RS) in every cell is switched globally between read and write operations to prevent sneak currents without loss of parallelism. (**B**) Four examples of parallel updates executed on the IFG crossbar, with grayscale indicating update strength. The corresponding programming vectors V_i^W and V_j^W are indicated at the edges of the rows and columns, respectively (purple). (**C**) The IFG array is mapped to a two-layer neural network used to classify XOR logic. (**D**) XOR classification of the input X = [1, 0]. The conductance state of each element G_{ij} is plotted explicitly using grayscale. The input X = [1, 0] values are shown in blue. The output of the first layer Y^T = [1, 1] (orange), where T denotes matrix transpose, is sent to the second layer. The final layer output Z = [1] is shown in green. (**E**) Simulated ANN accuracy for MNIST handwritten digit classification for 1024-by-1024–sized IFG array (black) and ideal device accuracy (green). (**F** to **H**) Simulated energy (F), latency (G), and area (H) for a hybrid IFG-CMOS accelerator with eight-bit, four-bit, and two-bit accuracy shown in orange, blue, and green, respectively. The IFG array is compared directly with an optimized SRAM-based digital accelerator using a 14- to 16-nm technology node.

Most critically, IFG prototype networks demonstrate that low-voltage electrochemical systems can be engineered to execute highly efficient learning and inference in memory. The scalable electrical characteristics of IFG open a path toward neuromorphic computers that could surpass their digital counterparts with a projected nearly three orders of magnitude improvement in energy efficiency. Such highly efficient neuromorphic computers could extend ANN learning to new low-power environments such as edge computing (e.g., mobile and wearable devices) or support adaptive neural algorithms that enable continuous learning through the life cycle of a product. The proposed IFG memory concept can be generalized to a wide variety of electrochemical systems (i.e., metal oxide cathodes, new semiconducting polymers, Cu-based CBM, and so on) that could provide even greater performance with improved CMOS compatibility and that are yet unexplored for analog memory applications.

REFERENCES AND NOTES

1. Y. LeCun, Y. Bengio, G. Hinton, *Nature* **521**, 436–444 (2015).
2. X. Xu *et al.*, *Nat. Electron.* **1**, 216–222 (2018).
3. A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, N. Andrew, in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta, D. McAllester, Eds. (Proceedings of Machine Learning Research, 2013), vol. 28, pp. 1337–1345.
4. N. P. Jouppi *et al.*, in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)* (ACM, 2017), vol. 1, pp. 1–12.
5. M. J. Marinella *et al.*, *IEEE J. Emerg. Sel. Top. Circuits Syst.* **8**, 86–101 (2018).
6. G. W. Burr *et al.*, *IEEE Trans. Electron Devices* **62**, 3498–3507 (2015).
7. D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams, *Nature* **453**, 80–83 (2008).
8. S. Agarwal *et al.*, in *2016 International Joint Conference on Neural Networks (IJCNN)* (IJCNN, 2016), pp. 929–938.
9. C. Li *et al.*, *Nat. Commun.* **9**, 2385 (2018).
10. C. Li *et al.*, *Nat. Electron.* **1**, 52–59 (2018).
11. E. J. Fuller *et al.*, *Adv. Mater.* **29**, 1604310 (2017).
12. Y. van de Burgt *et al.*, *Nat. Mater.* **16**, 414–418 (2017).
13. M. T. Sharbati *et al.*, *Adv. Mater.* **30**, e1802353 (2018).
14. C. S. Yang *et al.*, *Adv. Funct. Mater.* **28**, 1804170 (2018).
15. J. Tang *et al.*, paper presented at the 2018 IEEE International Electron Devices Meeting, San Francisco, CA, 1 to 5 December 2018.
16. Z. Wang *et al.*, *Nat. Mater.* **16**, 101–108 (2017).
17. S. T. Keene *et al.*, *J. Phys. D Appl. Phys.* **51**, 224002 (2018).
18. S. T. Keene, A. Melianas, Y. van de Burgt, A. Salleo, *Adv. Electron. Mater.* **5**, 1800686 (2019).
19. S. Agarwal *et al.*, *Front. Neurosci.* **9**, 484 (2016).
20. R. Midya *et al.*, *Adv. Mater.* **29**, 1604457 (2017).
21. S.-G. Park *et al.*, in *2012 Electron Devices Meeting (IEDM 2012)* (IEEE, 2012), pp. 20.28.21–20.28.24.
22. D. Ielmini, F. Nardi, C. Cagli, *Appl. Phys. Lett.* **96**, 053503 (2010).
23. E. Stavrinidou *et al.*, *Adv. Mater.* **25**, 4488–4493 (2013).
24. B. Cheng *et al.*, *Commun. Phys.* **2**, 28 (2019).
25. Y. Zhu, X. He, Y. Mo, *ACS Appl. Mater. Interfaces* **7**, 23685–23693 (2015).
26. P. Pavan, R. Bez, P. Olivo, E. Zanoni, *Proc. IEEE* **85**, 1248–1271 (1997).
27. M. Prezioso *et al.*, *Nature* **521**, 61–64 (2015).
28. Y. K. Jung *et al.*, *SID Symp. Dig. Tech. Pap.* **47**, 707–710 (2016).
29. R. B. Jacobs-Gedrim *et al.*, in *2017 IEEE International Conference on Rebooting Computing (ICRC)* (IEEE, 2017), pp. 1–10.
30. S. Choi *et al.*, *Nat. Mater.* **17**, 335–340 (2018).

ACKNOWLEDGMENTS

We gratefully acknowledge assistance with device characterization software development from C. Bayley and U. Sohi and assistance with high-speed data acquisition equipment from R. B. Jacobs-Gedrim and D. Muratore. We thank Solvay for providing us with the Aquivion electrolyte. **Funding:** This work was supported in part by Sandia's Laboratory-Directed Research and Development (LDRD) Program under the Hardware Acceleration of Adaptive Neural Algorithms (HAANA) Grand Challenge. E.J.F. and A.A.T. were also supported by Nanostructures for Electrical Energy Storage (NEES-II), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under award number DESC0001160. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the U.S. government. A.S. and S.T.K. acknowledge financial support from the National Science Foundation and the Semiconductor Research Corporation, E2CDA award no. 1739795. Additionally, S.T.K. acknowledges the Stanford Graduate Fellowship fund for support. A.M. gratefully acknowledges support from the Knut and Alice Wallenberg Foundation (KAW 2016.0494) for postdoctoral research at Stanford University. This work was in part performed at the Stanford Nano Shared Facilities (SNSF) and the nano@Stanford (SNF) labs, which are supported by the National Science Foundation as part of the National Nanotechnology Coordinated Infrastructure under award ECCS-1542152. **Author contributions:** E.J.F., S.T.K., A.M., S.A., A.S., and A.A.T. conceptualized the devices and experiments. E.J.F., S.T.K., A.M., Y.L., and Y.T. investigated the materials and devices. S.A. and M.J.M. investigated the architecture and neural network performance. A.M., Y.T., and S.T.K. fabricated redox-transistor devices and arrays. Z.W. fabricated CBMs. E.J.F., S.T.K., A.M., A.A.T., and A.S. prepared the manuscript. A.A.T., A.S., J.J.Y., M.J.M., and C.D.J. acquired the financial support. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/364/6440/570/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 and S2
References (*31*–*33*)

4 January 2019; accepted 15 April 2019
Published online 25 April 2019
10.1126/science.aaw5581

Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing

Elliot J. Fuller, Scott T. Keene, Armantas Melianas, Zhongrui Wang, Sapan Agarwal, Yiyang Li, Yaakov Tuchman, Conrad D. James, Matthew J. Marinella, J. Joshua Yang, Alberto Salleo and A. Alec Talin

Science **364** (6440), 570-574.
DOI: 10.1126/science.aaw5581originally published online April 25, 2019

Ionic floating-gate memories

Digital implementations of artificial neural networks perform many tasks, such as image recognition and language processing, but are too energy intensive for many applications. Analog circuits that use large crossbar arrays of synaptic memory elements represent a low-power alternative, but most devices cannot update the synaptic weights uniformly or scale to large array sizes. Fuller *et al.* developed an integrated device, ionic floating-gate memory, that has the gate terminal of a redox transistor electrically connected to a diffusive memristor. This low-power device enabled linear and symmetric weight updates in parallel over an entire crossbar array at megahertz rates over 10^9 write-read cycles.

Science, this issue p. 570