First Demonstration of Ge Ferroelectric Nanowire FET as Synaptic Device for Online Learning in Neural Network with High Number of Conductance State and G_{max}/G_{min}

Wonil Chung, Mengwei Si, and Peide D. Ye*
School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA
*Tel: 1-765-494-7611, Fax: 1-765-496-6443, email: yep@purdue.edu

***Abstract*—** In this paper, optimum weight update scheme for improved linearity and asymmetry of channel conductance potentiation and depression in a Germanium ferroelectric (FE) nanowire FET (NWFET) was experimentally demonstrated and simulated for the first time. It was found that -5 V, 320 pulses and +5 V, 256 pulses both with 50 ns pulse width were the optimum pulsing conditions for potentiation and depression process, respectively. With the optimized scheme, non-linearity for potentiation and depression were extracted to be α_p = 1.22 and α_d = -1.75, respectively resulting in asymmetry ($|\alpha_p - \alpha_d|$) of 2.97 based on models embedded in MLP simulator and NeuroSim [1]. G_{max}/G_{min} ratio (few hundreds) and number of conductance states (> 256) are both very large. 9 alternating consecutive conductance updates (potentiation followed by depression) were executed to observe variability in conductance profiles. Multilayer perceptron neural network was simulated over 1 million MNIST images with extracted experimental parameters which yielded in online learning accuracy of ~ 88 %.

I. Introduction

Due to introduction of processors with higher performance and faster parallel computing capabilities, brain-inspired synaptic device networks caught much attention for various real-life applications. Specifically, e-NVM (emerging non-volatile memory) such as resistive [2], [3], phase change [4] or ferroelectric [5]–[7] devices are studied towards non-von Neumann architectures. Specifically, ferroelectric (FE) devices, mostly with $Hf_{0.5}Zr_{0.5}O_2$ (HZO) due to its high compatibility with CMOS platform, are being studied actively related with negative capacitance (NC) devices [8], [9]. Studies on FE switching speed follow this trend [10], [11]. Partial polarization of a FE device makes it possible to serve as a synaptic device when optimized properly [5]. Motivation for using e-NVM for online learning in deep neural network (DNN) is related to their ability to retain the weight data and to locally process the input via multiplication of weights in the form of conductance. Subsequent addition of weighted currents are read with peripheral circuitry [1] and the processed data is used as the input of the following layer after non-linearization. Sequential processing of data from one layer to the following layer can be noted as forward propagation as shown in Fig. 1. During online training process, back propagation is used to update the less-accurate synaptic devices' weights. Iterative cycles of forward and back propagation increase the accuracy as it undergoes training epochs. For efficient and adaptive operation of the network, linear and symmetrical conductance profile is highly preferred. Fig. 2 depicts various possible non-ideal effects related to e-NVM's programming process including nonlinearity, asymmetry, stochasticity and large variability in conductance values [4]. Although it was reported that the e-NVM-based neural networks are relatively robust to conductance's stochasticity and variability [4], nonlinear and asymmetric profiles significantly limit the accuracy. To overcome such detrimental influence, different pulsing schemes were proposed as depicted in Fig. 3 [5], [7]. However, if pulses are not *identical* throughout the programming process, an additional step of accessing the weight value is needed every time an update takes place to find the appropriate pulse at that specific level compromising the efficiency. In this paper, we demonstrate Germanium FE NWFET as a synaptic device with high number of conductance states and G_{max}/G_{min}. *Identical* conductance update pulsing schemes were optimized for improved linearity and asymmetry. Ge FE NWFETs can then be configured into pseudo-crossbar array as shown in Fig. 4 for practical implementation towards the DNN.

II. Experiments

Germanium-on-insulator (GeOI) wafer with 100 nm of Ge layer on top of SiO_2 (400 nm) and Si handling wafer was used for the fabrication of the Ge FE NWFETs. Typical wafer cleaning and mesa isolation was processed, followed by ion implantation. Fin structure was first defined with SF_6-based dry etching and nanowire was released by partial etching of the underlying SiO_2. 3-step ALD deposition was conducted starting with 1 nm of Al_2O_3 and subsequent post-oxidation to form thin GeO_x (~1 nm) under the Al_2O_3 for better interface quality. 10 nm of HZO and additional 1 nm of Al_2O_3 capping layer were deposited. Then 500 °C post deposition annealing (PDA) was performed to enhance the ferroelectricity within the HZO stack. Source and drain were formed using recessed S/D technique [12]. Ohmic annealing and gate, source and drain metallization was finished in the last step. Fig. 5 presents more details related to the fabrication steps. Fig. 5 (a) and (b) visualize the 3D structure of the device and false-colored SEM images are shown in Fig. 6 (a)-(c). Multiple parallel nanowires form a single device and a typical transfer curve of such device can be seen in Fig. 7. Large negative hysteresis (-4 V) is observed showing ferroelectric switching. Fatigue measurement on our ALD HZO was done with 10^9 PUND (Positive Up, Negative Down) pulses to verify our HZO's reliability (Fig. 8). Details on our HZO film and devices' operation are elaborated in our previous reports [8], [13], [14].

III. Results and Discussion

As seen in Fig. 4, FeFETs can be configured into a pseudo-crossbar design. When a row is selected, and subsequent potentiation and depression pulses are fed into the gate of a FeFET, already-saved conductance value can be newly programmed through partial polarization switching (Fig. 9). Input data fed (Fig. 4, purple line) into the FeFET's channel results in weighted current and can be read externally. To optimize the programming pulses, measurement set-ups were prepared as configured in Fig. 10. Fig. 11 is the real-time monitored potentiation process after applying a single -8.75 V pulse (75 ns) to the gate of FeFET. Current was measured using an oscilloscope through current amplifier (Fig. 10 (a)).

Then, pulse voltage optimization was done with fixed pulse widths. As seen in Fig. 12, with a fixed pulse width (1 μs), increasing the potentiation voltage from -1 V to -4.5 V reduces the number of pulses needed to reach the maximum conductance (G_{max}, ~ 80 μS). Fixing the pulse level at -4.5 V, pulse widths were swept from 100 ns to 1 μs and similar trend in conductance profile could be observed (Fig. 13) where longer pulses reduce the number of pulses to maximize the conductance. Same procedure was executed for depression pulses (Fig. 14). It could be clearly seen from Fig. 15 that if the pulse conditions (either voltage or pulse width) are not properly optimized, it gives highly non-linear and asymmetric conductance profiles. Non-linearity for potentiation and depression were extracted to be α_p = 5.72 and α_d = -9, respectively yielding asymmetry ($|\alpha_p - \alpha_d|$) of 14.72. These values were extracted through curve-fitting model embedded in MLP (Multilayer perceptron) simulator (+NeuroSim) [1] for fair benchmarking among reported HZO-based FeFET synapse devices.

Freezing of a device (stuck-on conductance) were occasionally observed when excessively strong pulses (significantly larger number of pulses, long pulse widths, or larger pulse voltages) were delivered similar to the reported PCRAM-based study [4] limiting subsequent conductance programming. The report concludes that the contribution of these non-ideal circumstances doesn't compromise the accuracy too much. Restoration was possible via initialization (#1 in Fig. 9) but using intermediate conductance values in the first place rather than exploiting the whole maximum conductance range would be helpful in preventing the devices from becoming nonresponsive in the cost of reduced G_{max} and number of states.

Considering various factors, pulse width was reduced to 50 ns as seen in Fig. 16 for better $\alpha_{p,d}$ and lower asymmetry. Pulse period was 500 μs and conductance sampling was done 50 μs after each pulse. Conductance values were sampled at single point (V_G = 0 V) instead of sweeping V_G to minimize the unwanted effect of measurement V_G on the programmed polarization state. It was found that when our Ge FE pNWFET (L= 105 nm, W= 32 nm, H= 26 nm) was subjected to 50 ns, V_G = $\mp$5 V (- for potentiation, + for depression), it resulted in significantly improved linearity (Fig. 17) of α_p= 1.22 and α_d= -1.75 (Asymmetry = $|\alpha_p - \alpha_d|$ = 2.97). Respective pulse numbers were chosen to acquire symmetric operation between potentiation (320 pulses) and depression (256 pulses) resulting in effective control of the conductance without introducing non-responsive devices. Smaller pulse number for depression implies that the polarization switching is more sensitive to depression (+5 V) than potentiation (-5 V). To investigate the variability of the optimized pulse scheme, 9 consecutive cycles of continuously alternating potentiation (-5 V, 50 ns, 320 cycles) and depression (+5 V, 50 ns, 256 cycles) were executed. Fig. 18 (a) is the accumulated conductance profiles and Fig. 18 (b) shows the overlapped profiles. It could be observed that this pulse scheme was effective repetitively yielding reliable conductance profiles without serious non-ideal curves. Since number of conductance states were high in our case (320 for potentiation and 256 for depression), multiple pulses could be tied in the form of a pulse train to reduce the number of states for applications that require lower number of states. If 10 pulses (-5 V, 50 ns) are delivered as a pulse train, it will result in 320/10 = 32 states (5 bits) with 10 times larger ΔG step. The G_{max}/G_{min} ratio is also considered as an important parameter in training simulation. Low ratio causes degradation in training accuracy which makes higher ratio more preferable [1], [3]. Our devices show excellent ratio in the range of hundreds because of low G_{min} < 1 μS and high G_{max} ~ 200 μS (Fig. 18).

Fig. 19 shows the improvement in non-linearity due to transition from non-optimized (Fig. 15) to optimized (Fig. 18) scheme. Both α_p and α_d approach the desired targeted values of +1 and -1 (ideally both 0) [1]. Fig. 20 compares the asymmetry and training accuracy of online learning through multilayer perceptron (MLP) neural network architecture (400 input, 100 hidden, 10 output neurons) [1] before and after conductance profile optimization. 1 million hand written digit images (MNIST, Modified National Institute of Standards and Technology, cropped 20 × 20 pixels) were trained for 125 epochs of training. With dramatic improvement in linearity (and thus asymmetry), high accuracy of ~ 88 % could be achieved. It could be further improved if various parameters in the simulator such as learning rates between layers, number of synapses, number of hidden layers are optimized more precisely. Fig. 21 summarizes device performance metrics of reported HZO-based FeFET synaptic devices. All three studies use HZO as main FE dielectric layer and results using identical pulses are shown for comparison.

IV. Conclusion

In this paper, we have reported the first experimental and simulation demonstration of Ge FE NWFET as a synaptic device for online learning in a neural network with optimized pulse schemes. Separate *identical* pulsing conditions for potentiation (-5V, 50 ns, 320 cycle) and depression (+5V, 50 ns, 256 cycle) were found respectively which gave significantly improved linearity and symmetry in conductance profiles. These conditions were effective in preventing devices from becoming non-responsive since the combination of pulse voltage, pulse width and number of pulses was optimized to prevent the freezing. As a result, improved linearity (α_p/α_d= 5.72/-9 → 1.22/-1.75) and asymmetry (14.72 → 2.97) in conductance profiles could be observed. Learning accuracy after training 1 million MNIST images over 125 epochs gave ~ 88 %. It can be concluded that precise optimization of pulsing conditions can affect the conductance update profiles significantly.

Acknowledgment

The authors appreciate Pragya R. Shrestha, Jason P. Campbell and Kin P. Cheung at National Institute of Standards and Technology (NIST) for demonstration of real-time monitoring of conductance during potentiation pulses. The work is supported by SRC and Lam Research.

References

[1] P. Y. Chen et al., *IEDM*, 2017. [2] J. Woo et al., *EDL*, vol. 37, no. 8, pp. 994–997, 2016. [3] S. Yu et al., *IEDM* ,2015. [4] G. W. Burr et al., *IEDM*, 2014. [5] M. Jerry et al., *IEDM*, 2017. [6] M. Seo et al., *EDL*, vol. 39, no. 9, pp. 1445-1448, 2018. [7] S. Oh et al., *EDL*, vol. 38, no. 6, pp. 732–735, 2017. [8] W. Chung et al., *IEDM*, 2017. [9] M. Si et al., *Nat. Nanotechnol.*, vol. 13, no. 1, pp. 24–28, Jan. 2018. [10] Z. Krivokapic et al., *IEDM* ,2017. [11] J. Muller et al., *EDL*, vol. 33, no. 2, pp. 185–187, 2012. [12] H. Wu et al., *IEDM*, 2014. [13] W. Chung et al., *VLSI*, 2018. [14] M. Si et al., *IEDM*, 2017.

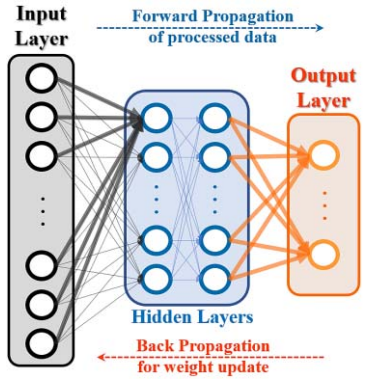


Fig. 1. Basic operation diagram of online training scheme with back propagation for weight update.

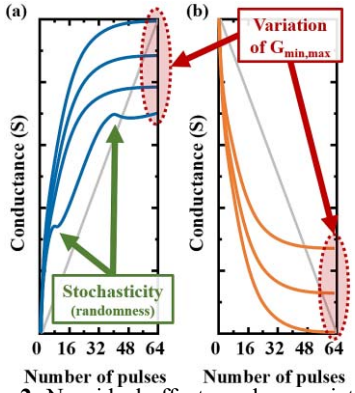


Fig. 2. Non-ideal effects such as variation in $G_{max,min}$, stochasticity, non-linear and asymmetric G profile are shown.

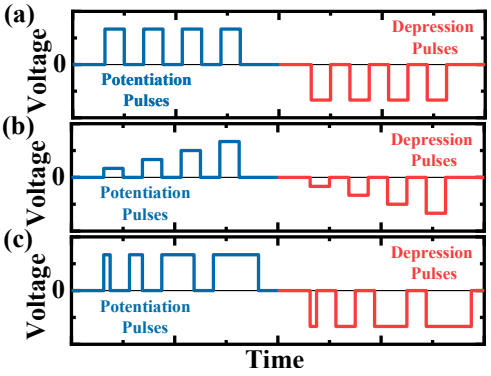


Fig. 3. Various possible pulses for potentiation and depression. **(a)** Identical pulses, **(b)** different pulse levels and **(c)** pulse widths.

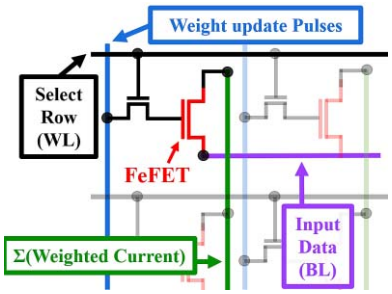


Fig. 4. Pseudo-crossbar scheme showing weight update (blue) and data processing (purple → green) in a row selected by WL.

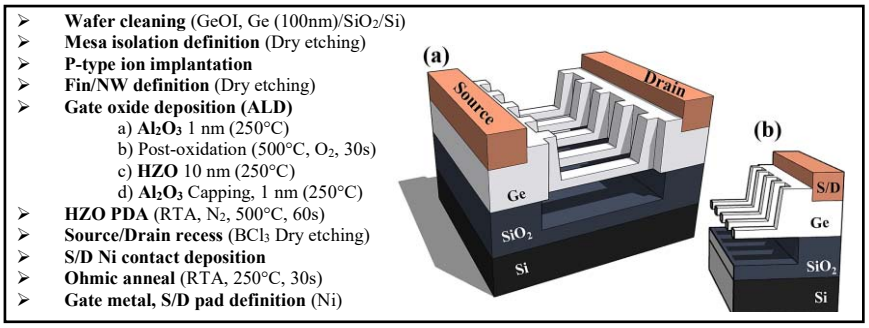


Fig. 5. Key process steps for fabrication of a Ge FE NWFET neuromorphic device. **(a)** 3D Structure of the device and **(b)** its cross-sectional view of nanowires are illustrated.

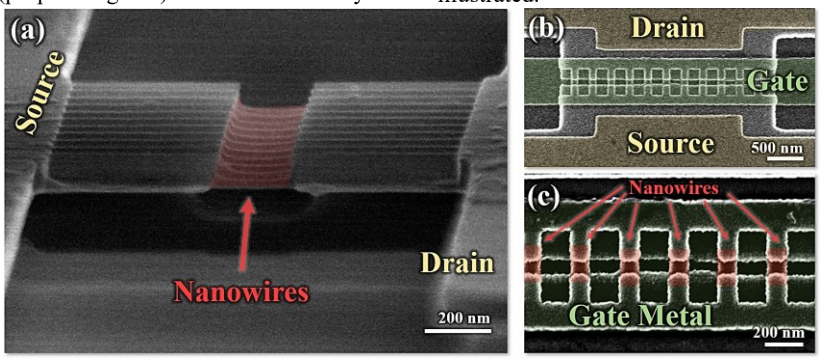


Fig. 6. SEM images of fabricated Germanium nanowire structures **(a)** viewed from the side before ALD ferroelectric oxide deposition and **(b)** top after completion of all fabrication processes. Nanowires connect source and drain regions and air-gap exists below the nanowire. **(c)** Zoom-in image of (b) shows multiple nanowires in parallel.

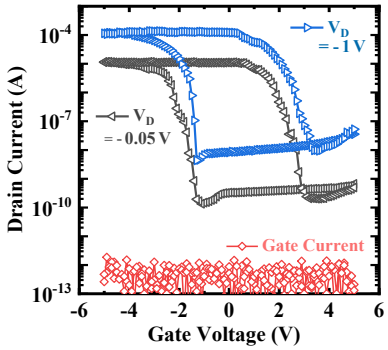


Fig. 7. I_D-V_G curve of Ge FE pNWFET (L= 105 nm, W= 50 nm, H= 26 nm) shows a clear negative ferroelectric hysteresis of approximately -5 V and negligible I_G.

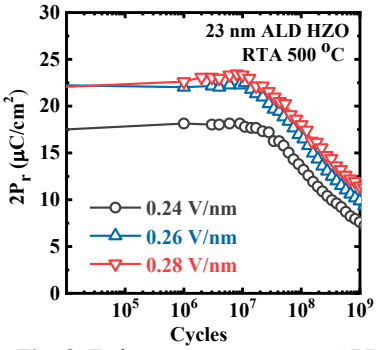


Fig. 8. Fatigue measurement on ALD Ferroelectric HZO capacitor was done with PUND pulses.

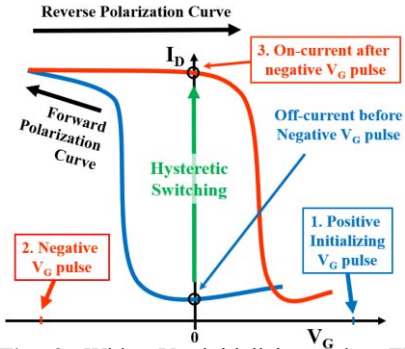


Fig. 9. With +V_G initializing pulse, FE pFET follows the forward polarization curve but a large -V_G switches the polarization and raises the I_D.

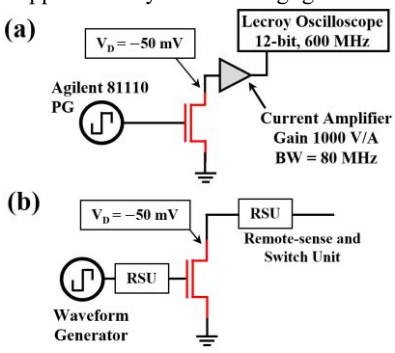


Fig. 10. (a) Set-up used for real-time conductance update during potentiation. **(b)** Set-up for conductance sampling between potentiation and depression pulses.

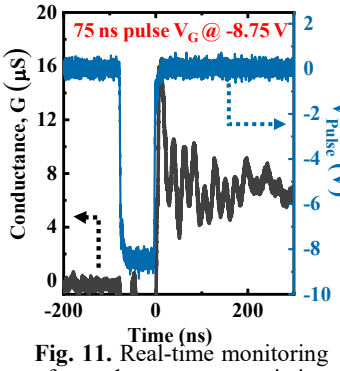


Fig. 11. Real-time monitoring of conductance potentiation with set up in Fig. 11 (a).

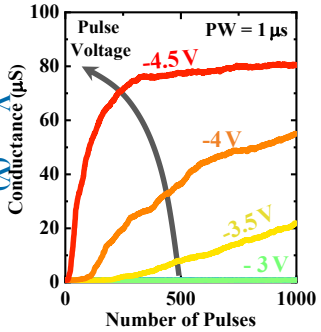


Fig. 12. Potentiation profile with fixed pulse width (1 μs). Higher voltage increases potentiation rate.

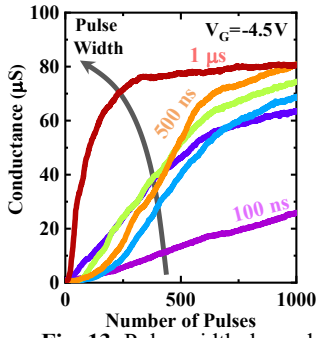


Fig. 13. Pulse width dependent potentiation profile with fixed V_G= -4.5 V. Longer pulses increase the potentiation rate.

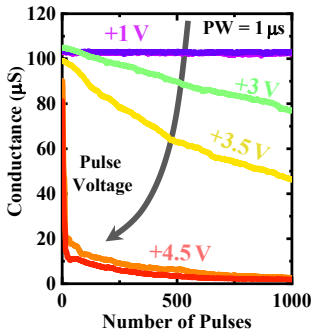


Fig. 14. Pulse voltage dependent depression profile with fixed pulse width of 1 μs showing similar trend as Fig. 12.

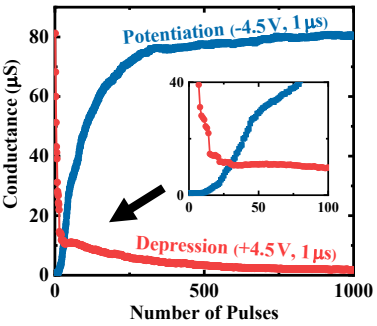


Fig. 15. Without optimized pulse conditions (±4.5 V, 1 μs), it results in highly non-linear and asymmetric profile

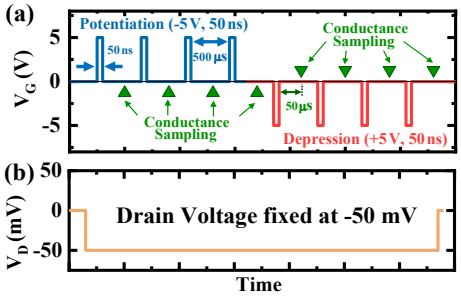


Fig. 16. (a) Optimized potentiation (-5 V, 50 ns) and depression (+5 V, 50 ns) pulses for the fabricated Ge FE pNWFET (L = 105 nm, W = 32 nm, H = 26 nm) **(b)** V_D is fixed at -50 mV.

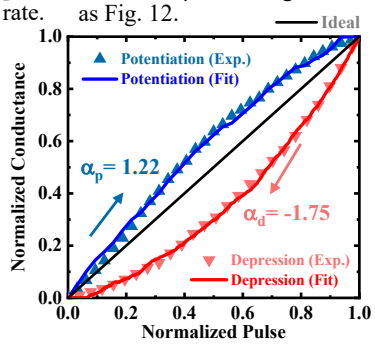
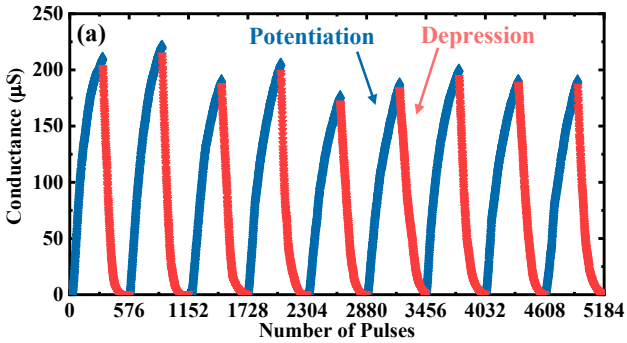


Fig. 17. The best nonlinearity coefficient from pulse scheme in Fig. 16 using reported model [1]. Only 10 % of pulses are displayed for better visualization



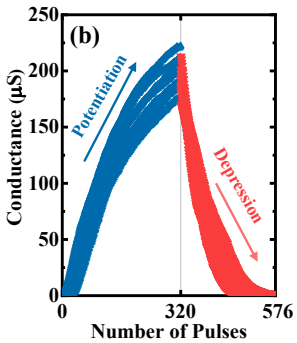


Fig. 18. (a) 9 cycles of consecutive alternating potentiation (-5 V, 50 ns, 320 pulses) and depression (+5 V, 50 ns, 256 pulses) give highly repetitive conductance profiles. **(b)** Overlapped curves from (a) show some conductance variation over multiple programming cycles.

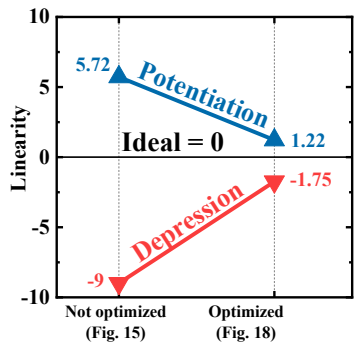


Fig. 19. Non-linearity comparison extracted from both optimized and not optimized pulses.

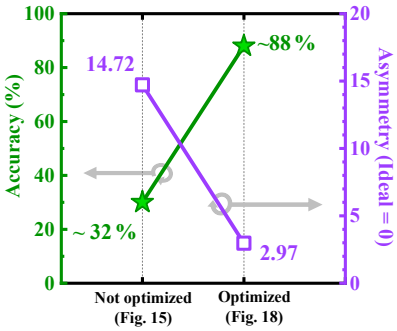


Fig. 20. Improvement in training accuracy after optimization. Total of 1 million cropped 20×20-pixel MNIST images were trained using MLP+NeuroSim V2.0 [1].

	This Work	[5]		[6]
Device	**Ge FE Nanowire pFET**	Si FE Planar nFET		Si FE Junctionless nFinFET
Gate Stack (Thickness, nm)	GeO_X (~1) + HZO (10) + Al_2O_3 (2)	HZO (10) + SiO_2 (0.8)		HZO (8.5) + SiO_2 (1.5)
Device Dimension	**L = 105 nm, W = 32 nm**	**L = 600 nm, W = 20,000 nm**		**L = 120 nm, W = 50 nm**
# States (Pot./Dep.)	**320 / 256**	20	32	> 32
Pot. Pulse (Type)	**(Identical) 50 ns**, 5 V	**(Identical)** 75 ns, 3.7 V	**(Varying)** 75 ns, 2.85 ~ 4.45 V	**(Identical)** 100 μs, 3.7 V
Dep. Pulse (Type)	**(Identical) 50 ns**, -5 V	**(Identical)** 75 ns, -3.2 V	**(Varying)** 75 ns, -2.1 ~ -3.8 V	**(Identical)** 100 μs, -3.2 V
Non-linearity (α_p/α_d)	1.22 / -1.75	5.54 / -8.08	1.75 / 1.46	1.58 / - 7.57
Asymmetry ($\lvert\alpha_p$-$\alpha_d\rvert$)	**2.97**	**13.62**	**0.29**	**9.15**
G_{max}/G_{min}	**Few hundreds**	~ 8	**45**	**4.98**
Accuracy (# of trained images)	~ 88 % (1 Million)	N/A	~ 90 % (1 Million)	~ 80 % (3 Million)

Fig. 21. Benchmark of various reported FeFET-based synapse devices for online learning. Lower non-linearity, asymmetry coefficients and higher on/off ratio are preferred. Cycle to cycle variation during our potentiation and depression process is < 1 %. Accuracy can be further increased with better optimized simulation conditions including various learning rates and circuit parameters.