Technological Benchmark of Analog Synaptic Devices for Neuroinspired Architectures

Pai-Yu Chen
Arizona State University

Shimeng Yu
Georgia Institute of Technology

Editor's note:
In this article, the authors present a circuit-level macro model ("NeuroSim" simulator) to estimate circuit-level performance of neuroinspired architectures to facilitate design space exploration. The model is used to analyze the impact of analog synapse device characteristics on the performance of a two-layer multilayer perceptron (MLP) neural network and identify critical device properties (on/off ratio and asymmetry, in this case) to guide technology development.
—An Chen, Semiconductor Research Corporation

■ **Recent advances in** machine/deep learning algorithms have achieved tremendous success in speech and image recognition, implemented with conventional graphics processing units and/or field programmable gate arrays. Because of the limited on-chip memory resources, traditional von Neumann architecture is inadequate for fast training and/or real-time classification. In recent years, several custom CMOS ASIC hardware accelerators have been developed (e.g., MIT's Eyeriss [1] and Google's tensor processing unit [2]), where SRAM is used to implement the synapses. Typically, the weight information of a single synapse is stored using multiple binary SRAM cells, which is area-inefficient (with cell size $100F^2$~$200F^2$, F is the lithography feature size). As a result, part of the weights may have to be stored off-chip (i.e., in DRAM), introducing the bottleneck of the off-chip memory access. To replace SRAM, analog synaptic devices (or analog synapses) are considered as promising candidates due to their compact device structure and the ability to store "analog" weights. One type of analog synapses is based on emerging nonvolatile memory (eNVM) devices, e.g., resistive memories [3]–[6] and phase change memory (PCM) [7], [8]. These eNVM-based synapses are two-terminal with cell size $4F^2$~$12F^2$, and they can represent the analog weight with their multilevel conductance states, where the transition between conductance states are triggered by electrical inputs. Its detailed physical mechanism can be different for different types of eNVM devices. Generally, the conductance can be increased and decreased with positive and negative programming voltage pulses. Recently, there has been remarkable progress in the array-level demonstration of the essential neural computation operations including weighted sum and weight update [9]–[12]. The other type of analog synapses is based on ferroelectric field-effect transistor (FeFET) [13], [14]. FeFET synapses are three-terminal

Digital Object Identifier 10.1109/MDAT.2018.2890229

Date of publication: 28 December 2018; date of current version: 29 May 2019.

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

like a conventional transistor, but with its gate dielectric replaced by a ferroelectric material that has multiple domains of polarization. With programming voltage pulses applied on the gate, part of the polarization direction can be changed, enabling gradual tuning of the threshold voltage and, thereby, the channel conductance to store analog weights.

The most compact and simplest array structure to form a weight matrix is the crossbar array structure with eNVM analog synapses [15]. As shown in Figure 1a, the weighted sum (matrix-vector multiplication) can be performed in a parallel fashion with the input vector being the voltages and the weighted sum being the output currents [16]. An ideal analog synaptic device behavior assumes a linear update of the conductance (or weight) with the programming voltage pulses. As shown in Figure 1b, however, the realistic devices reported in the literature do not follow such ideal trajectory, exhibiting "nonideal" properties such as nonlinear and asymmetric weight increase/decrease, limited precision, and finite ON/OFF ratio. Such nonideal behaviors commonly exist in today's synaptic devices. In this article, we use the developed NeuroSim+ simulator [17] to analyze the impact of these nonideal device properties on the learning accuracy and investigate the design tradeoffs with SRAM and analog synapses-based neuroinspired architectures. The new materials presented in this article, beyond [17], include more synaptic device candidates such as PCM and FeFET; and an improved weight update scheme to skip the unnecessary weight update operations in order to reduce the latency and energy of analog synapses.

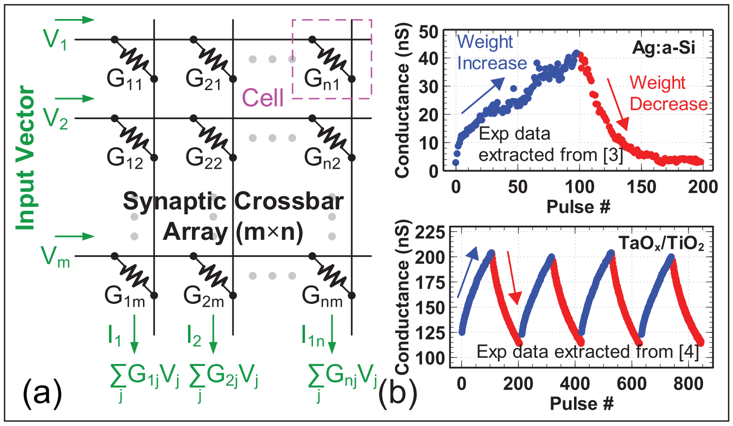

Figure 1. (a) Weighted-sum operation in an eNVM-based synaptic crossbar array structure. (b) Weight update of the eNVM-based synaptic devices.

NeuroSim architecture

Overview

NeuroSim is a circuit-level macro model developed in C++ that can be used to estimate the area, latency, dynamic energy, and leakage power of neuromorphic hardware accelerators to facilitate the design space exploration. The framework of NeuroSim follows the principles of CACTI [18] for SRAM cache and NVSim [19] for NVM, while NeuroSim is dedicated to support neuroinspired architectures. The hierarchy of the simulator consists of different levels of abstraction from the memory cell parameters and transistor technology parameters, to the gate-level subcircuit modules and then to the array architecture. Figure 2a shows an overview of the high-level architecture with neuromorphic hardware accelerator to implement neural networks (NNs). A synaptic core is specifically designed for the weighted sum and weight update. The computation within the core could be analog, but the digital communication is enforced between the cores. The synaptic array is the core unit of weighted-sum computation and the array periphery helps transform the results to digital [with analog-to-digital converter (ADC)]. On the other hand, the neuron periphery is responsible for nonlinear activation function and communication between synaptic cores.

Circuit architectures of synaptic cores

NeuroSim supports both SRAM, analog eNVM, and analog FeFET synaptic cores, which are shown in Figure 2b–d. As SRAM cells can only store binary bits, we group multiple SRAM cells along the row as one synapse to represent a higher weight precision. Similar to conventional SRAM for read and write, the weighted-sum and weight update operation in the SRAM synaptic core are based on row-by-row fashion. The input vector is encoded using multiple clock cycles to represent its precision. For each row, an input vector bit of 1 means the row will be selected for read; otherwise, the row will be skipped. After the memory data are read by the sense amplifier (S/A), adder and register are used to accumulate the partial weighted sum in a row-by-row fashion. The adder and shift register pair at the bottom performs shift and add of the weighted-sum result at each input vector bit cycle to get the final weighted sum. For the write operation, all the cells on the same row can be updated at the same time, and the new weight will be provided from the input of the write driver.

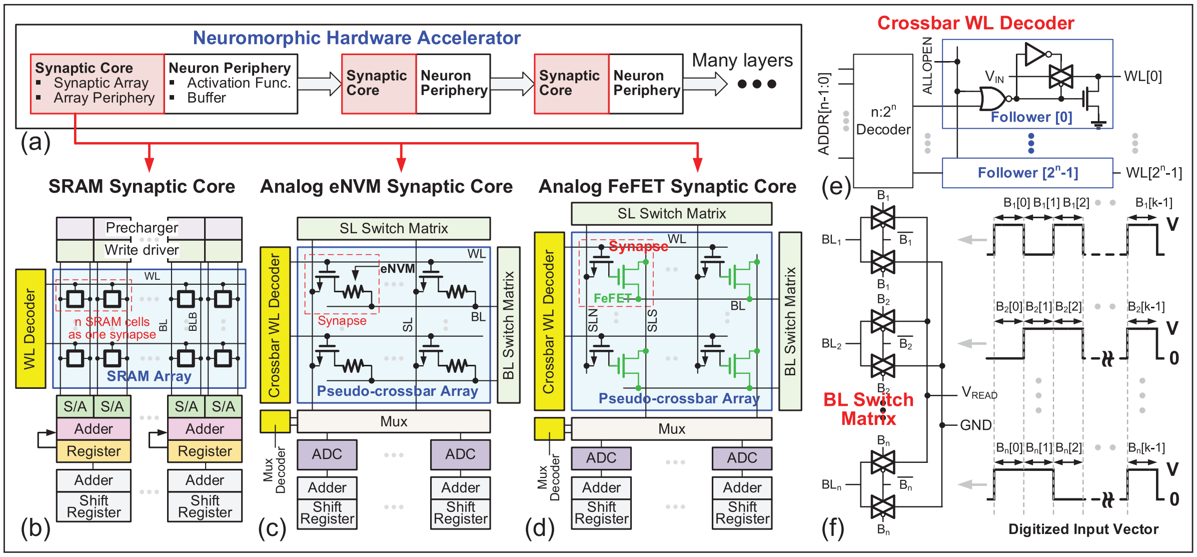

Figure 2. (a) Overview of high-level architecture with neuromorphic hardware accelerator. (b)–(d) Circuit block diagram of SRAM, analog eNVM, and analog FeFET synaptic cores. (e) Circuit schematic of the crossbar WL decoder. (f) Circuit schematic of the BL switch matrix and its control signals in a binary bit stream to represent the precision of input vector.

In NeuroSim, the analog eNVM synaptic core supports the pseudocrossbar array architectures. As shown in Figure 2c, this architecture is modified from the conventional one-transistor one-resistor array, with perpendicular bit lines (BLs) and source lines (SLs) to enable the weighted-sum operation [20], where the transistor can prevent write interference between cells. The word line (WL) decoder is also modified to be "crossbar WL decoder" by attaching the follower circuits to every output row of the traditional decoder [21], as shown in Figure 2e. If *ALLOPEN* = 1, the crossbar WL decoder will activate all the WLs; otherwise, it will function as a traditional WL decoder. The switch matrix consists of transmission gates that are connected to each row or column, and the input vector signal in a binary bit stream can control these transmission gates to enable multiple rows or columns, as shown in Figure 2d. In this way, the read voltage (V_{READ}) can pass to the BLs and the weighted sums are read out through SLs in parallel. To convert these analog weighted-sum currents to digital outputs, we use the read circuit [22] as the ADC to employ the principle of the integrate-and-fire neuron model. As the size of ADC is typically larger than the column pitch, multiple columns may share one ADC to improve the area efficiency. However, this inevitably increases the weighted-sum latency as multiple-read-cycle (or time multiplexing) is needed. The weight update in eNVM is performed row by row with the write pulses coming from the SL switch matrix. The weight update requires two phases because there are two voltage polarities for weight increase and decrease. On the other hand, the FeFET synaptic core is different from the eNVM one in the synaptic array structure, as shown in Figure 2d. It also has an access transistor for each cell to prevent programming on other unselected rows during row-by-row weight update. As FeFET is a three-terminal device, it needs two separate SLs: source linvve south for the weighted sum and source line north for the weight update, respectively.

Transistor and memory cell models

At the device level, NeuroSim is featured with various design options in transistor technologies and synaptic devices. The transistors can be configured to be high-performance or low-standby-power type with different technology nodes from 130 nm down to 7 nm, where FinFET is used at 14 nm and beyond. The transistor models are calibrated based on predictive technology model [23]. Important parameters in transistor models include device W/L, the operating and threshold voltage, gate and parasitic capacitance (per unit area), and

NMOS/PMOS saturation/off current density across different temperatures. Based on these parameters, the area and intrinsic RC model of standard logic gates (INV, NAND, and NOR) can be calculated, and thus, the circuit-level performance metrics of each subcircuit module can be estimated. The design of SRAM, eNVM, and FeFET cells in NeuroSim is also flexible. We use conventional 6T SRAM, where all transistors' W/L can be adjusted. On the other hand, eNVM and FeFET cells have parameters such as max/min conductance, read/write voltage and pulsewidth, number of multilevel (precision), and/or *I*–*V* nonlinearity degree.

Benchmark with NeuroSim+ framework

To benchmark the performance of various synaptic devices on NNs, NeuroSim is integrated with a two-layer multilayer perceptron (MLP) NN with synaptic device properties incorporated into the weights. The entire framework is named "NeuroSim+," which is able to evaluate the online learning accuracy as well as the circuit-level performance such as area, latency, energy, and leakage at the run-time of the algorithm [17].

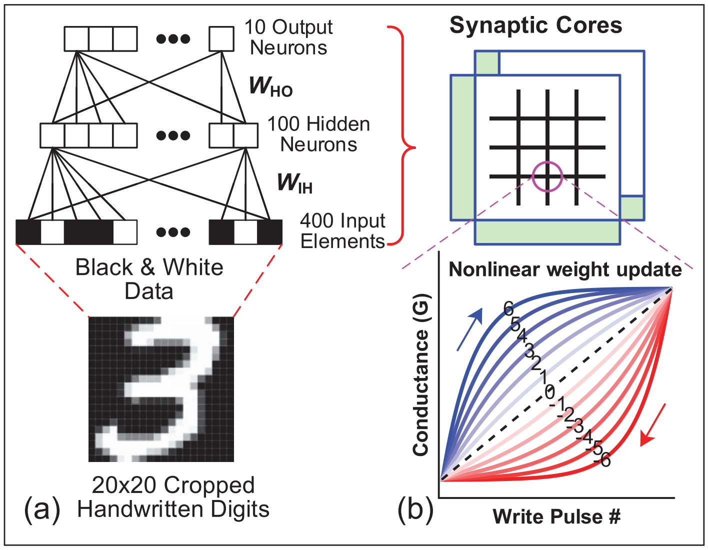

Figure 3. (a) The two-layer MLP NN. The input MNIST images are cropped and encoded into black/white data for simplification. (b) In NeuroSim+, the weights W_{IH} and W_{HO} are mapped to synaptic cores. For analog synapses, the device properties include the nonlinear weight update (shown here), conductance ON/OFF ratio, cycle-to-cycle weight update variation, and so on.

Adapt MLP network to hardware

As shown in Figure 3a, we use MNIST handwritten digits [20] as the training and testing data sets in the MLP network, and the network topology is 400(input layer) –100(hidden layer) –10(output layer), where 400 neurons of the input layer correspond to a 20 × 20 MNIST image (converted to black/white and edge cropped), and 10 neurons of the output layer correspond to 10 classes of digits. The two-layer MLP NN is mapped to two synaptic cores (Figure 3b) with neuron peripheries serving as connection paths between the cores (not shown). Such a simple two-layer MLP can achieve 96%~97% in online learning in the software baseline, which is not as high as reported (>98%) [24] due to the simplicity made for hardware implementation, where the neuron node is modularized to truncate the weighted sum to 1-bit output value through a low-precision activation function for the input of the next neuron node. Moreover, it should be noted that the synaptic devices can only represent positive weights, thus a mapping from the algorithm's weight (–1~1) to device's weight (0~1) is required. In neuron peripheral circuits, the array's weighted-sum result will be mapped back to the algorithm's weighted-sum result by subtracting the sum of input vector elements.

Impact of nonideal synaptic device properties

For analog synapses, we consider several nonideal synaptic device properties. To analyze the effect of nonlinear weight update, we define a set of nonlinear curves labeled with nonlinearity values from 6 to –6 for both the potentiation (weight increase) and depression (weight decrease), as shown in Figure 3b. In particular, the positive conductance change (G+) and negative conductance change (G–) with the number of pulses (P) are described with the following equations:

$$G_+ = B\left(1-e^{\left(-\frac{P}{A}\right)}\right)+G_{\min} \tag{1}$$

$$G_- = -B\left(1-e^{\left(\frac{P-P_{\max}}{A}\right)}\right)+G_{\max} \tag{2}$$

$$B = (G_{\max}-G_{\min})/\left(1-e^{\frac{-P_{\max}}{A}}\right), \tag{3}$$

where $G_{\max}$, $G_{\min}$, and $P_{\max}$ are directly extracted from the experimental data, which represents the maximum conductance, minimum conductance, and the maximum pulse number required to switch the device between the minimum and maximum conductance states; *A* is the parameter that controls the nonlinear behavior of weight update; and B is simply a function of A that fits the functions within the range of $G_{\max}$, $G_{\min}$, and $P_{\max}$. A and B may be different in (1) and (2).

The potentiation and depression will not necessarily follow the same trajectory, resulting in asymmetry with the positive nonlinearity value for potentiation and negative nonlinearity for depression. More experimental results of eNVMs today [3]–[6] show asymmetry in that the potentiation and the depression have positive and negative nonlinearities, respectively. These nonlinearity values can be extracted from the experimental data. During weight update, the device's conductance is tuned within a confined conductance range, and only a finite number of conductance states are available due to the weight precision. Ideally, the lowest conductance state (OFF-state) should be low enough to represent the zero weight in the algorithm, making the dynamic range (conductance ON/OFF ratio) sufficiently large. In reality, the ON/OFF ratio is always finite and normally not large enough. On top of the nonlinear weight update curves, there are also considerable weight update variations from one pulse to another within one device. The effect of this cycle-to-cycle variation refers to as the variation in conductance change at every programming pulse.

To quantify the impact of the aforementioned nonideal device properties, we performed sensitivity analyses in online learning using our simulator. Figure 4a shows the requirement of weight precision. The result suggests that 6-bit weight is required for online learning (at least for MNIST data set). Figure 4b shows the learning accuracy with different ON/OFF ratios. Limited ON/OFF ratio (<10) will degrade the accuracy of online learning. Figure 4c shows the impact of weight update nonlinearity and asymmetry. The result shows that the asymmetry (positive potentiation P and negative depression D) is the key factor that degrades the accuracy, and high nonlinearity can be tolerated if P/D has the same polarity, which agrees with the results in [25]. However, for common situations where P/D is positive/negative, the impact of nonlinearity on the online learning accuracy is very critical. High accuracy can only be achieved with small nonlinearity (<1). The device's cycle-to-cycle variability is modeled as a random Gaussian-like fluctuation to the conductance after each programming pulse, and the amount of cycle-to-cycle variation (σ) is expressed in terms of percentage of the entire weight range. As shown in Figure 4d, small cycle-to-cycle variation (<2%) can alleviate the degradation of learning accuracy by high nonlinearity. The reason may be attributed to the random disturbance that aids convergence of the weights to an optimal weight pattern (i.e., to help the system jump out of local minima). Thus, synaptic devices with nonlinear weight update behavior may perform better than expected if they exhibit a little noisy weight update. However, too large variation (>2%) overwhelms the deterministic weight update amount defined by the algorithm and, thus, is harmful to the learning accuracy.

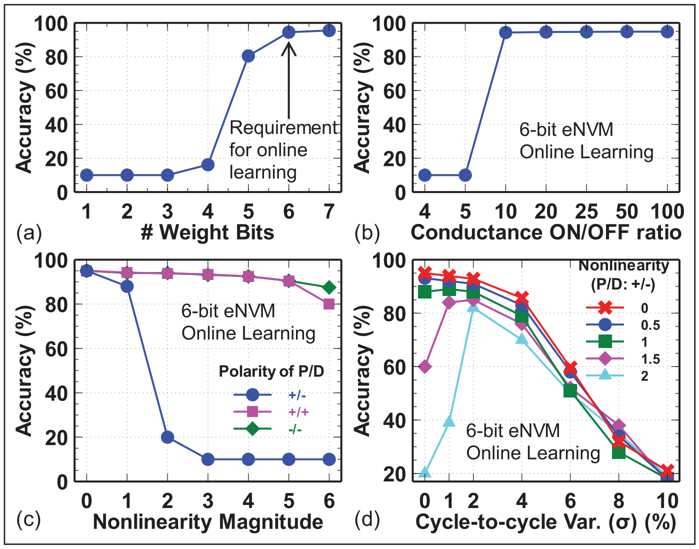

Figure 4. The impact of (a) weight precision, (b) conductance ON/OFF ratio, (c) weight update asymmetry/nonlinearity, and (d) weight update cycle-to-cycle variation in online learning.

Benchmark results and discussion

Table 1 surveys the representative analog synaptic devices in the literature with the extracted aforementioned device properties. Based on these parameters, NeuroSim+ was used to evaluate the system-level performance metrics such as learning accuracy, area, latency, energy, and leakage power for online learning with 1 million MNIST images being trained. The benchmark results show that all analog eNVM devices fail to achieve good accuracy (>90%). The cause of degradation can be largely attributed to the devices' poor ON/OFF ratio. It is observed that for an ON/OFF ratio <10, the devices cannot perform well in the learning no matter how good other parameters are. This agrees with the results

Table 1. Specs and online learning performance of different synaptic devices.

	Analog eNVM synapses					Analog FeFET synapses		Digital synapse
Device type	Ag:a-Si [3]	TaO$_x$/TiO$_2$ [4]	PCMO [5]	AlO$_x$/HfO$_2$ [6]	GST PCM [7]	HZO FeFET [13]	HZO FeFET [14]	6-bit SRAM
# of conductance states	97	102	50	40	100-120	32	32	--
Nonlinearity (weight increase/decrease)	2.4/-4.88	1.85/-1.79	3.68/-6.76	1.94/-0.61	0.105/2.4	2.53/1.83	1.545/1.755	--
R_{ON}	26 MΩ	5 MΩ	23 MΩ	16.9 kΩ	4.71 kΩ	559.28 kΩ	500 kΩ	--
ON/OFF ratio	12.5	2	6.84	4.43	19.8	45	~1300	--
Weight increase pulse	3.2V/300µs	3V/40ms	-2V/1ms	0.9V/100µs	0.7V (avg.)/ 6µs	3.65V (avg.)/ 75ns	2.17V (avg.)/ 50µs	--
Weight decrease pulse	-2.8V/300µs	-3V/10ms	2V/1ms	-1V/100µs	3V (avg.)/ 125ns	-2.95V (avg.)/ 75ns	-1.62V (avg.)/ 50µs	--
Cycle-to-cycle variation (σ)	3.5%	<1%	<1%	5%	1.5%	<1%	<1%	--
Online learning accuracy	~73%	~10%	10%	~41%	~87%	~90%	~90%	~94%
Area	1072.0 µm^2	1071.3 µm^2	1071.3 µm^2	3657.2 µm^2	7233.0 µm^2	1190.4 µm^2	1193.5 µm^2	10311 µm^2
Latency (naïve)	4.20E8 s	3.57E10 s	7.00E8 s	5.60E7 s	4.39E6 s	3.36E4 s	2.24E7 s	7.76 s
Energy (naïve)	87.94 mJ	65.86 mJ	29.4 mJ	150 mJ	1.52 J	98.01 mJ	38.39 mJ	6.98 mJ
Latency (optimized)	64200 s	2.845 s	5.2507 s	443.98 s	413.0 s	1.2924 s	479.6 s	3.7049 s
Energy (optimized)	14.81 mJ	0.17 mJ	0.17 mJ	146.19 mJ	1.34 J	0.21 mJ	0.28 mJ	3.3 mJ
Leakage power	35.29 µW	35.29 µW	35.29 µW	35.29 µW	35.29 µW	35.29 µW	35.29 µW	1.1 mW

in Figure 4b. The second critical parameter is the asymmetry/nonlinearity. Even the Pr0.7Ca0.3MnO3 (PCMO) device [5] has a slightly better ON/OFF ratio than the AlO$_x$/HfO$_2$ one [6]; its high nonlinearity restrains itself from converging to the desired conductance during weight update, leading to poor accuracy of 10%. In contrast, the learning accuracy of both FeFET devices is much better (~90%), owing to their large ON/OFF ratio. Even though their nonlinearities are not small, the degradation can be less significant because the potentiation and depression are symmetric, as observed in Figure 4c. It should be pointed out that PCM [7] and FeFET [13], [14] used nonidentical pulses with increasing pulse amplitude/widths to update the weights, which may complicate the peripheral circuit design.

Despite that SRAM can achieve better accuracy (~94%) than all analog synapses, it typically requires 10× area and >30× leakage power consumption. However, some analog synapses such as AlO$_x$/HfO$_2$ [6] and Ge2Sb2Te5 (GST) PCM [7] have less advantage in the area due to their small R_{ON}, where the transistor W/L in peripheral circuits (such as Mux and switch matrixes) needs to be sized larger to prevent the noticeable current-resistance drop. On the other hand, it is found in analog synapses that most of the latency and energy are dominated by the weight update, and they are far too large compared to those in SRAM, making analog synapses unfavorable for the online learning. This is because we have used a naive scheme for the weight update, where all cells in each operation need to go through the full number of pulse cycles (essentially the worst case) even if the cells do not have to be updated ($\Delta W = 0$). To optimize this scheme, in this article, we propose using the maximum ΔW's number of cycles in each weight update operation. If all the cells in an operation do not need an update ($\Delta W = 0$), this operation can be skipped. Table 1 shows the latency and energy with both the naive and optimized schemes. In the optimized scheme, the latency in analog synapses is significantly reduced, indicating ΔW are often small or zero. In TaO$_x$/TiO$_2$ [4] and PCMO [5] devices, the reduction ratios are extremely large because these devices basically could learn nothing (most $\Delta W = 0$). Similarly, the energy can also be greatly reduced in the optimized scheme with skipping operations, thus saving the dynamic energy in charging the array wires and circuits. The only exceptions are AlO$_x$/HfO$_2$ [6] and GST PCM [7]. Their energy reduction is much less because their R_{ON} is small, thus the array static energy (consumed by cells) dominates rather than the dynamic energy. If the programming pulse is short enough (<100 ns), analog synapses can be superior to SRAM in every aspect of the circuit-level performance with the optimized weight update scheme, as observed from the results of Hf0.5Zr0.5O2 FeFET [13].

We have developed an integrated framework, namely, NeuroSim+, which can evaluate the learning performance of neuroinspired architectures with different device technologies. We have analyzed the impact of nonideal device properties and benchmarked several representative analog synapses such as resistive memories, PCM, and FeFET in a two-layer MLP. The results suggest that degradation of learning accuracy is mainly due to small ON/OFF ratio and large asymmetry and nonlinearity in weight update. Finally, the optimized weight update scheme is proposed to minimize the latency and energy overhead by skipping redundant pulse cycles and even operations during training. ■

Acknowledgments

This work was supported in part by NSF-CCF-1552687, NSF-CCF-1740225, NSF/SRC E2CDA program, and in part by ASCENT, one of the six SRC/DARPA JUMP Centers. We would like to thank Prof. S. Datta from the University of Notre Dame for discussions on the ferroelectric field-effect transistor.

References

[1] Y.-H. Chen, T. Krishna, J. Emer, and V. Sze, “Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks,” in *Proc. IEEE Int. Solid-State Circ. Conf.*, 2016, pp. 262–263.

[2] N. P. Jouppi et al., “In-datacenter performance analysis of a tensor processing unit,” in *Proc. ACM/IEEE Int. Symp. Comput. Archit.*, 2017.

[3] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, “Nanoscale memristor device as synapse in neuromorphic systems,” *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, 2010.

[4] L. Gao et al., “Fully parallel write/read in resistive synaptic array for accelerating on-chip learning,” *Nanotechnology*, vol. 26, no. 45, pp. 455204, 2015.

[5] S. Park et al., “Neuromorphic speech systems using advanced ReRAM-based synapse,” in *Proc. IEEE Int. Elect. Dev. Meet.*, 2013, pp. 625–628.

[6] J. Woo et al., “Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems,” *IEEE Electr. Dev. Lett.*, vol. 37, no. 8, pp. 994–997, 2016.

[7] D. Kuzum, R. G. Jeyasingh, B. Lee, and H.-S. P. Wong, “Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing,” *Nano Lett.*, vol. 12, no. 5, pp. 2179–2186, 2011.

[8] M. Suri et al., “Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction,” in *Proc. IEEE Int. Elect. Dev. Meet.*, 2011, pp. 79–82.

[9] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. K. Likharev, and D. B. Strukov, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” *Nature*, vol. 521, no. 7550, pp. 61–64, 2015.

[10] 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. Electr. Dev.*, vol. 62, no. 11, pp. 3498–3507, 2015.

[11] P. Yao et al., “Face classification using electronic synapses,” *Nat. Commun.*, vol. 8, p. 15199, 2017.

[12] C. Li et al., “Analogue signal and image processing with large memristor crossbars,” *Nat. Electr.*, vol. 1, pp. 52–59, 2018.

[13] M. Jerry et al., “Ferroelectric FET analog synapse for acceleration of deep neural network training,” in *Proc. IEEE Int. Elect. Dev. Meet.*, 2017, pp. 139–142.

[14] S. Oh et al., “HfZrO x-based ferroelectric synapse device with 32 levels of conductance states for neuromorphic applications,” *IEEE Electr. Dev. Lett.*, vol. 38, no. 6, pp. 732–735, 2017.

[15] S. Yu, “Neuro-inspired computing with emerging non-volatile memory,” in *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[16] M. Hu, H. Li, Q. Wu, and G. S. Rose, “Hardware realization of BSB recall function using memristor crossbar arrays,” in *Proc. Design Autom. Conf.*, 2012, pp. 498–503.

[17] P.-Y. Chen, X. Peng, and S. Yu, “NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures,” in *Proc. IEEE Int. Elect. Dev. Meet.*, 2017, pp. 135–138. [Online]. Available: https://github.com/neurosim/MLP_NeuroSim_V2.0

[18] S. J. Wilton and N. P. Jouppi, “CACTI: An enhanced cache access and cycle time model,” *IEEE J. Solid-State Circ.*, vol. 31, no. 5, pp. 677–688, 1996.

[19] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, “NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory,” *IEEE Trans. Comput.-Aided Design Integr. Circ. Syst.*, vol. 31, no. 7, pp. 994–1007, 2012.

[20] S. Yu, P.-Y. Chen, Y. Cao, L. Xia, Y. Wang, and H. Wu, “Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect,” in *Proc. IEEE Int. Elect. Dev. Meet.*, 2015, pp. 451–454.

[21] L. Xia et al., "MNSIM: Simulation platform for memristor-based neuromorphic computing system," in *Proc. ACM/IEEE Design, Autom. Test Europe Conf. Exhi.*, 2016, pp. 469–474.

[22] D. Kadetotad et al., "Parallel architecture with resistive crosspoint array for dictionary learning acceleration," *IEEE J. Emerg. Select. Topics Circ. Syst.*, vol. 5, no. 2, pp. 194–204, 2015.

[23] Predictive Technology Model. [Online]. Available: http://ptm.asu.edu/

[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[25] S. Agarwal et al., "Resistive memory device requirements for a neural algorithm accelerator," in *Proc. Int. Joint Conf. Neu. Nets.*, 2016, pp. 929–938.

Pai-Yu Chen is a Software Engineer at Synopsys, Sunnyvale, CA, USA. Chen has a BS in electrical engineering from National Taiwan University, Taipei, Taiwan, an MSE in electrical engineering from the University of Texas, Austin, TX, USA, and a PhD degree in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2018.

Shimeng Yu is an Associate Professor of electrical and computer engineering at Georgia Institute of Technology, Atlanta, GA, USA. Yu has a BS degree in microelectronics from Peking University, Beijing, China, and an MS and PhD in electrical engineering from Stanford University, Stanford, CA, USA.

■ Direct questions and comments about this article to Shimeng Yu, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; shimeng.yu@ece.gatech.edu.