Received 10 November 2020; revised 5 December 2020 and 8 December 2020; accepted 13 December 2020. Date of publication 16 December 2020; date of current version 28 January 2021. The review of this article was arranged by Editor J. Wang.

Digital Object Identifier 10.1109/JEDS.2020.3045194

Improving the Performance of Charge Trapping Memtransistor as Synaptic Device by Ti-Doped HfO_2

YU-CHE CHOU, WAN-HSUAN CHUNG, CHIEN-WEI TSAI, CHIN-YA YI, AND CHAO-HSIN CHIEN (Member, IEEE)

Institute of Electronics, National Chiao Tung University, Hsinchu 30010, Taiwan

CORRESPONDING AUTHOR: C.-H. CHIEN (e-mail: chchien@faculty.nctu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan under Grant MOST 107-2221-E-009-095-MY3, and in part by "Center for the Semiconductor Technology Research" under Grant MOST-108-3017-F-009-003.

ABSTRACT In this work, we improved the performance of germanium (Ge) channel Charge Trapping MemTransistors (CTMTs) as synaptic device by using Ti-doped HfO_2 as charge trapping layer (CTL). We manipulated the amount of Ti dopant within the HfO_2 CTL to perform the band engineering by varying the Hf/Ti cycle ratio in atomic layer deposition (ALD). The content of Ti was quantified and the energy band structures of the gate stack was constructed with the aid of transmission electron microscope (TEM) images and X-ray photoelectron spectroscopy (XPS) analysis. We then fabricated the charge trapping capacitors and characterized their memory characteristics such as memory windows. By the implementation of amphoteric trap model, thermal activated electron retention model and advanced charge decay model, the trap distribution of the CTL was extracted. Finally, we fabricated the CTMTs with Ti-doped HfO_2 as the CTL and characterized their performance as synaptic device such as nonlinearity of depression and potentiation and also conductance on/off ratio. We used NeuroSim simulator with multilayer perceptron and convolutional neural network models to evaluate the pattern recognition accuracy of neural network hardware accelerator using CTMTs as synaptic devices and benchmarked the performance of our CTMT with those of other types of synaptic devices.

INDEX TERMS Germanium, dielectric materials, neural network hardware, analog memories, artificial intelligence, MOSFETs, pattern recognition, semiconductor memories.

I. INTRODUCTION

Since the von Neumann architecture was firstly proposed by Burks *et al.* [1], it has been the supreme guideline of computer architectures. The explosive growth of computational power based on von Neumann architecture has driven the improvement of the entire human society throughout whole information era. Nevertheless, an new form of algorithm called artificial neural network (ANN) [2] had burgeoned in 90s including multilayer perceptron (MLP) [3], support vector machine (SVM) [4] and convolutional neural network (CNN) [5] which created a whole new type of request for computational power. Therefore, a new kind of computer architecture, i.e., neuromorphic architecture, was proposed in the meantime [6]. Although the concept of neuromorphic architecture was proposed in 90s, lacking the implementation of synaptic device made it remain in concept for a long period. Therefore, the later-blooming artificial neural networks such as AlexNet [7], GoogLeNet [8], VGG [9] and ResNet [10] are all based on the hardware with von Neumann architecture such as graphic processing units (GPU) [11], field-programmable gate arrays (FPGA) [12] and tensor processing units (TPU) [13] etc. With these off-the-shelf hardware platforms, ANN has dominated various application fields such as pattern recognition, object detection, computer vision and speech recognition. Nevertheless, the bottlenecks of von Neumann

architecture for ANN implementation are on-chip memory capacity, off-chip memory bandwidth and latency [14]. Generally speaking, all of off-the-shelf hardware platforms with von Neumann architecture mentioned above uses layers of cache which consists of static random-access memory (SRAM) as on-chip memory and dynamic random-access memory (DRAM) as off-chip memory. Although the scaling of SRAM can benefit from advanced technology node, its cell area per bit is still around 100-200 F^2 which limits its capacity to a few megabytes. It is insufficient for storing the parameters of ANN algorithms which are typically hundreds of megabytes [15]. Another drawback of the von Neumann architecture on ANN applications is its inefficient use of energy and memory bandwidth. The energy and bandwidth are mostly consumed by the movements of data from memory units to arithmetic units or the calculated results into memory units rather than the computation itself [16]. Therefore, the concept of neuromorphic architecture [6] was brought to the table again. But in this time, with the implementations of synaptic devices, neuromorphic computing known as in-memory computing has become a popular topic in various research field. There are various types of synaptic device such as flash memory [17], phase change memory [18], resistive random access memory [19]–[23], magnetic random access memory [24], [25], ferroelectric field-effect transistor [26] and SONOS type device [27]. Even though these candidates have been shown obvious advantage in the ANN applications, however, the degradation of recognition accuracy caused by the nonlinearity of weight-to-pulse and preneuron-to-postneuron relations have been a great challenge in recent studies[26], [28], [29]. Therefore, we have proposed the Charge Trapping MemTransistor (CTMT) beforehand [30], [31] which has comprehensive improvements in the abovementioned criteria.

In this study, we further improved the performance of CTMT on germanium (Ge) as synaptic device by using Ti-doped HfO_2 as charge trapping layer (CTL). We chose Ge substrate because it has high electron mobility [32]. Therefore, we can achieve the same drive current under lower bias to reduce the energy consumption during read-out operations. More importantly, the Ge/GeO_2 gate stack has smaller energy barrier than Si/SiO_2 gate stack which depicts higher injection current. Therefore, using Ge/GeO_2 gate stack can not only enlarge the memory window but also enhance the program speed [33]. We chose Al_2O_3 as the tunneling layer because of its high bandgap (E_G), conduction band offset (ΔE_C) and capability of stopping the diffusion of Ge into the high-κ dielectric, which would lead to higher interfacial trap density [34]. We adjusted the amount of Ti dopant within the HfO_2 CTL to perform the band engineering on it by utilizing the atomic layer deposition (ALD) system and adjusting the Hf/Ti cycle ratio during the deposition of CTL. We firstly quantified the Ti content and constructed the energy band structures of the gate stack of CTMTs with the aid of transmission electron microscope (TEM) images and X-ray photoelectron spectroscopy (XPS) analysis. We then fabricated the charge trapping capacitors (CTCs) with various CTLs as gate stack. We characterized their memory characteristics such as memory windows and retention. Based on the amphoteric trap model [35], thermal activated electron retention model [36] and advanced charge decay model [37], we performed the retention measurement under elevated temperature and extracted the trap distribution within the energy band. The effect of the amount of Ti dopant was then discussed. We then fabricated CTMTs using Ti-doped HfO_2 as CTL and characterized their synaptic characteristics such as the nonlinearity of pulse-number-to-weight relation, number of weight states and conductance on/off ratio. Finally, we utilized the NeuroSim [38], [39] simulator with MLP and CNN as ANN models to benchmark the CTMT as synaptic device against other types of synaptic device [19]–[23], [26] and our previous CTMT [31].

II. FABRICATION OF CHARGE TRAPPING CAPACITORS

First, the (100)-oriented p-type Ge substrate was cleaned with diluted hydrofluoric acid. A GeO_2 interfacial layer (IL) and high-κ gate stack were then deposited using a plasma-enhanced ALD (PEALD) system in situ. A 40-cycle Al_2O_3 was deposited as tunneling layer (TL) by PEALD using O_2 plasma as oxygen source. A 60-cycle HfTiO was deposited as the CTL by thermal ALD using H_2O precursor as oxygen source. We adjusted the Hf/Ti cycle ratio during the deposition of CTL to vary the amount of Ti dopant within the CTL. The cycle ratio between HfO_2 and TiO_2 in the case of H9T1, H19T1 and H29T1 was 9:1, 19:1 and 29:1, respectively. In the case of HT, there was no H_2O precursor injection between the injections of Hf precursor (TDMAH) and Ti precursor (TDMAT) and the cycle ratio of TDMAH and TDMAT precursor is 1:1. In the case of HfO_2, pure HfO_2 was used as CTL for reference. A 60-cycle Al_2O_3 was then deposited as barrier layer (BL) by PEALD. Afterwards, a 400°C 60s rapid thermal annealing was performed as post-deposition annealing for improving the quality of the high-κ dielectric, and then a 50 nm TiN was deposited using physical vapor deposition (PVD) and patterned as gate metal. Finally, a 10 nm Ti and a 300 nm Al were deposited with PVD as the backside contact. Fig. 1 shows the process flow and TEM image of the gate stack. The process steps of CTC are colored in red. The physical thickness of TL, CTL and BL is 3.4, 4.8 and 5.2 nm, respectively.

$$\text{Ti content}(\%) = \frac{A_{Ti,2p}/S_{Ti,2p}}{A_{Hf,4d}/S_{Hf,4d} + A_{Ti,2p}/S_{Ti,2p}} \quad (1)$$

where A is the area of the XPS peak and S is the sensing factor of the orbital

III. PHYSICAL AND MEMORY CHARACTERISTICS OF CHARGE TRAPPING CAPACITORS

We firstly used the XPS analysis to determine the content of Ti because in the case of H19T1, H29T1 and HT, the content

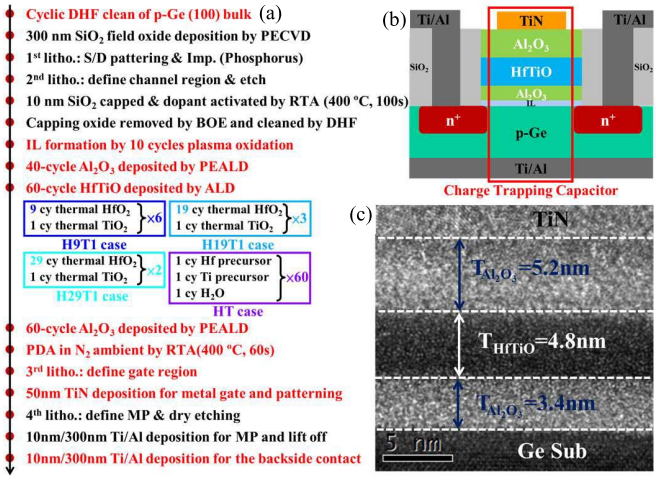

FIGURE 1. **(a) Process flow of CTCs and CTMTs. The process steps of CTC are colored in red. (b) Illustration of the structure of CTC and CTMT. The structure of CTC is framed by red line. (c) TEM image of the gate stack.**

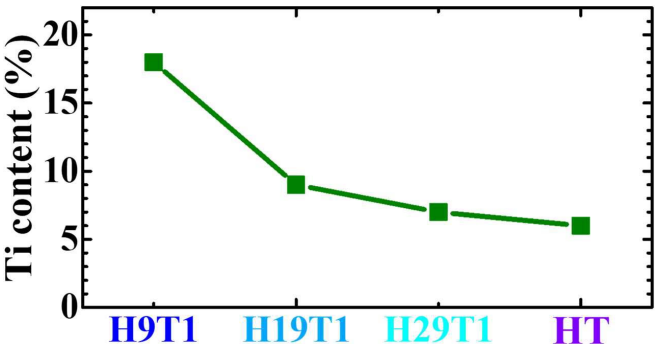

FIGURE 2. **Ti content versus Hf/Ti cycle ratio extracted by XPS.**

of Ti was too inadequate to detect by energy-dispersive X-ray spectroscopy (EDS) of our TEM system. The Ti content was defined by (1). Through the area of XPS peak divided by its own sensing factor, we are able to extract the Ti content within the HfTiO high-κ dielectrics. The Ti content versus Hf/Ti cycle ratio plot is shown in Fig. 2. The Ti content was reduced from 18% in the case of H9T1 to 7% in the case of H29T1 because of the increment of Hf/Ti cycle ratio from 9:1 to 29:1. However, the Ti content was further reduced to 6% in the case of HT because of the lack of H_2O precursor injection between the injection of TDMAH and TDMAT. Therefore, there is no sufficient surface bonding for TDMAT to deposit on the surface which led to further reduction of Ti content within the HfTiO layer. We then extracted the energy band structure of the gate stacks with the results of XPS analysis [40]. The illustration of the energy band structure of the gate stack is shown in Fig. 3(a) and the plot of conduction band offset of HfTiO and Al_2O_3 ($\Delta E_{C,AH}$) and the bandgap of HfTiO versus Hf/Ti cycle ratio are shown in Fig. 3(b). The increment of $\Delta E_{C,AH}$ with the reduction of Ti content was mainly attributed to the reduction of the bandgap of HfTiO. Therefore, the energy

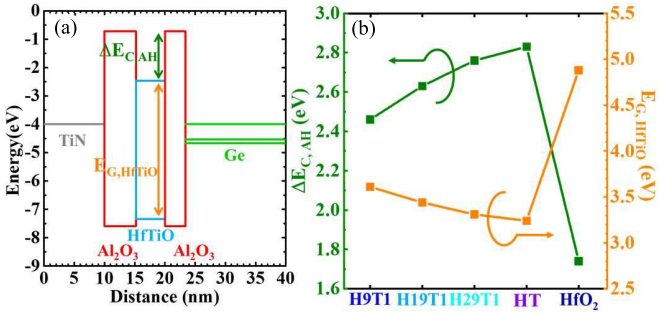

FIGURE 3. **(a) Illustration of the energy band structure of the gate stack of CTMT. (b) conduction band offset of HfTiO and Al_2O_3 ($\Delta E_{C,AH}$) and bandgap of HfTiO versus Hf/Ti cycle ratio.**

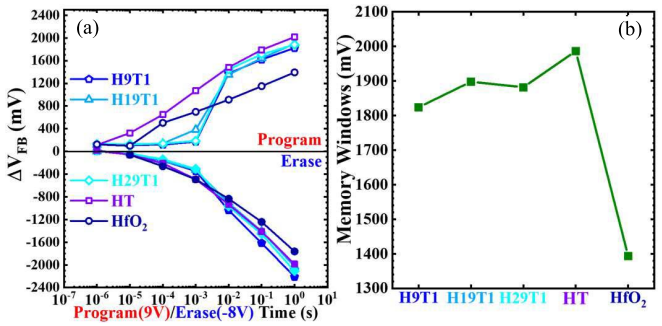

FIGURE 4. **(a) Flat-band-voltage shift (ΔV_{FB}) to program/erase pulse width plot. (b) Memory windows of CTCs.**

barrier for the electrons trapped in the CTL was enlarged with the reduction of Ti content within the HfTiO layer.

After the physical analyses of the gate stacks, we then characterized the writing characteristics of CTCs including the flat-band-voltage shift (ΔV_{FB}) to pulse width and memory windows which are shown in Figs. 4(a) and (b). Comparing with the case of HfO_2 as reference [31], using HfTiO as CTL can enlarge the memory windows. In the case of H9T1, H19T1, H29T1 and HT, the memory window is enlarged by 430.7, 504.8, 488.7 and 593.5 mV. Among all cases, the case of HT outperformed other cases in all pulse width which can not only improve the memory window but also program speed. Next, the methodology of the extraction of the trap distribution is based on the amphoteric trap model [35], the thermal activated electron retention model [36] and the advanced charge decay model [37]. In the amphoteric trap model [35], three types of trap state and the relation between trap states caused by charge capturing and releasing were clearly defined. The thermal activated electron retention model [36] considered only neutral trap state with no charge captured and negative trap state after capturing an electron. In the thermal activated electron retention model [36], three ways of electrons escaping from trap were considered including thermal excitation, trap-to-band tunneling and trap-to-trap tunneling. However, when the temperature elevates above 450K, the thermal excitation dominates due to its lower time constant and the excitation rate is affected by the energy level of traps [36]. In

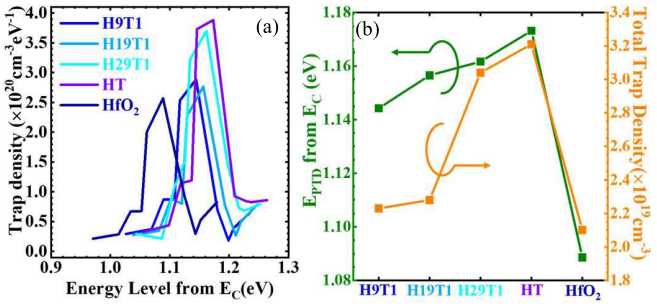

FIGURE 5. (a) Trap distribution within the energy band. (b) Energy level from E_C with peak trap density and total trap density versus Hf/Ti cycle ratio.

the advanced charge decay model [37], the relation between excitation rate and energy level of traps was derived and the trap distribution could be extracted by the retention measurement under elevated temperature. Therefore, we performed the retention measurement at elevated temperature and extracted the trap distribution within the energy band. Fig. 5(a) shows the trap distribution within the energy band in each case. Fig. 5(b) shows the energy level with peak trap density (E_{PTD}) from conduction band (E_C) and total trap density versus Hf/Ti cycle ratio. With the reduction of Ti content within the HfTiO, the E_{PTD} shifted away from E_C which is beneficial for data retention. Moreover, the excessive Ti cycles could result in proper TiO_2 formation. Obviously, this did not fit our purpose because we intended to create charge trap centers by adding Ti as dopants. Therefore, the total trap density increases with the reduction of Ti content which consists with the result of memory-window enlargement shown in Fig. 4(b).

IV. SYNAPTIC CHARACTERISTICS OF CHARGE TRAPPING MEMTRANSISTORS

After characterizing the physical and memory characteristics of CTC, we then fabricated the CTMTs with the gate stack of HfO_2, H29T1 and HT cases. The process flow is shown in Fig. 1. We firstly characterized the synaptic characteristics of CTMTs. By applying positive/negative pulse on the gate of CTMT, we are able to shift the threshold voltage (V_T) positively/negatively because of the trapped charges in CTL. Therefore, the channel conductance of CTMT as synaptic device which represents the weight stored in the synapse will reduce/increase. The illustration of the operation principle of CTMT as synaptic devices is shown in Fig. 7(c). By applying multiple positive/negative pulses, we can manipulate the reduction/increment of the conductance which is called depression/potentiation of the weight.

From the device perspective, there are three key factors which affect the recognition accuracy. First one is the nonlinearity of the potentiation (α_p) and depression (α_d) of the weight [29]. The closer to zero the nonlinearity is, the more linear the potentiation or depression is. The more linear the potentiation or depression is, the more precise the weight is during the training phase. Second one is the conductance

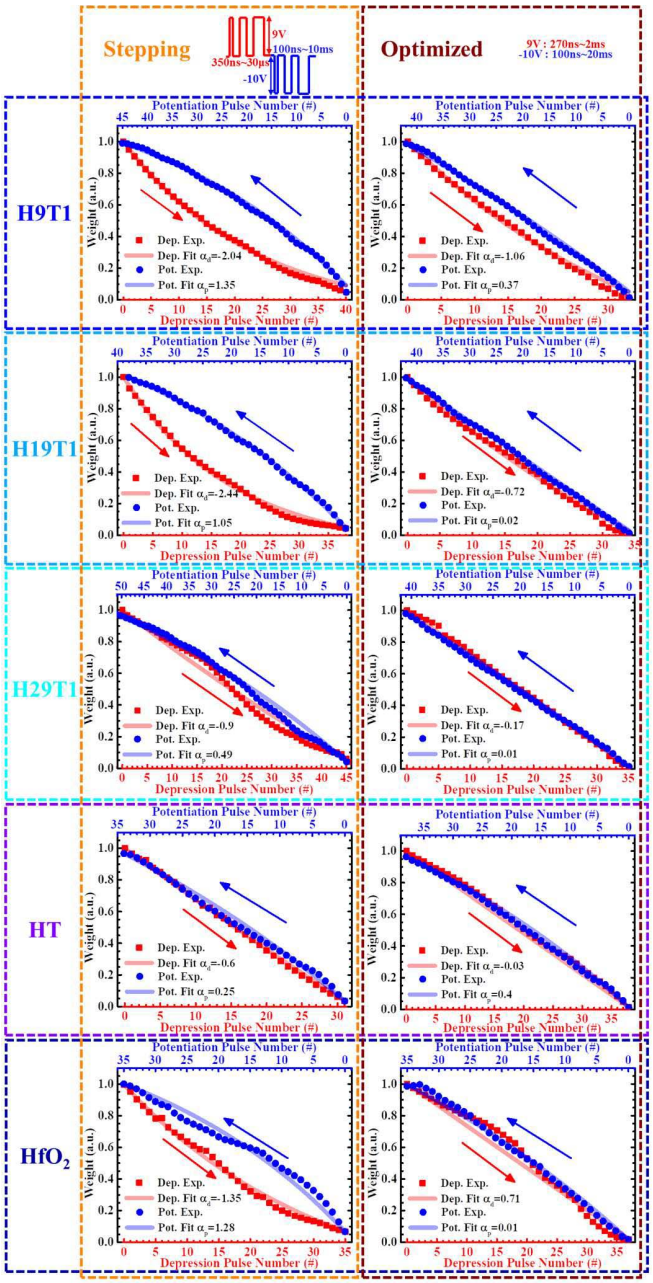

FIGURE 6. Weight depression/potentiation curves of H9T1, H19T1, H29T1, HT and HfO_2 cases under stepping and optimized pulse scheme.

on/off ratio. In practice, we normalize the maximum conductance to one to represent the maximum weight. Therefore, the conductance on/off ratio represents the minimum value of the weight that a synaptic device can store. Ideally, the minimum value of the weight is zero which indicates that the conductance on/off ratio is infinite. However, it is impractical for any kind of synaptic device. Therefore, the higher the impractical ratio is, the smaller the minimum weight is. The smaller the minimum weight is, the larger the usable range of the weight is which leads to higher recognition accuracy. Last one is the weight precision which is the number of conductance states in the depression/potentiation of the weight.

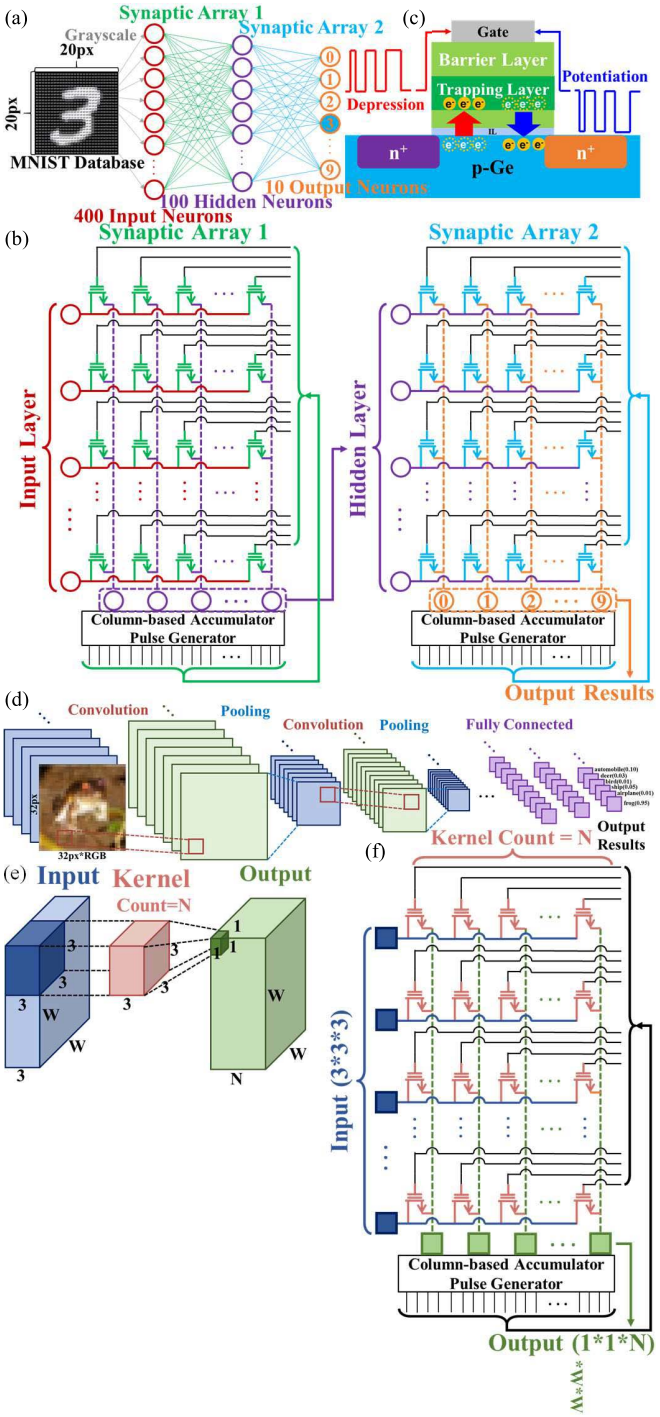

FIGURE 7. **(a) Illustration of MLP network with one hidden layer for MNIST database recognition. (b) Architecture of CTMT synaptic array for MLP hardware acceleration derived from (a). (c) Operation principle of CTMT as synaptic devices. (d) Illustration of CNN network using VGG-8 model for CIFAR-10 database recognition. (e) Illustration of convolution operation in VGG-8 model. (f) Architecture of CTMT synaptic array for convolution operation derived from (e).**

According to our simulation by NeuroSim [38], [39] with MLP and CNN as ANN models, there is an accuracy saturation occurring when the precision of the weight is higher than 5 bits, i.e., the number of states is greater than 32 state.

TABLE 1. **Parameters of MLP network.**

	Input Elements	Output Elements	Number of Synapses
Layer 1	20*20	100	20*20*100
Layer 2	100	10	100*10

The weight depression/potentiation curves of CTMTs are shown in Fig. 6. In order to achieve low nonlinearity and high conductance on/off ratio, we implemented non-identical pulse schemes with pulse width modulation. Comparing to the pulse height modulation, the pulse width modulation we implemented in this work has less complexity in peripheral, but it will cause higher latency. By applying the stepping pulse scheme in which pulse width increased in a constant difference with the increase of pulse number, we were able to achieve low nonlinearity in the case of H29T1 and HT. However, in the case of H9T1 and H19T1, once the Ti dopant was excessive, there was no improvement in nonlinearity. The (α_p, α_d) of H9T1, H19T1, H29T1, HT and HfO_2 case was (1.35, −2.04), (1.05, −2.44), (0.49, −0.9), (0.25, −0.6) and (1.28, −1.35) respectively. The conductance on/off ratio in the case of H9T1, H19T1, H29T1, HT and HfO_2 under the stepping pulse scheme was 20.8, 22.5, 24.8, 28.7 and 18.1 respectively. These results indicated that using HfTiO as CTL in CTMT can improve both linearity of weight depression/potentiation and conductance on/off ratio because of its trap distribution within the energy band and total trap density accordingly. In order to further improve the nonlinearity and conductance on/off ratio, we applied the optimized pulse scheme. The range of pulse width is listed in Fig. 6. Within this range, we changed the increment of pulse width based on the stepping pulse scheme according to the amount of change in weight after applying previous pulse. Therefore, the nonlinearity was optimized by the optimized increment of pulse width. The on/off ratio was also improved because of the larger increment of pulse width. The (α_p, α_d) of H9T1, H19T1, H29T1, HT and HfO_2 case was (0.37, −1.06), (0.02, −0.72), (0.01, −0.17), (0.4, −0.03) and (0.71, 0.01) respectively. The conductance on/off ratio in the case of H9T1, H19T1, H29T1, HT and HfO_2 under the optimized pulse scheme was 69.2, 71.4, 76.5, 82.2 and 66.2 respectively. The CTMTs with HfTiO as CTL still had better nonlinearity and conductance on/off ratio than the one with HfO_2 as CTL under optimized pulse scheme only if the Ti dopant was not excessive. The nonlinearity and conductance on/off ratio of H29T1, HT and HfO_2 cases are shown in Figs. 8(a) and (b). The H9T1 and H19T1 cases were skipped for the further investigation due to the insufficient improvement in nonlinearity.

Finally, we benchmarked the pattern recognition accuracy the CTMTs by NeuroSim [38], [39]. We used two kinds of models to inference two different datasets. First, we constructed the MLP model with 400 input neurons, one hidden layer with 100 hidden neurons and 10 output neurons for MNIST, a 28px×28px hand-written-number-pattern dataset, inference. The model is illustrated in Fig. 7(a) and

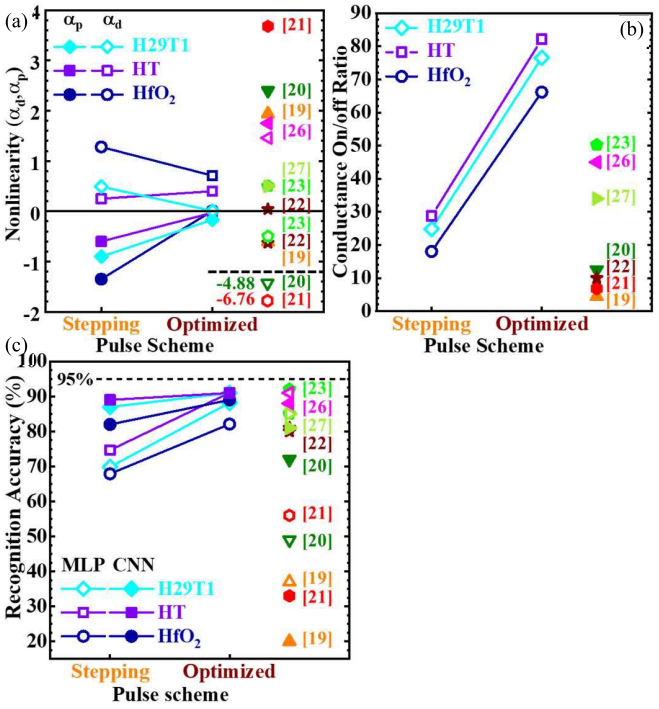

FIGURE 8. (a) Plot of nonlinearity (α_p, α_d) versus pulse schemes with three kinds of gate stacks. (b) Plot of conductance on/off ratio versus pulse schemes with three kinds of gate stacks. (c) Plot of recognition accuracy of MLP network on MNIST dataset and CNN (VGG-8) network on CIFAR-10 dataset versus pulse schemes with three kinds of gate stacks.

TABLE 2. Parameters of VGG-8 model.

	Input Dimension	Kernel Dimension	Kernel Count	Pooling	Fully Connected
Layer 1	32*32*3	3*3*3	128	N	N
Layer 2	32*32*128	3*3*128	128	Y	N
Layer 3	16*16*128	3*3*128	256	N	N
Layer 4	16*16*256	3*3*256	256	Y	N
Layer 5	8*8*256	3*3*256	512	N	N
Layer 6	8*8*512	3*3*512	512	Y	N
Layer 7	1*1*8192	1*1*8192	1024	N	Y
Layer 8	1*1*1024	1*1*1024	10	N	Y

the parameters of the model are shown in Table 1. It takes two synaptic arrays to fully implement this MLP model. The first one connects the input layer and the hidden layer and is comprised of 400×100 synapses. The second one connects the hidden layer and the output results and is comprised of 100×10 synapses which is illustrated in Fig. 7(b). By applying the stepping pulse scheme, the recognition accuracy in the case of H29T1, HT and HfO_2 was 69.9%, 74.7% and 67.9% respectively. By applying the optimized pulse scheme, the recognition accuracy in the case of H29T1, HT and HfO_2 was 88.2%, 91.1% and 82.1%, respectively. HT case has the highest recognition accuracy because of its low nonlinearity and highest conductance on/off ratio. Next, we constructed the CNN model with VGG-8 model for CIFAR-10, a 32px×32px×RGB color-pattern dataset,

TABLE 3. Benchmark of synaptic devices.

Type of Synaptic device	AlO_x/HfO_2 RRAM [19]	Ag:a-Si RRAM [20]	PCMO RRAM [21]	TaO_x/HfO_x RRAM [22]	Ag:SiGe EpiRAM [23]	HZO FeFET [26]	SONOS GSD [27]	HfO_2 CTMT [31]	HfTiO CTMT (This work)
# of states	40	97	50	128	64	32	30	35	**35**
Nonlinearity (α_p/α_d) [29]	1.94/ -0.61	2.4/ -4.88	3.68/ -6.76	0.04/ -0.63	0.5/ -0.5	1.75/ 1.46	0.5/ - (est.)	0.06/ -0.89	**0.17/ 0.01**
On/off ratio	4.43	12.5	6.84	10	50.2	45	34 (est.)	75.3	**82.2**
R_{ON} (Ω)	16.9 k	26 M	23 M	100 k	81 k	559.3 k	87 M	2.08 M	**1.26 M**
Area for MLP (μm^2)	21760.0	6292.3	6292.4	8663.1	9144.3	7032.6	-	7032.6	7032.6
Accuracy of MLP on MNIST (%)	~20	~72	~33	~80	92	88	~81	87	**91**
Accuracy of VGG-8 on CIFAR-10 (%)	37	49	56	81	85	91	~85	89	**91**

Numbers in green mean the strengths while numbers in red mean the weaknesses.

inference which is shown in Fig. 7(d). The model consists of convolution, pooling and fully connected layer; parameters are shown in Table 2. The main operation to accelerate here is the convolution operation which is shown in Fig. 7(e). The dimension of the synaptic array to perform a convolution operation with the input dimension of W×W×3, the kernel dimension of 3×3×3 and N as the count of kernels is (3×3×3)×N. The number of rows to perform one convolution operation is W×W and will produce W×W rows of output with the dimension of 1×1×N. Therefore, the dimension of output is W×W×N. The architecture of CTMT synaptic array for the convolution operation mentioned above is shown in Fig. 7(f). The recognition accuracy in the case of H29T1, HT and HfO_2 was 87%, 89% and 82% under the stepping pulse scheme respectively. The recognition accuracy of H29T1, HT and HfO_2 case was 91%, 91% and 89% under the optimized pulse scheme respectively. The HT case also has the highest recognition accuracy in the CNN model and the optimized pulse scheme can further improve the recognition accuracy. Last but not least, Table 3 shows the benchmark of synaptic devices. Our CTMT outperforms other types of synaptic devices in the categories of nonlinearity and conductance on/off ratio and its recognition accuracy is close to the best-in-class.

V. CONCLUSION

In this work, we have discussed the effect of using Ti-doped HfO_2 as CTL on Ge CTMTs as synaptic devices. We first characterized the physical and memory characteristics of CTCs in terms of energy band structure, memory window and trap distribution. Using Ti-doped HfO_2 as the CTL of CTC enlarges the memory window because it has larger $\Delta E_{C,AH}$, deeper trap distribution and higher trap density. Next, we characterized the synaptic characteristics of CTMTs in terms of nonlinearity and conductance on/off ratio. Using Ti-doped HfO_2 as the CTL of CTMT can reduce the nonlinearity and increase the conductance on/off ratio because of its trap distribution and trap density. We then used NeuroSim to simulate the recognition accuracy of neural networks using CTMTs as synaptic device. Using Ti-doped HfO_2 as the CTL of CTMT can improve the recognition accuracy because of its lower nonlinearity and higher conductance on/off ratio. Last, we benchmarked our CTMT

with other synaptic devices. The CTMT outperforms former alternative candidates, such as RRAM, FeFET in terms nonlinearity, conductance on/off ratio and recognition accuracy. We believe our proposed CTMT is the one with great potential for the synaptic device in ANN application.

REFERENCES

[1] A. W. Burks, H. H. Goldstine, and J. Von Neumann, *Preliminary Discussion of the Logical Design of an Electronic Computing Instrument*. Princeton, NJ, USA: Inst. Adv. Study, 1946.

[2] Y. Xin, "Evolving artificial neural networks," *Proc. IEEE*, vol. 87, no. 9, pp. 1423–1447, Sep. 1999, doi: 10.1109/5.784219.

[3] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[5] 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, Nov. 1998, doi: 10.1109/5.726791.

[6] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, Oct. 1990, doi: 10.1109/5.58356.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran, 2012, pp. 1097–1105.

[8] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1–9.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: arXiv:1409.1556.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[11] S. Chetlur *et al.*, "cuDNN: Efficient primitives for deep learning," 2014. [Online]. Available: arXiv:1410.0759.

[12] G. Lacey, G. W. Taylor, and S. Areibi, "Deep learning on FPGAs: Past, present, and future," 2016. [Online]. Available: arXiv:1602.04283.

[13] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," *presented at the Proc. 44th Annu. Int. Symp. Comput. Archit.*, Toronto, ON, Canada, 2017, pp. 1–12. [Online]. Available: https://doi.org/10.1145/3079856.3080246

[14] G. Indiveri and S. Liu, "Memory and information processing in neuromorphic systems," *Proc. IEEE*, vol. 103, no. 8, pp. 1379–1397, Aug. 2015, doi: 10.1109/JPROC.2015.2444094.

[15] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, Feb. 2018, doi: 10.1109/JPROC.2018.2790840.

[16] G. Desoli *et al.*, "14.1 A 2.9TOPS/W deep convolutional neural network SoC in FD-SOI 28nm for intelligent embedded systems," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Feb. 2017, pp. 238–239, doi: 10.1109/ISSCC.2017.7870349.

[17] O. Fujita and Y. Amemiya, "A floating-gate analog memory device for neural networks," *IEEE Trans. Electron Devices*, vol. 40, no. 11, pp. 2029–2035, Nov. 1993, doi: 10.1109/16.239745.

[18] S. B. Eryilmaz *et al.*, "Experimental demonstration of array-level learning with phase change synaptic devices," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, USA, Dec. 2013, pp. 1–4, doi: 10.1109/IEDM.2013.6724691.

[19] J. Woo *et al.*, "Improved synaptic behavior under identical pulses using AlOx/HfO2 bilayer RRAM array for neuromorphic systems," *IEEE Electron Device Lett.*, vol. 37, no. 8, pp. 994–997, Aug. 2016, doi: 10.1109/LED.2016.2582859.

[20] 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, Apr. 2010, doi: 10.1021/nl904092h.

[21] S. Park *et al.*, "Neuromorphic speech systems using advanced ReRAM-based synapse," in *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, USA, Dec. 2013, pp. 1–4, doi: 10.1109/IEDM.2013.6724692.

[22] W. Wu *et al.*, "A methodology to improve linearity of analog RRAM for neuromorphic computing," in *Proc. IEEE Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2018, pp. 103–104, doi: 10.1109/VLSIT.2018.8510690.

[23] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nat. Mater.*, vol. 17, no. 4, pp. 335–340, Jan. 2018, doi: 10.1038/s41563-017-0001-5.

[24] D. Fan and S. Angizi, "Energy efficient in-memory binary deep neural network accelerator with dual-mode SOT-MRAM," in *Proc. IEEE Int. Conf. Comput. Design (ICCD)*, Boston, MA, USA, Nov. 2017, pp. 609–612, doi: 10.1109/ICCD.2017.107.

[25] Y. Luo, X. Peng, and S. Yu, "MLP+NeuroSimV3.0: Improving on-chip learning performance with device to algorithm optimizations," *presented at the Proc. Int. Conf. Neuromorphic Syst.*, Knoxville, TN, USA, 2019, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3354265.3354266

[26] M. Jerry *et al.*, "Ferroelectric FET analog synapse for acceleration of deep neural network training," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, Dec. 2017, pp. 1–4, doi: 10.1109/IEDM.2017.8268338.

[27] J. Bae, S. Lim, B. Park, and J. Lee, "High-density and near-linear synaptic device based on a reconfigurable gated schottky diode," *IEEE Electron Device Lett.*, vol. 38, no. 8, pp. 1153–1156, Aug. 2017, doi: 10.1109/LED.2017.2713460.

[28] S. Yu, P. 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. Electron Devices Meeting (IEDM)*, Washington, DC, USA, Dec. 2015, pp. 1–4, doi: 10.1109/IEDM.2015.7409718.

[29] P. Chen *et al.*, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Austin, TX, USA, Nov. 2015, pp. 194–199, doi: 10.1109/ICCAD.2015.7372570.

[30] Y. Chou, C. Tsai, C. Yi, W. Chung, and C. Chien, "Impact of the crystal phase of ZrO_2 on charge trapping memtransistor as synaptic device for neural network application," *IEEE J. Electron Devices Soc.*, vol. 8, pp. 572–576, 2020, doi: 10.1109/JEDS.2020.2993859.

[31] Y.-C. Chou, C.-W. Tsai, C.-Y. Yi, W.-H. Chung, S.-Y. Wang, and C.-H. Chien, "Neuro-inspired-in-memory computing using charge-trapping memtransistor on germanium as synaptic device," *IEEE Trans. Electron Devices*, vol. 67, no. 9, pp. 3605–3609, Sep. 2020, doi: 10.1109/TED.2020.3008887.

[32] C.-W. Chen, J.-Y. Tzeng, C.-T. Chung, H.-P. Chien, C.-H. Chien, and G.-L. Luo, "High-performance germanium p- and n-MOSFETs with NiGe source/drain," *IEEE Trans. Electron Devices*, vol. 61, no. 8, pp. 2656–2661, Aug. 2014, doi: 10.1109/TED.2014.2327620.

[33] Z.-H. Ye, K.-S. Chang-Liao, L.-J. Liu, J.-W. Cheng, and H.-K. Fang, "Enhanced programming and erasing speeds of charge-trapping flash memory device with Ge channel," *IEEE Electron Device Lett.*, vol. 36, no. 12, pp. 1314–1317, Dec. 2015, doi: 10.1109/LED.2015.2495344.

[34] Y. Tsai *et al.*, "Improving thermal stability and interface state density of high-κ stacks by incorporating Hf into an interfacial layer on p-Germanium," *IEEE Electron Device Lett.*, vol. 37, no. 11, pp. 1379–1382, Nov. 2016, doi: 10.1109/LED.2016.2613999.

[35] J. Robertson, "The electronic properties of silicon nitride," *Philos. Mag. B*, vol. 44, no. 2, pp. 215–237, Aug. 1981, doi: 10.1080/01418638108222558.

[36] Y. Yang and M. H. White, "Charge retention of scaled SONOS nonvolatile memory devices at elevated temperatures," *Solid-State Electron.*, vol. 44, no. 6, pp. 949–958, Jun. 2000. [Online]. Available: https://doi.org/10.1016/S0038-1101(00)00012-5

[37] T. H. Kim, J. S. Sim, J. D. Lee, H. C. Shin, and B.-G. Park, "Charge decay characteristics of silicon-oxide-nitride-oxide-silicon structure at elevated temperatures and extraction of the nitride trap density distribution," *Appl. Phys. Lett.*, vol. 85, no. 4, pp. 660–662, Jul. 2004, doi: 10.1063/1.1773615.

[38] P. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067–3080, Dec. 2018, doi: 10.1109/TCAD.2018.2789723.

[39] X. Peng, S. Huang, H. Jiang, A. Lu, and S. Yu, "DNN+NeuroSim V2.0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," 2020. [Online]. Available: arXiv:2003.06471.

[40] M. Perego, G. Seguini, and M. Fanciulli, "Energy band alignment of HfO_2 on Ge," *J. Appl. Phys.*, vol. 100, no. 9, Aug. 2006, Art. no. 093718, doi: 10.1063/1.2360388.