IEEE Journal on Exploratory Solid-State Computational Devices and Circuits

Received 18 February 2019; accepted 23 February 2019. Date of publication 8 March 2019; date of current version 31 May 2019.

Digital Object Identifier 10.1109/JXCDC.2019.2902409

Using Floating-Gate Memory to Train Ideal Accuracy Neural Networks

SAPAN AGARWAL[1] (Member, IEEE), DIANA GARLAND[2], JOHN NIROULA[2], ROBIN B. JACOBS-GEDRIM[2] (Member, IEEE), ALEX HSIA[2], MICHAEL S. VAN HEUKELOM[2], ELLIOT FULLER[1], BRUCE DRAPER[2], and MATTHEW J. MARINELLA[2] (Senior Member, IEEE)

[1]Sandia National Laboratories, Livermore, CA 94550 USA
[2]Sandia National Laboratories, Albuquerque, NM 87123, USA

CORRESPONDING AUTHOR: S. AGARWAL (sagarwa@sandia.gov).

This work was supported by the Department of Energy (DOE) Advanced Manufacturing Office and the Laboratory Directed Research and Development program at Sandia National Laboratories.

ABSTRACT Floating-gate silicon-oxygen-nitrogen-oxygen-silicon (SONOS) transistors can be used to train neural networks to ideal accuracies that match those of floating-point digital weights on the MNIST handwritten digit data set when using multiple devices to represent a weight or within 1% of ideal accuracy when using a single device. This is enabled by operating devices in the subthreshold regime, where they exhibit symmetric write nonlinearities. A neural training accelerator core based on SONOS with a single device per weight would increase energy efficiency by 120×, operate 2.1× faster, and require 5× lower area than an optimized SRAM-based ASIC.

INDEX TERMS Analog, flash, floating gate, memristor, neural network (NN), neuromorphic, silicon-oxygen-nitrogen-oxygen-silicon (SONOS), training.

I. INTRODUCTION

Analog accelerators promise to improve the energy and latency of training a neural network (NN) by more than a 100× over an optimized ASIC [1]. Analog matrix operations are used to process each memory element in parallel and thereby eliminate data movement, as illustrated in Fig. 1 [2]. However, this requires devices with high resistance, low write variability, and low write nonlinearity [3]. Resistive memory devices have been used to represent synaptic weights, but the write variability and asymmetric write nonlinearity in current resistive memory device technology prevent the weights from being learned to high accuracy [3], [4]. Algorithmic and circuit techniques help improve accuracy [5], [6], but NN accuracy is not ideal. Novel lithium [7] and polymer [8] based devices with excellent analog properties have been demonstrated but will require continued work to integrate into modern CMOS foundries. In this paper, we show that a conventional floating-gate memory, commonly available in foundries, can be used train an NN to within 1% of that achieved with floating-point weights on MNIST data set (ideal accuracy). It has been shown that floating-gate memories can be used to create accurate inference accelerators [9], [10]. We extend this to online training. Furthermore, the recently demonstrated periodic carry technique with multiple cells per weight [5] enables training to ideal accuracy. We also estimate that an 8-bit floating gate-based accelerator will have training energy, latency, and area advantages of 120×, 2.1×, and 5×, respectively, versus performing the same training tasks with an optimized SRAM-based ASIC.

In order to accelerate NN training using backpropagation, three kernels need to be accelerated: vector–matrix multiplication (VMM), matrix–vector multiplication (MVM), and outer product update (OPU) [2], as shown in Fig. 1. To accelerate both VMM and MVM, the source needs to be connected to the rows and the drain connected to the columns (or vice versa). During the OPU (parallel write), this configuration requires an access transistor for each memory cell to disconnect the drain from the rows. The access transistor prevents hot electron injection and junction breakdown. It also prevents large currents from flowing between the source and drain, which would cause unacceptable energy consumption and parasitic voltage drops in an array.

II. DEVICE CHARACTERIZATION

The silicon-oxygen-nitrogen-oxygen-silicon (SONOS) memory cell illustrated in Fig. 2 was fabricated and characterized. The binary memory operation is illustrated in Fig. 3. A reasonable *I–V* memory window is shown. Using longer

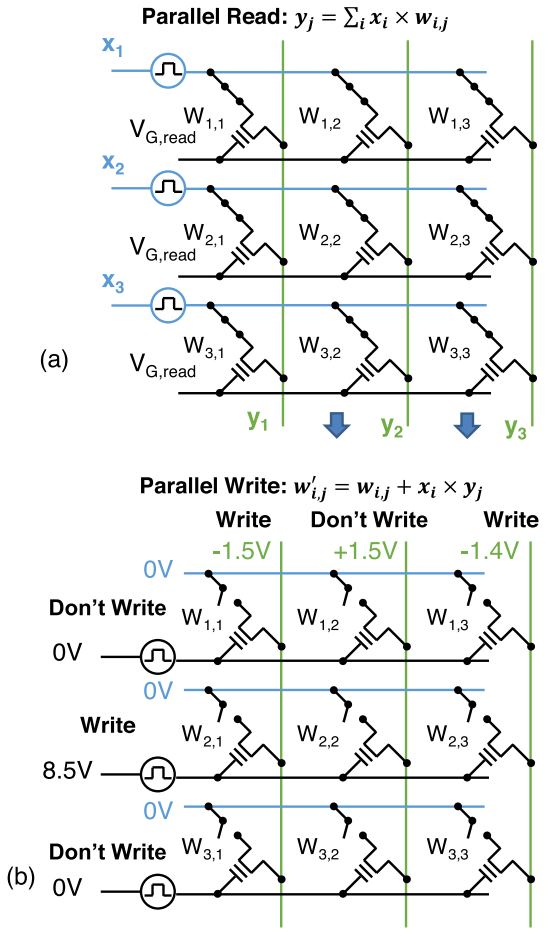

FIGURE 1. **(a) VMM is illustrated. A fixed read voltage is applied to all gate terminals. Access transistors (drawn as a switch) are biased on. Pulses of varying lengths are applied along the rows, and the resulting current is integrated along the columns. The transpose MVM can be performed by applying pulses to the columns and reading along the rows. (b) Parallel write, or OPU, for a 10 V write is shown with the corresponding biases labeled. The access transistors are open circuited. Selected devices see up to the full 10 V across V_{GS}, while unselected devices see a maximum of 7 V across V_{GS}. The last column has a write voltage of −1.4 V, applying 9.9 V across $W_{2,3}$ resulting in smaller state change than a full 10 V write. The amount written can be controlled by varying the voltage or pulse length.**

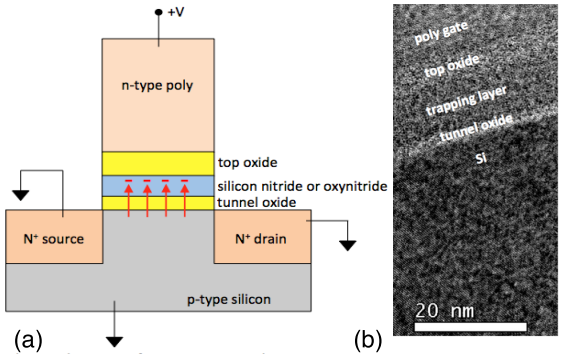

FIGURE 2. **(a) SONOS memory is schematically illustrated. (b) Transmission electron micrograph of the gate-stack is shown. The channel length of the device is 1.2 μm, and the channel width is 7 μm. The oxygen-nitrogen-oxygen (ONO) layer was grown in a tunnel oxidation furnace (VTR-20) in a dilute nitrous oxide (N_2O) atmosphere at 750 °C.**

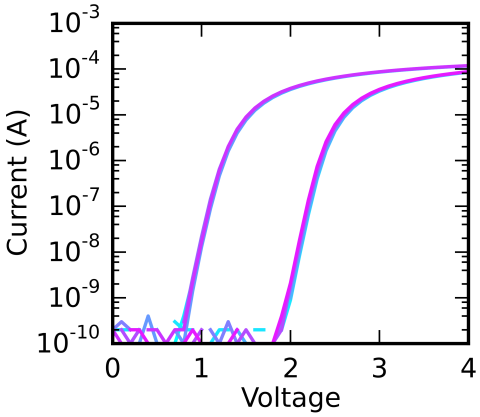

FIGURE 3. **Binary memory window of the SONOS cell is shown. Alternating −11 V, 2.5 ms erase pulses and 10 V, 2.5 ms program pulses are applied. The pulse lengths can be increased to further increase the memory window.**

write pulses or higher voltages can give a larger memory window. In Fig. 4, we characterize the analog properties of the device for different write voltages. The write voltage used determines the number of analog states and write linearity. Write pulses of $V_{GS} = -11$ V for 10 μs and $V_{GS} = +10$ V for 10 μs were chosen as the lowest voltages that give a reasonable G_{high}/G_{low} ratio and high linearity in the conductance versus pulse characteristic. The threshold shift during the analog write is illustrated in Fig. 5 and is only about 200 mV. This is because only a $\sim 10 \times G_{high}/G_{low}$ ratio is needed for analog operation.

To analyze the effect of drain bias while programing the cell in an array, we investigated the effects of different source–drain configurations, including $V_{DS} = 0$ V, $V_{DS} =$ 3 V, and floating/high-Z (Fig. 6). Ideally, $V_{DS} = 0$ during write. To achieve a condition close to this, an access transistor is used to float the drain, resulting in the drain floating condition. To see what would happen without an access transistor, $V_{DS} = \pm 3$ was also applied across the drain. Fortunately, changing V_{DS} does not significantly affect the state written as both the source and body are grounded. This indicates that there is potential for writing without an access transistor to float the drain. Nevertheless, we model an access device in subsequent area projections to eliminate parasitic currents during a write and to improve reliability by preventing hot electron injection. Eliminating the transistor would require redesigning the floating-gate cell to limit the on-state current to limit the parasitic currents during a write.

It has also been verified that unselected devices do not change state under partial gate-bias conditions, with $V_{GS} = -8$ V for erase and $V_{GS} = +7$ V for program, as illustrated in Fig. 7. The access transistor only must block half the difference between the selected and unselected write voltages, reducing the size requirement of this transistor.

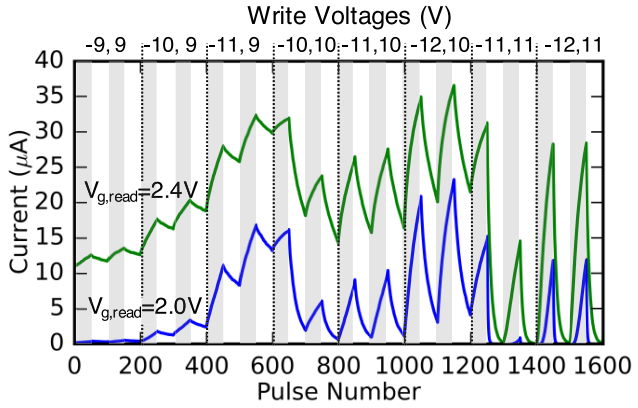

FIGURE 4. Alternating series of 50 erase pulses (gray shading), followed by 50 program pulses (white shading) are applied for different write voltages. $V_S = V_B = 0$ V and V_D is floating. After applying a write pulse, the conductance is measured at $V_{GS} = 2$ and 2.4 V and $V_{DS} = 0.1$ V. Write voltages of $V_{GS} = -11$ and 10 V give a reasonable on/off range and high write linearity. Increasing the erase voltage to −14 V broke the device.

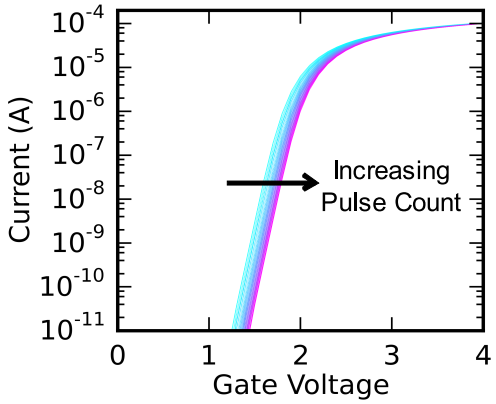

FIGURE 5. 50 10 V, 10 μs set pulses are applied, and an *I–V* is measured after each pulse. During the analog write, the threshold only shifts by about 200 mV, instead of the full 1–2 V of a memory write.

If the write voltage is $V_{GS} = 10$ V and the unselected write voltage is 7 V, the access transistor will have to hold off 1.5 V.

The key limitation in NN training accuracy is the asymmetric nonlinearities during a write [3]. With an asymmetric nonlinearity, alternating program and erase pulses that can occur at the end of training cause the weight to decay to a midpoint value. Nevertheless, NN can train to high accuracy with symmetric write nonlinearities [3]. To optimize the write nonlinearity, the gate read voltage needs to be optimized, as shown in Fig. 8. Choosing the correct read gate voltage will have a dramatic impact on the NN work accuracy. As $V_{G,\text{read}}$ is lowered from 2.6 to 1.4 V, the nonlinearity changes from an asymmetric nonlinearity to a symmetric linearity. By lowering $V_{G,\text{read}}$, the device is operating in the subthreshold regime. In this regime, the magnitude of the change in conductance after a write pulse primarily depends on the starting state and not the sign of the write voltage. Achieving a symmetric nonlinearity is critical to enabling high-accuracy training of NNs.

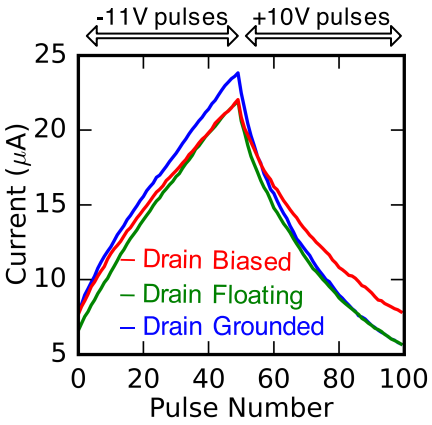

FIGURE 6. Changing the drain bias during a write does not affect the write properties. In all cases, $V_{body} = 0$ V. The biased case corresponds to the conditions without an access device: $V_S = 0$ and $V_D = 3$ V during program and −3 V during erase.

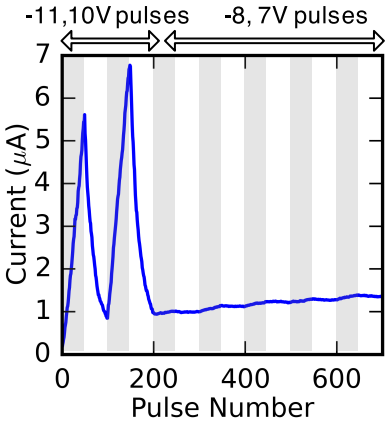

FIGURE 7. Alternating series of erase (gray shading) and program (white shading) pulses are applied. Lowering the write voltage from $V_{GS} = -11$ and 10 to −8 and 7 V inhibits significant state change.

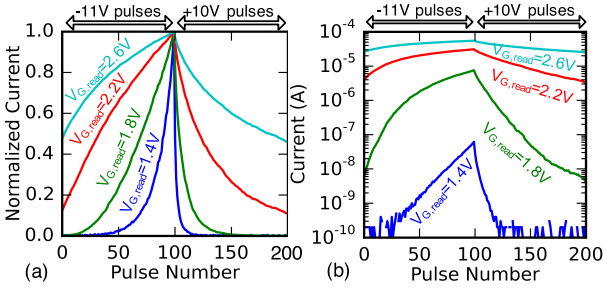

FIGURE 8. 100 erase pulses ($V_{GS} = -11$ V, 10 μs) followed by 100 program pulses ($V_{GS} = 10$ V, 10 μs) are applied, and the current is measured at different read gate voltages. (a) Normalized current and (b) absolute value of the current are shown. Decreasing $V_{G,\text{read}}$ significantly reduces the write nonlinearity and changes the nonlinearity from an asymmetric nonlinearity to a symmetric nonlinearity.

To characterize the analog statistics, a series of increasing and decreasing pulses were applied, as illustrated in Figs. 9–11. The conductance versus pulse number is plotted in Fig. 9. In Fig. 10, the conductance change at

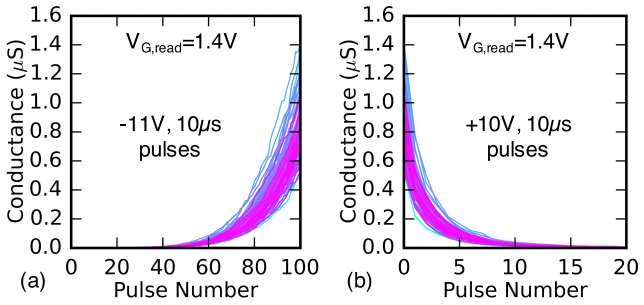

FIGURE 9. **Alternating series of 100 erase pulses followed by 100 program pulses are applied. The conductance after each pulse is read at V_{DS} = 100 mV, and the measurement is repeated 50 times to collect statistics.**

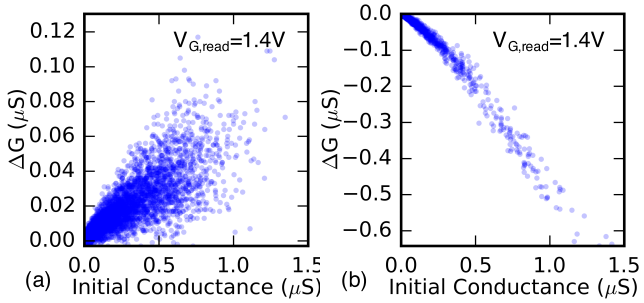

FIGURE 10. **At $V_{G,read}$ = 1.4 V, the conductance change is symmetric between program and erase (small changes at a low initial state and large changes at a high initial state) leading to a high training accuracy.**

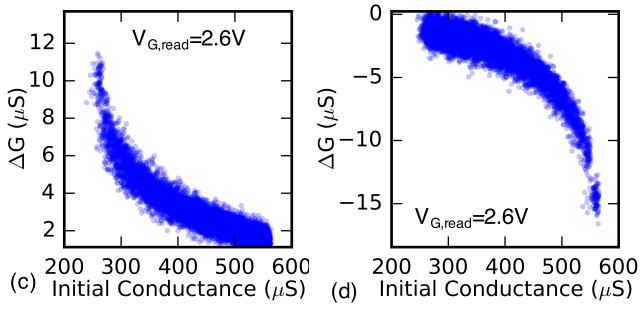

FIGURE 11. **At $V_{G,read}$ = 2.6 V, the conductance change in asymmetric between program and erase leading to lower training accuracy.**

$V_{G,\text{read}}$ = 1.4 V for different starting conductances is extracted from the pulsing data shown in Fig. 9. We see the symmetric write nonlinearity where the conductance change is directly proportional to the starting state. In Fig. 11(a), at $V_{G,\text{read}}$ = 2.6 V, this reverses resulting in an asymmetric nonlinearity. The asymmetric nonlinearity results in significantly lower training accuracies.

A remaining challenge is to understand analog endurance in a floating-gate device. A typical analog write pulse is only 0.1% or less of the length of a digital memory pulse [3], potentially increasing the endurance by three orders of magnitude or more. Furthermore, NN training is also resilient to occasional device failure [4]. If needed, it is also possible to tradeoff retention for endurance.

TABLE 1. **A/D and D/A converter properties.**

	Range	Bits
Row Input	-1 to 1	8
Col Output	-6 to 6	8
Col Input	-1 to 1	8
Row Output	-4 to 4	8
Row Update	-0.01 to 0.01	7
Col Update	-1 to 1	5

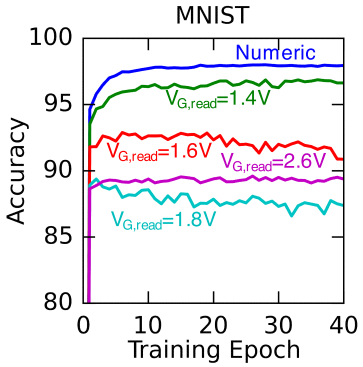

FIGURE 12. **The lower the read voltage is, the higher the training accuracy is. A single device is used per weight.**

TABLE 2. **Data set properties.**

Data set	#Training/Test Examples	Network Size
File Types[7]	4,501 / 900	256×512×9
MNIST [8]	60,000 /10,000	784×300×10

III. NEURAL NETWORK SIMULATION

To simulate the accuracy of an NN based on this SONOS device, a detailed system simulation was performed in CrossSim [3], [7], Sandia's analog crossbar simulator. We model the general purpose neuromorphic system in [3] where crossbars are used to perform matrix operations in analog, and the inputs and outputs are processed in digital. This requires digital-to-analog (D/A) and analog-to-digital (A/D) converters at the inputs and outputs as specified in Table 1. The bit precision and algorithmic input–output ranges used are given. They have a negligible (0.2%) impact on accuracy [5]. In order to model negative weights, a single device per weight is initially used, and a reference current is subtracted [3]. Two different two-layer NNs, summarized in Table 2, are simulated [11], [12]. Simulation details are explained in the supplementary information of [7]. It is assumed that write voltages or pulse lengths can be scaled to vary the amount written.

As shown in Fig. 12, by choosing the correct gate voltage, a good accuracy of 96.9% is achievable on MNIST. Representing negative numbers by taking the difference between two devices averages out some of the noise and increases the accuracy to 97.6% on MNIST. Using two devices per digit to represent negative numbers and two digits to represent a weight

TABLE 3. Area comparisons.

	8 bit	4 bit	2 bit
SRAM (μm^2)	836,000	814,000	800,000
ReRAM (μm^2)	75,000	46,000	41,000
SONOS (μm^2)	195,000	166,000	161,000

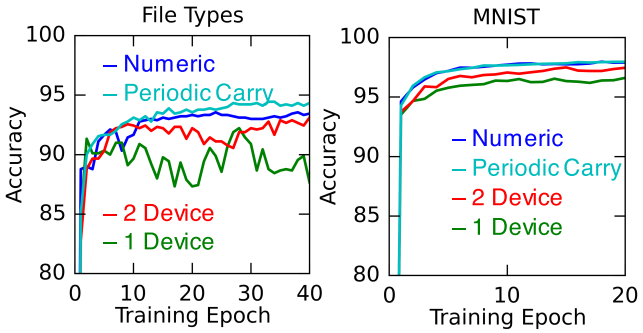

FIGURE 13. Training an NN with the SONOS device can reach good accuracies of around 96% on MNIST when using a single device but can reach ideal accuracies when using multiple devices with periodic carry. The one-device architecture uses a single device to represent a weight and subtracts a reference current. The two-device architecture takes the difference between two devices to represent negative numbers. The periodic carry architecture also uses two devices for the file-type data set and four devices for MNIST.

with periodic carry [5], an ideal device accuracy of 98.0% can be achieved, as shown in Fig. 13. We use a base 8, two-digit number system where the first digit represents numbers eight times larger than the second digit. Periodic carry allows one to take the advantage of both a parallel write and a place value number system. Normally, a carry must be computed after every addition if using multiple digits. This eliminates the benefit of the parallel update. Allowing for a part of an analog device's conductance range to represent a carry allows the carry from the second digit to the first digit to be computed only once every 1000 updates, thereby averaging out the cost of reading each memory element and adjusting the weights to perform a carry. We dedicate 50% of the conductance range of the lowest order digit to representing the carry.

For the file-type data set, only a single device is needed per digit, and using periodic carry actually results in higher accuracy than the numeric floating-point calculation (likely due to noise finding a more optimal solution).

IV. ARCHITECTURAL EVALUATION

One of the key drawbacks of using a floating-gate memory for an analog accelerator is that it requires a far larger area and voltage versus a ReRAM. Nevertheless, it is still possible to achieve significant system-level advantages relative to an optimized digital SRAM-based ASIC. To understand this, the architectural-level analysis in [1] was modified to use a 1024 × 1024 SONOS array. The energy, area, and latency of a neural core that performs the three key matrix operations, VMM, MVM, and OPU, were modeled. A 14-/16-nm process was modeled for the digital logic and interconnects. We assume that the SONOS cell can scale to 28 nm and estimate a gate capacitance of 100 aF and cell area of 0.053 μm^2 based on existing 28-nm floating-gate transistors [13], [14]. We also assume that it is possible to optimize the channel to give the high resistance (100 MΩ) needed for large-scale arrays. The access transistors are assumed to have the same area and capacitance as the floating-gate cell. Finally, writing the array requires large high-voltage transistors that can support 11 V. Based on [15], high-voltage vertical transistors can be fabricated in an area of 1.44 μm^2 and a capacitance of 7.44 fF. These transistors are 9% of the core area. If needed, larger planar high-voltage transistors can be used without drastically changing the overall area. We assume that a future process will be able to integrate the needed transistors on a single substrate as commercial 28-nm embedded flash is already in development. The ReRAM- and SRAM-based accelerators and device properties are described in detail in [1]. The SRAM-based accelerator is based on a 1-MB cache synthesized using a cache generator targeting the 14-/16-nm PDK. The ReRAM is assumed to have a 100-MΩ on state, 35-aF capacitance, 10× on/off ratio, and a 1.8 V write voltage. The resulting energy, area, and latency relative to digital SRAM-based accelerator and analog ReRAM-based accelerator are summarized in Tables 3 and 4 for the accelerator. For an 8-bit floating-gate training accelerator, 70% of the write energy is due to the CV^2 energy of charging wires to 10 or 11 V. The very low write currents result in negligible contributions to the write energy. The SONOS read latency is comparable to ReRAM as the timing is dominated

TABLE 4. Energy and latency comparisons.

	VMM			MVM			OPU			Total		
	8 bit	4 bit	2 bit	8 bit	4 bit	2 bit	8 bit	4 bit	2 bit	8 bit	4 bit	2 bit
Energy – SRAM (nJ)	2850	2237	1848	4855	4241	3852	4300	3673	3274	12,000	10,150	8974
Energy – ReRAM (nJ)	12.8	1.00	0.44	12.8	1.00	0.44	2.2	1.00	0.46	27.9	2.66	1.35
Energy – SONOS (nJ)	14.4	2.25	1.5	14.4	2.25	1.5	71.5	30.9	10.6	100	35.4	13.6
Latency – SRAM (µs)	4	4	4	32	32	32	8	8	8	44	44	44
Latency – ReRAM (µs)	0.384	0.024	0.011	0.384	0.024	0.011	0.512	0.032	0.032	1.28	0.080	0.054
Latency – SONOS (µs)	0.402	0.032	0.014	0.402	0.032	0.014	20	20	20	20.80	20.06	20.02

by the A/D and D/A converters. However, 96% of the total latency is due to the slow write speed of SONOS. Nevertheless, the large parallelism afforded by an analog accelerator allows for the total SONOS latency to still be 2× faster than an SRAM-based accelerator. Latency can be decreased by trading off retention for a faster write or by using a device with a steeper subthreshold swing that allows for a larger conductance change with a smaller threshold shift.

Only 57% of the area is due to the SONOS cell and the access transistor, indicating that the array area is reasonably balanced with the area of the rest of the circuitry. If higher area efficiency is desired, two 3-D integration options can be explored. High-density (1.8 μm pitch) face-to-face interconnects [16] could be used to connect two wafers, one with digital logic and one with high-voltage and floating-gate transistors to reduce the area by 50%. The 3-D interconnect capacitance would be less than the row or column capacitance in the SONOS array. Following [17], 3-D nand arrays could also be used to store multiple layers of an NN in the same 2-D area. Each individual SONOS cell shown in Fig. 1 could be replaced by a column in a 3-D nand array.

V. CONCLUSION

Floating-gate memories, currently available in commercial foundries, are a compeling near-term option for analog training accelerators. This paper has demonstrated lower write noise and write nonlinearity than alternative resistive memories, allowing for training to ideal accuracies on MNIST. Despite the high voltage and slow writes, the energy, area, and latency of an 8-bit floating-gate neural accelerator is still 120×, 5.0×, and 2.1× better, respectively, than an optimized digital ASIC counterpart. The high accuracies are enabled by operating the devices in the subthreshold regime giving symmetric write nonlinearities. Any three-terminal transistor-based device should be able to operate in this favorable regime.

ACKNOWLEDGMENT

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. This paper describes objective technical results and analysis. Any subjective opinions do not necessarily represent the views of the U.S. Department of Energy or the U.S. Government.

REFERENCES

[1] M. J. Marinella *et al.*, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 86–101, Mar. 2018.

[2] S. Agarwal *et al.*, "Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding," *Front. Neurosci.*, vol. 9, p. 484, Jan. 2016.

[3] S. Agarwal *et al.*, "Resistive memory device requirements for a neural algorithm accelerator," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2016, pp. 929–938.

[4] G. W. Burr *et al.*, "Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015.

[5] S. Agarwal *et al.*, "Achieving ideal accuracies in analog neuromorphic computing using periodic carry," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2017, pp. T174–T175.

[6] I. Boybat *et al.*, "Improved deep neural network hardware-accelerators based on non-volatile-memory: The local gains technique," in *Proc. IEEE Int. Conf. Rebooting Comput. (ICRC)*, Nov. 2017, pp. 1–8.

[7] E. J. Fuller *et al.*, "Li-ion synaptic transistor for low power analog computing," *Adv. Mater.*, vol. 29, no. 4, 2017, Art. no. 1604310.

[8] Y. Van de Burgt *et al.*, "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing," *Nature Mater.*, vol. 16, no. 4, pp. 414–418, Apr. 2017.

[9] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *IEDM Tech. Dig.*, Dec. 2017, pp. 6.5.1–6.5.4.

[10] S. Ramakrishnan and J. Hasler, "Vector-matrix multiply and winner-take-all as an analog classifier," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 2, pp. 353–361, Feb. 2014.

[11] Y. LeCun, C. Cortes, and C. J. Burges. *The MNIST Database of Handwritten Digits*. Accessed: Jan. 12, 2018. [Online]. Available: http://yann.lecun.com/exdb/mnist

[12] J. A. Cox, C. D. James, and J. B. Aimone, "A signal processing approach for cyber data classification with deep neural networks," *Procedia Comput. Sci.*, vol. 61, pp. 349–354, Nov. 2015.

[13] Y. K. Lee *et al.*, "High-speed and logic-compatible split-gate embedded flash on 28-nm low-power HKMG logic process," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2017, pp. T202–T203.

[14] Y. Taito *et al.*, "A 28 nm embedded split-gate MONOS (SG-MONOS) flash macro for automotive achieving 6.4 GB/s read throughput by 200 MHz no-wait read operation and 2.0 MB/s write throughput at Tj of 170°C," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 213–221, Jan. 2016.

[15] K. Sakui and T. Endoh, "Design impacts on NAND flash memory core circuits with vertical MOSFETs," in *Proc. IEEE Int. Memory Workshop*, May 2010, pp. 1–4.

[16] S.-W. Kim *et al.*, "Ultra-fine pitch 3D integration using face-to-face hybrid wafer bonding combined with a via-middle through-silicon-via process," in *Proc. IEEE 66th Electron. Compon. Technol. Conf. (ECTC)*, May/Jun. 2016, pp. 1179–1185.

[17] P. Wang *et al.*, "Three-dimensional nand flash for vector-matrix multiplication," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8571188&isnumber=4359553. doi: 10.1109/TVLSI.2018.2882194.

Authors' photographs and biographies not available at the time of publication.