Exploiting Hybrid Precision for Training and Inference: A 2T-1FeFET Based Analog Synaptic Weight Cell

Xiaoyu Sun[1], Panni Wang[1], Kai Ni[2], Suman Datta[2], and Shimeng Yu[3]
[1]Arizona State University, Tempe, AZ 85281, USA [2]University of Notre Dame, Notre Dame, IN 46556, USA
[3]Georgia Institute of Technology, Atlanta, GA 30332, USA Email: shimeng.yu@ece.gatech.edu

***Abstract*—** In-memory computing with analog non-volatile memories (NVMs) can accelerate both the in-situ training and inference of deep neural networks (DNNs) by parallelizing multiply-and-accumulate (MAC) operations in the analog domain. However, the in-situ training accuracy suffers from unacceptable degradation due to undesired weight-update asymmetry/nonlinearity and limited bit precision. In this work, we overcome this challenge by introducing a compact Ferroelectric FET (FeFET) based synaptic cell that exploits hybrid precision for in-situ training and inference. We propose a novel hybrid approach where we use modulated "volatile" gate voltage of FeFET to represent the least significant bits (LSBs) for symmetric/linear update during training only, and use "non-volatile" polarization states of FeFET to hold the information of most significant bits (MSBs) for inference. This design is demonstrated by the experimentally validated FeFET SPICE model and co-simulation with the TensorFlow framework. The results show that with the proposed 6-bit and 7-bit synapse design, the in-situ training accuracy can achieve ~97.3% on MNIST dataset and ~87% on CIFAR-10 dataset, respectively, approaching the ideal software based training.

I. INTRODUCTION

DNNs have made remarkable advances in cognitive tasks such as image and speech recognition. However, the energy-efficiency and speed of DNN training is highly limited by moving the data back and forth between the memory and the processor in conventional von Neumann hardware. To overcome this challenge, in-memory computing, where computing is done at the location of the data storage, has been proposed to accelerate the computation. To store a large number of DNN weights on-chip, logic process compatible NVM devices are attractive where synaptic weights are encoded as their analog conductance values. There are array-level experimental demonstrations for training/inference with resistive random-access memory (RRAM) [1-2] and phase change memory (PCM) [3]. However, training with these NVMs suffers from unacceptable accuracy degradation due to various non-idealities including limited dynamic range, variation, and most importantly asymmetric/nonlinear weight update [4]. For example (Fig. 1), in filamentary RRAM [5], the excessive asymmetry/nonlinearity between positive and negative update leads to a poor accuracy ~41% for MNIST dataset. While interfacial RRAM [6] exhibits improved nonlinearity with higher accuracy ~73%, the programming pulse width is on the orders of ms due to the slow diffusion process of ions or vacancies. A recent discovery of partial polarization switching in ferroelectric-FET (FeFET) [7] provides highly symmetric weight update leading to an accuracy ~90%, but "non-identical" pulses must be applied for conductance tuning, which increases the peripheral circuitry complexity. Despite recent progress, these hardware implementations are not competitive with the software training accuracy ~98% even for MNIST dataset.

Motivated by the observation that in a DNN algorithm a relatively higher precision (larger than 6-bit) is necessary during training to accumulate the incremental weight change, while a lower precision (less than 2-bit) is sufficient during inference to achieve a reasonably good accuracy [8], we introduce a synaptic weight cell design in this work that combines two CMOS transistors and one FeFET (2T1F) for training and inference with hybrid precision. During training, the "volatile" modulated gate voltage of FeFET is used to represent LSBs for symmetric and linear update. After training process is complete, the information of LSBs is discarded, only MSBs are preserved by "non-volatile" polarization states of FeFET for inference. We demonstrate a 6-bit/7-bit synapse design (2-bit MSBs + 4-bit/5-bit LSBs) for MNIST/CIFAR-10 dataset and benchmark with a LeNet-5-like/VGG-like convolutional neural network (CNN). The SPICE simulation result with the experimental validated FeFET model and TSMC 65nm PDK is coupled with the TensorFlow framework, showing that the learning accuracy could achieve ~97.3%/~87%, approaching the ideal software training.

II. 2T1F SYNAPTIC WEIGHT CELL DESIGN

A. 2T1F Analog Synaptic Weight Cell

Fig. 2-3 show the schematic and fundamental principle of the proposed 2T1F synaptic weight cell. The FeFET gate capacitor serves as an analog memory for LSBs, which is charged/discharged by the corresponding pull-up pFET and pull-down nFET. Thus, the LSBs of the weight can be encoded to the channel conductance of the FeFET by modulating the gate voltage (V_G) while keeping the FeFET working in the triode region as shown in Fig. 3(a). During weight update, pulses are applied to the gate of pFET/nFET for positive/negative updates while keeping these two transistors working in saturation region to ensure the charging/discharging current is independent of V_G. With the pulse amplitudes that generate the balanced charging and discharging current, the positive/negative update of LSBs is expected to be symmetric. The MSBs can be encoded to different FeFET polarization (thus channel conductance) states without overlapping LSBs within each MSB state. For example, assuming 2-bit MSBs (i.e., 4 polarization states) as shown in Fig. 3(b), the V_G dynamic range [V_A, V_B] which is constrained by the linear region overlap of multiple polarization states, determines the number of update steps (i.e., the bitwidth of LSBs). Fig. 3(c) illustrates different

scenarios of updating LSBs and MSBs. If V_G increases beyond V_B, the consequential read-out current I_D will be larger than the reference current (ref. 2 in Fig. 3), requiring a programming towards S_2 state to transfer the weight information to MSBs, then the LSBs can be continuously updated within S_2 state and V_G prefers to be reset to the certain level that maintains the same I_D to prevent the information loss of LSBs. Similarly, if I_D decreases below ref. 2, the FeFET requires a programming towards S_1 state.

With the proposed synaptic weight cell, the modified DNN training flow is shown in Fig. 4. For each training batch, update LSBs by applying certain pulses to modulate V_G based on the value of ΔW calculated through stochastic gradient descent (SGD) based backpropagation algorithm [8]. Due to the limited V_G dynamic range and capacitor leakage, the information of LSBs needs to be occasionally transferred to MSBs to prevent the information loss. As a result, for every N batches, we need to transfer the weight, i.e., program the FeFET to the corresponding state according to the read-out current level. After the weight-transfer, V_G prefers to be reset to the certain level that maintains the same channel conductance to recover the residual information of LSBs. However, this step requires a high-precision ADC (equals the total bitwidth of weights) which is too power- and area-hungry in practice. Therefore, we only reset the V_G to $(V_A+V_B)/2$ to avoid high-precision ADCs at the expense of inducing possible residual errors. The impact of these errors on learning accuracy is investigated in Section III.

B. Implementation of 2-bit MSBs + 4-bit LSBs Synapse

First, we demonstrate the implementation of 6-bit synapse (2-bit MSBs + 4-bit LSBs) as an example. The FeFET utilizes multi-domain polarization switching dynamics in ferroelectric $Hf_{0.5}Zr_{0.5}O_2$ (HZO) gate dielectric to gradually tune the threshold voltage of the underlying channel by the application of programming pulses to the gate. Fig. 5 shows the measured I_D-V_G characteristics of our fabricated HZO FeFET with tunable V_{th}. We adopt the FeFET SPICE model from our prior work [9], where the model consists of a conventional MOSFET model by BSIM 4, and a ferroelectric switching by Preisach dynamic model. The SPICE model accurately captures the experimental P-V loop (Fig. 6). Fig. 7(a) shows the pulse scheme and the simulated corresponding remnant polarization charge that result in 4 states shown in Fig. 7(b), which serve as 2-bit MSBs. The dynamic range of V_G is set to be [1.44V, 1.76V], with a pulse width of 5ns that leads to ΔV_G of 20mV per update pulse, 4-bit LSBs can be achieved. The voltage bias schemes for update/read operations of the 2T1F weight cell are summarized in Fig. 8. The equivalent weight update curve of the 6-bit synapse is shown in Fig. 9. However, because the charging and discharging current cannot always be the same in practical circuits, ΔV_G per update pulse is not ideally the same at different V_G, resulting a slight nonlinearity as shown in Fig. 10, but the weight update is still symmetric as the maximum difference of ΔV_G between positive update and negative update is only ~5% of one LSB step. The nonlinearity is observed to be less than +1/+1 as defined in Fig. 11, which is much better than those asymmetric and nonlinear NVM devices [4] as compared in Fig. 12.

III. Results and Discussion

We benchmark the performance of the proposed hybrid 6-bit 2T1F synapse by incorporating the aforementioned synaptic characteristics into TensorFlow simulation with a CNN, which is a variation of LeNet-5 (Fig. 13), for MNIST dataset. The learning accuracy of ~98.5% from ideal software training with 6-bit weights is utilized as the baseline. With the slight nonlinearity in 2T1F design, the accuracy can achieve ~98.3% (Fig. 14). Then we investigate the impact of residual errors caused by occasional weight-transfer on the accuracy. Fig. 15 shows the simulation results of V_G leakage with different starting V_G, the inset figure shows that it takes 1.64ms for V_G to leak by one LSB step (20mV) in the worst case (starting V_G = 2V). Assuming the training time per batch (forward + backward + update, batch size is 100) is ~7 µs, the maximum transfer interval becomes ~230 batches. Fig. 16 shows the training accuracy curve with transfer interval of 100, 200, and 300 batches. When the transfer interval is 100 batches, the accuracy can only achieve ~96%, when the transfer interval is 200 batches and 300 batches, the accuracy can reach ~97.3% and ~98.0% respectively, showing slight degradations compared to 98.3%. The reason is that if the absolute accumulated ΔW within one transfer interval is less than half of one MSB step (8 LSB steps), which fails to trigger the MSBs state change, the weight will be reset back after weight transfer as the V_G will be reset to $(V_A+V_B)/2$ as aforementioned. Fig. 17 shows the percentage of effective $|\Delta W|$ (>8 LSB steps) during first, second, and third weight-transfer operations as an example. A larger interval leads to a larger percentage of effective $|\Delta W|$. Given the fact that weights tend to be stabilized through training process, a dynamic transfer interval (increasing through training) is preferred to fully recover the accuracy. Fig. 18 shows the impact of FeFET polarization state variation on the learning accuracy. A small variation (<2.5%) does not hurt the accuracy as it may help on compensating the residual errors caused by non-ideal weight-transfer. The degradation becomes unacceptable when variation exceeds 5%. By directly reducing the LSBs tuning step to 10 mV, which results in a 7-bit synapse (2-bit MSBs + 5-bit LSBs), we estimated the learning accuracy on the more complex CIFAR-10 dataset with a VGG-like CNN. Fig 19 shows that the accuracy can achieve ~87% without noise, and ~88% with noise in one LSB step due to the random fluctuation of a 10mV step in practice.

Fig. 20 compares this work to recent works with "volatile" capacitor-based design. The work [10] using 1T1C is totally volatile thus could not support inference. While the work [11], which combines 2 PCM cells with a 3-transistor-1-capacitor structure, is suitable for both training and inference, it has relatively larger cell size and higher programming energy.

IV. Conclusion

We introduce a compact 2T1F synaptic weight cell design that combines the benefits of capacitor-based symmetric weight update for LSBs during training and NVM based long-term weight storage for MSBs during inference. A 6-bit/7-bit synapse is demonstrated for MNIST/CIFAR-10 dataset, which can achieve accuracy of ~97.3%/~88%, approaching that of the ideal software based training.

ACKNOWLEDGMENT: This work is supported by ASCENT, one of the six SRC/DARPA JUMP centers.

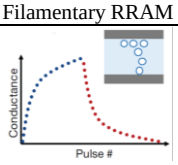
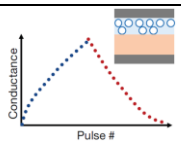
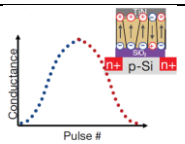
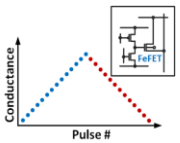
Analog Synapse Devices				
Type	Filamentary RRAM	Interfacial RRAM	Ferroelectric FET	This work
Weight update behavior	 [5]	 [6]	 [7]	
Symmetry	Low	Medium	Medium	High
Programming	Identical Pulses	Identical Pulses	Non-identical Pulses	Identical Pulses
Speed	10-100ns	100µs-10ms	50ns-100ns	5ns
MNIST Accuracy	~41%	~73%	~90%	~97.3%

Fig. 1. Comparison of analog synapses for on-chip in-situ learning. The proposed 2T1F design exhibits the desired characteristics including highly symmetric/linear weight update, fast and identical update pulses, allowing fast training of neural networks with high accuracy.

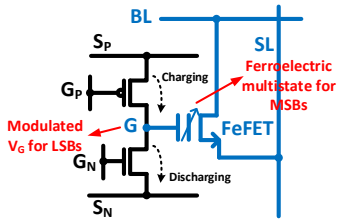


Fig. 2. Schematic of the proposed 2T1F weight cell design. The LSBs of weight are encoded to the conductance of the FeFET by modulating V_G while the MSBs are encoded to different FeFET polarization states.

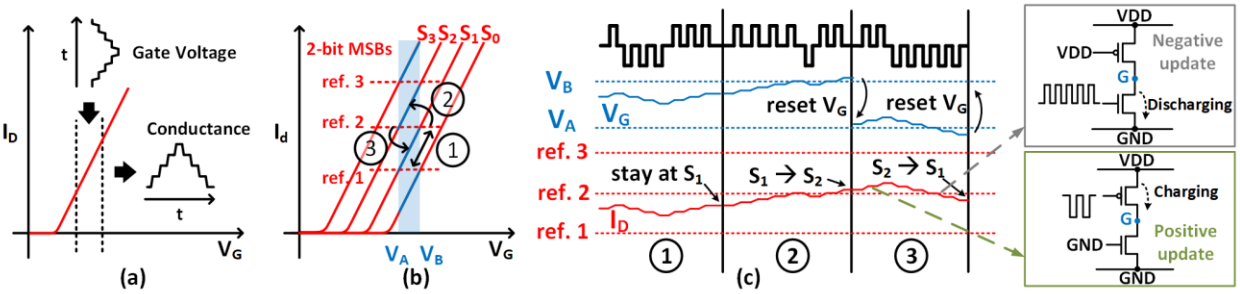


Fig. 3 (a) The LSBs of weight are linearly encoded to the conductance value of the FeFET by modulating the gate voltage while keeping the FeFET in the triode region. (b) The MSBs are encoded to different FeFET polarization states without overlapping of LSBs within each MSB state. (c) Illustration of updating LSBs within a FeFET polarization state and updating MSBs depending on the corresponding read-out current level. V_G is preferred to be reset to the certain level to maintain the same I_D after the update of MSBs to prevent the information loss of LSBs.

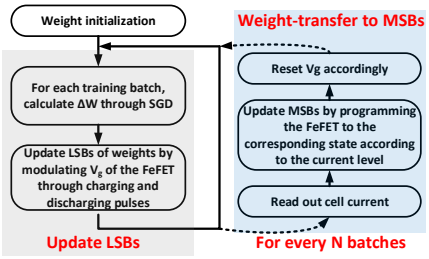


Fig. 4. The training flow chart. For each training batch, update LSBs by applying charging/discharging pulses to modulate V_G based on the value of ΔW. For every N batches, program the FeFET to the corresponding polarization state according to the read-out current level, namely weight-transfer.

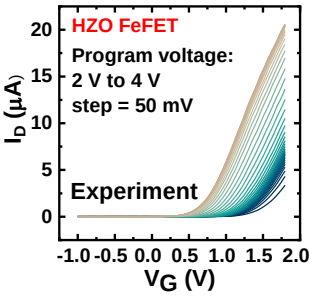


Fig. 5. Measured I_D-V_G characteristics of our fabricated HZO FeFET [9] for program voltage from 2V to 4V, showing tunable threshold voltage.

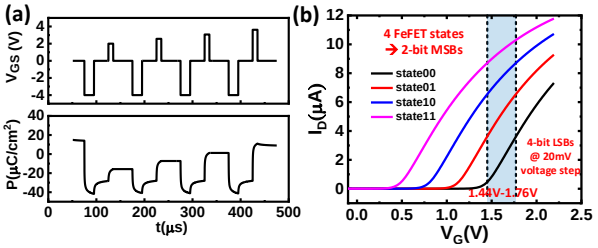


Fig. 7. (a) The pulse scheme and the corresponding remnant polarization charge that generates 4 FeFET states. (b) Simulated I_D vs. V_G curve of different FeFET states. 4 polarization states serve as 2-bit MSBs. The dynamic range of V_G is set to be [1.44V, 1.76V], with a pulse width of 5ns that leads to ΔV_G of 20mV per update pulse, 4-bit LSBs can be achieved.

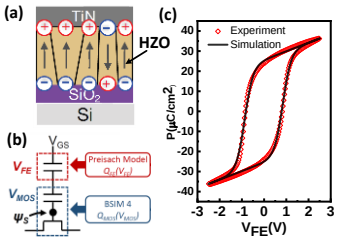


Fig. 6. (a) Schematic of partially switching of HZO ferroelectric domains. (b) The FeFET model [9] consists of the conventional MOSFET modeled by BSIM 4 and the ferroelectric modeled by the dynamic Preisach model. (c) The model accurately captures the experimental P-V loop.

Operation	Update LSBs	Update MSBs	Read
BL	GND	GND/4V	V_{IN}
SL	GND	GND/4V	GND
S_P	VDD	2V-4V/VDD	1.6V
G_P	Pulse/VDD	GND/VDD	GND
S_N	GND	GND/GND	GND
G_N	GND/Pulse	GND/VDD	GND
Substrate	GND	GND/4V	GND

Fig. 8. The voltage bias scheme for different operations for 2T1F weight cell. For the update of MSBs, V1/V2 means the voltage for program/erase operation respectively. For the update of LSBs, V1/V2 means the voltage for charging and discharging, respectively.

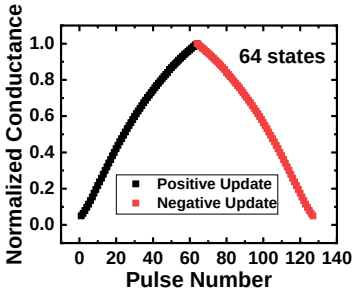


Fig. 9. The equivalent conductance update curve of the 6-bit synapse realized by 2T1F weight cell, showing much improved symmetry and linearity between positive update and negative update.

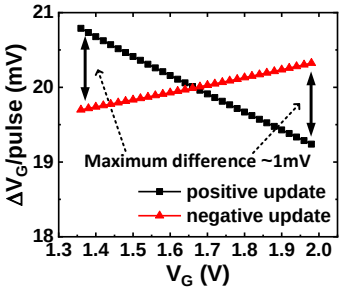


Fig. 10. The ΔV_G per pulse during positive update and negative update as a function of V_G. The maximum difference of ΔV_G between two directions is only ~1mV (5% of one LSB step), suggesting symmetry in weight update.

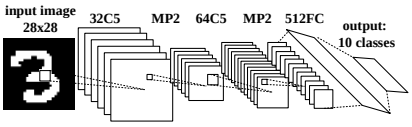


Fig. 13. Benchmark with a CNN on MNIST dataset. The adopted CNN is a variation of LeNet-5 with 32C5-MP2-64C5-MP2-512FC-10 configuration.

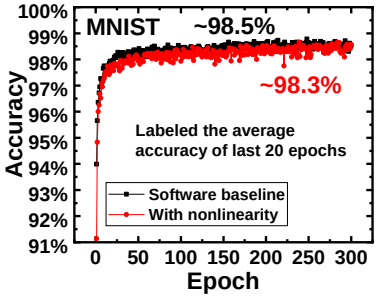


Fig. 14. The MNIST learning accuracy can achieve 98.3% with the slight nonlinearity of the proposed 2T1F design, showing 0.2% degradation compared to the ideal software training with 6-bit weights.

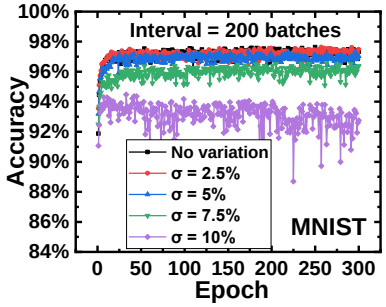


Fig. 18. The impact of FeFET polarization state variation on the MNIST learning accuracy. A small variation does not hurt the accuracy as it may help on compensating the residual errors caused by non-ideal weight-transfer. The degradation becomes unacceptable when variation exceeds 5%.

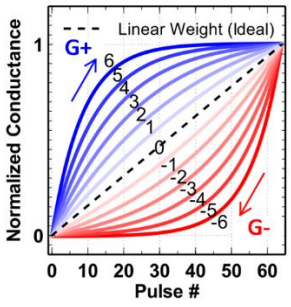


Fig. 11. Analog NVM device behavioral model [4] of the nonlinear/asymmetric weight update. The nonlinearity degree is labeled from +6 to -6.

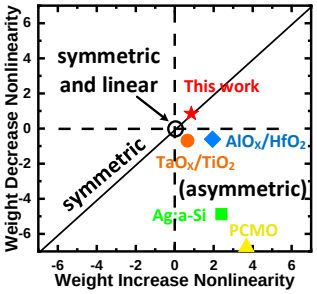


Fig. 12. Comparison of the asymmetry/linearity between this work and other NVM devices [4]. The nonlinearity of this work is fitted using the same model in Fig. 11 as used by other works.

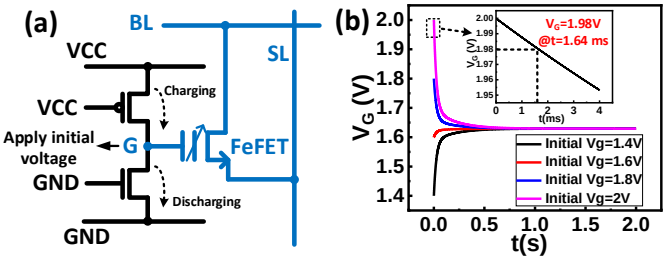


Fig. 15. (a) Circuit setup for leakage simulation. (b) Simulation results of V_G leakage with different starting V_G. The inset figure shows that it takes 1.64 ms for V_G to leak by one LSB step (20mV) in the worst case (starting V_G = 2V). Assuming the training time is ~7 µs/batch, the maximum transfer interval becomes ~230 batches, limited by the leakage.

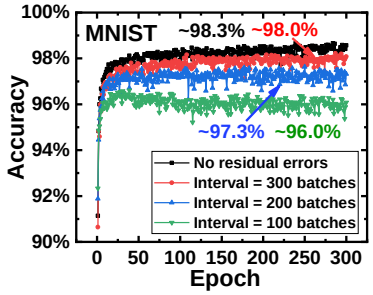


Fig. 16. The MNIST learning accuracy with weight- transfer interval of 100, 200, and 300 batches, achieving ~96.0%, ~97.3%, and ~98.0% respectively.

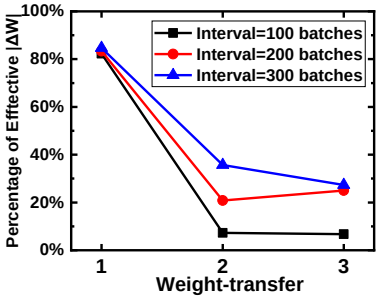


Fig. 17. The percentage of effective |ΔW| (>8 LSB steps) during first, second, and third weight-transfer with different number of interval batches. A larger interval leads to a larger percentage of effective |ΔW| to be accumulated, which benefits the training.

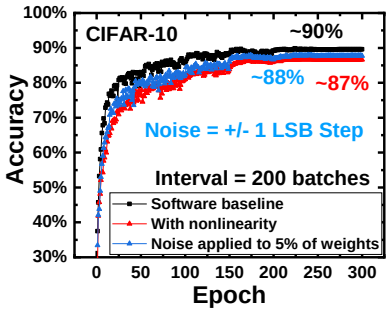


Fig. 19. The learning accuracy on CIFAR-10 dataset could achieve 87% w/o noise and 88% w/ noise using the proposed 7-bit synapse with a VGG-like CNN.

Work	[10]	[11]	This work
Weight Cell	3T1C	2PCM+ 3T1C	2T1F
Programming energy	Low	High	Medium
Area	Medium	High	Low
Training	✓	✓	✓
Inference	×	✓	✓

Fig. 20. Comparison between this work and recent works with the capacitor-based design. This work supports both training and inference with relatively lower programming energy and area.

References

[1] C.-C. Chang, et al., *IEDM*, 2017. [2] P. Yao, et al., *Nature Communications*, 2017. [3] G. W. Burr, et al., *IEDM*, 2014. [4] S. Yu, *Proc. IEEE*, 2018. [5] J. Woo, et al., *IEEE Electron Device Lett.*, 2016. [6] S. H. Jo, et al., *Nano Letters*, 2010. [7] M. Jerry, et al., *IEDM*, 2017. [8] S. Wu, et al., *ICLR*, 2018. [9] K. Ni, et al., *Symp. VLSI Tech.*, 2018. [10] Y. Li, et al., *Symp. VLSI Tech., 2018.* [11] S. Ambrogio, et al., *Nature*, 2018.