Using Floating Gate Memory to Train Ideal Accuracy Neural Networks

Sapan Agarwal, Member, IEEE, Diana Garland, John Niroula, Robin B. Jacobs-Gedrim, Member, IEEE, Alex Hsia, Michael S. Van Heukelom, Elliot Fuller, Bruce Draper, Matthew J. Marinella, Senior Member, IEEE

Abstract—Floating gate SONOS (Silicon-Oxygen-Nitrogen-Oxygen-Silicon) transistors can be used to train neural networks to ideal accuracies that match those of floating point digital weights on the MNIST dataset 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 120X, operate 2.1X faster and require 5X lower area than an optimized SRAM based ASIC.

Index Terms—neuromorphic, analog, SONOS, flash, neural network, floating gate, memristor, training

I. INTRODUCTION

Analog accelerators promise to improve the energy and latency of training a neural network (NN) by more than a 100X 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 prevents the weights from being learned to high accuracy[3, 4]. Algorithmic and circuit techniques help improve accuracy[5, 6], but neural network 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 a neural network to within 1% of that achieved with floating point weights on MNIST dataset (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

Submitted Dec 31st, 2018. This work was supported by Sandia National Laboratories’ Laboratory Directed Research and Development Program. 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-NA0003525.

S. Agarwal (email: sagarw@sandia.gov) and E. Fuller are with Sandia National Laboratories, Livermore, CA

D. Garland, J. Niroula, R. B., Jacobs-Gedrim, A. Hsia, M. S. Van Heukelom, B. Draper, and M. Marinella (email: mmarine@sandia.gov) are with Sandia National Laboratories, Albuquerque, NM
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 120X, 2.1X, and 5X 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 SONOS (Silicon-Oxygen-Nitrogen-Oxygen-Silicon) memory cell illustrated in Fig 2 was fabricated and characterized. The binary memory operation is illustrated in Fig 3. A reasonable 1V memory window is shown. Using longer 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 VGS = -11V, 10V give a reasonable on/off range and high write linearity. Increasing the erase voltage to -14V broke the device.

In Fig 5, 50 10V, 10µs 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 1-2V of a memory write.
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. If the write voltage is $V_{GS}=10$V and the unselected write voltage is 7V, the access transistor will have to hold off 1.5V.

The key limitation in neural network 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, neural networks 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 neural network work accuracy. As $V_{G,\text{read}}$ is lowered from 2.6V to 1.4V, 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 neural networks.

To characterize the analog statistics, a series of increasing and decreasing pulses were applied as illustrated in Figs 9-11. The conductance after each pulse is read at $V_{DS}=100$ mV and the measurement is repeated 50 times to collect statistics.

A remaining challenge is understanding 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, neural network training is also resilient to occasional device failure[4]. If needed, it’s also possible to tradeoff retention for endurance.

**III. NEURAL NETWORK SIMULATION**

To simulate the accuracy of a neural network 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 I. 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 reference current is subtracted[3]. Two different two-layer neural networks, summarized in Table II, are simulated [11, 12]. Simulation details are explained in the supplementary information of [7]. It’s assumed that write voltages or pulse lengths can be scaled to vary the amount written.

As seen 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 with periodic carry[5] an ideal device accuracy of 98.0% can be achieved as shown in Fig 13. We use a base 8, 2-digit number system where the first digit represents numbers 8 times larger than the second digit. Periodic carry allows one to take 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 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 types dataset, only a single device is needed per digit and using periodic carry actually results in a higher accuracy.
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 1024x1024 SONOS array. The energy, area, and latency of a neural core that performs the three key matrix operations, VMM, MVM, and OPU was modelled. A 14/16nm process was modelled for the digital logic and interconnects. We assume the SONOS cell can scale to 28 nm and estimate a gate capacitance of 100aF and cell area of 0.053µm² based on existing 28nm SONOS cell can scale to 28 nm and estimate a gate capacitance of 100aF and cell area of 0.053µm² based on existing 28nm vertical transistors can be fabricated in an area of 1.44 µm² and 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 a future process will be able to integrate the needed transistors on a single substrate as commercial 28nm 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 1MB 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, 10X on/off ratio and a 1.8V write voltage. The resulting energy, area, and latency relative to Digital SRAM and Analog ReRAM based accelerators is summarized in Tables III & IV for the accelerator. For an eight-bit floating gate training accelerator, 70% of the write energy is due to the CV² energy of charging wires to 10 or 11V. 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 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 2X 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 access transistor, indicating that the area is reasonably balanced with the area of the rest of the circuitry. If higher area efficiency is desired two 3D integration options can be explored. High density (1.8µm 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 3D interconnect capacitance would be less than the row or column capacitance in the SONOS array. Following [17], 3D NAND arrays could also be used to store multiple layers of a neural network in the same 2D area. Each individual SONOS cell in Fig 1 could be replaced by a column in a 3D NAND array.

V. CONCLUSION

Floating gate memories, currently available in commercial foundries, are a compelling near-term option for analog training accelerators. This work 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 120X, 5.0X and 2.1X 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.

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 Journal on Emerging and Selected Topics in Circuits and Systems, 2018.

[2] S. Agarwal et al., "Energy Scaling Advantages of Resistive Memory Crossbar Based Computation and its Application to Sparse Coding," Frontiers in Neuroscience, vol. 9, pp. 484, 2016, Art. no. 484.

[3] S. Agarwal et al., "Resistive memory device requirements for a neural algorithm accelerator," in 2016 International Joint Conference on Neural Networks (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," Electron Devices, IEEE Transactions on, vol. 62, no. 11, pp. 3498-3507, 2015.

[5] S. Agarwal et al., "Achieving ideal accuracies in analog neuromorphic computing using periodic carry," in VLSI Technology, 2017 Symposium on, 2017, pp. T174-T175: IEEE.
[6] I. Boybat et al., "Improved Deep Neural Network Hardware-Accelerators Based on Non-Volatile-Memory: The Local Gains Technique," in 2017 IEEE International Conference on Rebooting Computing (ICRC), 2017, pp. 1-8.

[7] E. J. Fuller et al., "Li-Ion Synaptic Transistor for Low Power Analog Computing," Advanced Materials, vol. 29, no. 4, p. 1604310, 2017.

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

[9] X. Guo et al., "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in Electron Devices Meeting (IEDM), 2017 IEEE International, 2017, pp. 6.5.1-6.5.4: IEEE.

[10] S. Ramakrishnan and J. Hasler, "Vector-Matrix Multiply and Winner-Take-All as an Analog Classifier," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 22, no. 2, pp. 353-361, 2014.

[11] Y. LeCun, C. Cortes, and C. J. Burges, The MNIST database of handwritten digits [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 Computer Science, vol. 61, pp. 349-354, // 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 VLSI Technology, 2017 Symposium on, 2017, pp. T202-T203: IEEE.

[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 Journal of Solid-State Circuits, vol. 51, no. 1, pp. 213-221, 2016.

[15] K. Sakui and T. Endoh, "Design impacts on NAND Flash memory core circuits with vertical MOSFETs," in 2010 IEEE International Memory Workshop, 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 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), 2016, pp. 1179-1185.

[17] P. Wang et al., "Three-Dimensional nand Flash for Vector-Matrix Multiplication," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, pp. 1-4, 2018.