Mimicking Synaptic Behaviors with Junctionless Transistor for Low Power Neuromorphic Computing

Md. Hasan Raza Ansari, Hanrui Li, and Nazek El-Atab*

SAMA Labs, Computer, Electrical and Mathematical Science and Engineering Division (CEMSE) King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia, email: nazek.elatab@kaust.edu.sa

***Abstract*—This work highlights the application of a junctionless (JL) transistor with charge trapping mechanism as an artificial synaptic device for neuromorphic computing. In this work, synapse behaviors ((short-term potentiation (STP), long-term potentiation (LTP), and depression (LTD))) have been validated and analyzed by storing the positive charges (holes) in the floating body and charge trapping nitride layer. JL device can be operated at a lower drain voltage (V_{DS} = 0.8 V) to trigger the band-to-band tunneling and impact ionization mechanisms. The device achieves a higher and linear conductance value, and the non-linearity value for LTP is 0.1, which is beneficial for neural networks. Estimated conductance values from the device are utilized to estimate the pattern recognition and achieve an accuracy of ~ 85 % with the CNN algorithm and CIFAR-10 datasets.**

Keywords—Junctionless, Charge Trapping Memory, LTP, LTD, Neural Network.

I. Introduction

Neuromorphic computing is showing a great interest in overcoming the limitation of von Neumann computing [1]–[4]. The development of energy-efficient and highly integrated electronic synapses for neuromorphic computing have been essential in trying to emulate adaptive learning and memory in biological neural networks [1]–[5]. Synapses are the basic units that connect neurons and create neuromorphic structures in biological systems [6]. The synapse features such as (paired-pulse facilitation (PPF), short-term potentiation (STP), long-term potentiation (LTP), long-term depression (LTD), and time-dependent spike plasticity (STDP) are mimicked through different electronic devices [7]–[10]. Among them, two terminal memristive devices for artificial synapses are attractive due to being analogous to biological synapses and can modulate synapse plasticity (strength). However, memristive devices have low endurance and reproducibility issues. Also, these devices are implemented by non-silicon devices, which have the issue of co-integration with conventional complementary metal oxide semiconductor (CMOS) architecture. These issues open the room for further improvement and realization of artificial synapses with three-terminal devices. The charge-trapping memory shows the analog behavior and is applicable for artificial synapses [11]. Conventional inversion mode metal oxide semiconductor field effect transistors (MOSFETs) face the issue of short channel effects and also the formation of an ultrasharp *p-n* junction at nanoscale regime during gate length scaling for high density and fast computing [12]–[17]. A junctionless transistor with the same doping type throughout the source to drain can overcome these issues [12]–[17].

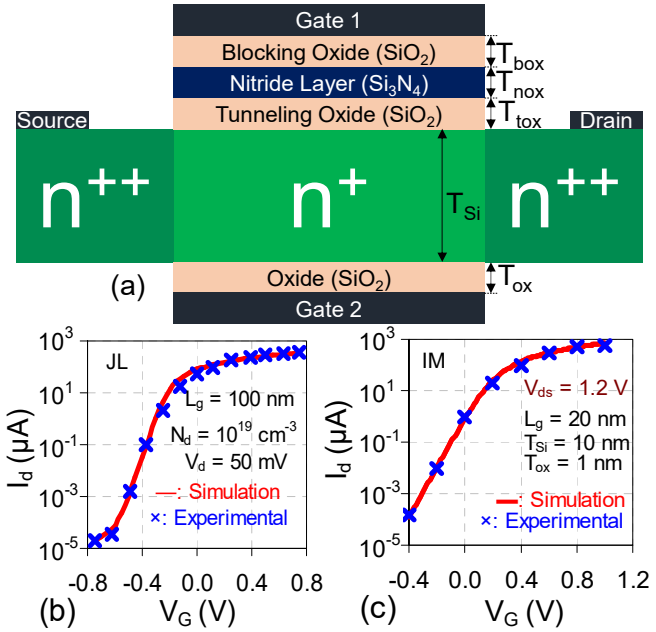

Fig. 1. (a) Schematic of a junctionless transistor with charge trapping mechanism as an artificial synapse for the neural network. Comparison of simulation transfer characteristics with experimental transfer characteristics of (b) junctionless (JL) [17] and (c) inversion mode (IM) [18] device.

In this work, a junctionless (JL) transistor is simulated with charge trapping mechanism, and its synaptic behaviors (STP, LTP, and LTD) have been validated. The advantage of the junctionless transistor is that it can be operated at a lower drain bias, which can trigger the Band to band tunneling and impact ionization [12], [19], [20]. Initially, the potentiation operation is based on the BTBT mechanism, which helps the device to consume less power and energy. In addition, the device conductance shows non-linearity (NL) values is 0.1 for LTP and 2.7 for LTD. The device is also feasible for demonstrating the pattern recognition simulation using the Canadian Institute For Advanced Research (CIFAR-10) dataset. Convolution Neural Network (CNN) algorithm shows outstanding performance for large datasets, and in this work, we have performed CNN algorithm simulation with the CIFAR-10 dataset and achieved an accuracy of 85 %.

This work was supported by the King Abdullah University of Science and Technology baseline fund.

2022 IEEE International Conference on Emerging Electronics (ICEE) | 978-1-6654-9185-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICEE56203.2022.10118253

978-1-6654-9185-3/22/$31.00 ©2022 IEEE

II. Simulation Methodology

Fig. 1(a) shows the double gate *n*-type junctionless (JL) transistor with charge trapping memory mechanism (oxide/nitride/oxide) for an artificial synaptic device. The device simulations are performed through Silvaco ATLAS [21] with calibrated models for junctionless (JL) [17] and inversion mode (IM) [18] devices as shown in Fig. 1(b) and (c), respectively. The device operation as an artificial synapse is based on charge generation, recombination, trapping, and de-trapping [9]. In order to capture the charge generation and recombination in the silicon film, non-local band-to-band tunneling (BTBT) and impact ionization (II) models are used. In addition, concentration dependent SRH, field dependent mobility model, bandgap narrowing model for highly doped semiconductor film, and bipolar model for the silicon-on-insulator device are used. The dynamic dynasonos model is incorporated along with the Fowler Nordheim (F-N) and Poole-Frankel models for charge trapping and de-trapping. The device parameters used for the simulation are illustrated in Table 1. The device is simulated with a gate length (L_g) of 100 nm and film thickness (T_{Si}) of 10 nm to deplete the carriers from the device underneath the gate. In this study, the front gate (Gate 1) is used as a charge trapping memory consisting of an oxide/nitride/oxide layer. The charge trapping memory is used to observe the long-term potentiation (LTP) and depression (LTD). Back gate (Gate 2) helps to store the holes in the floating body and shows the short-term potentiation (STP) behavior and transformation from STP to LTP.

Table 1. Device specification for synapse.

Device Parameters	Values
Gate Length (L_g)	100 nm
Silicon Channel thickness (T_{Si})	10 nm
Tunneling Oxide thickness (T_{tox})	2 nm (SiO_2)
Nitride Layer thickness (T_{nox})	2 nm (Si_3N_4)
Blocking Oxide thickness (T_{box})	2 nm (SiO_2)
Oxide thickness (T_{ox})	2 nm (SiO_2)
Gate 1 workfunction	*p*.poly silicon
Gate 2 workfunction	*p*.poly silicon
Channel doping (N_A)	10^{19} cm^{-3}
Source/Drain doping (N_D)	10^{20} cm^{-3}

III. Results And Discussion

A. Junctionless based as an artificial synapse

Fig. 2 demonstrates Atkinson's multi-store model of a human brain [22]. In the human brain, there are two types of memories: short-term (STM) and long-term (LTM). STM stores the information for a shorter period, while LTM stores the information for longer. Also, short-term memory can be transferred to long-term memory, depending on some events or rehearsing the event [22]. In this work, a junctionless transistor mimics these behaviors by applying a repetitive stimulus. Fig. 3(a) shows the transient analysis and trapped charges in the nitride layer during the potentiation operation of a JL based artificial synaptic device. For hardware neural networks with the CNN algorithm, we have simulated the device with 64 (6-bit) and 32 (5-bit) continuous pulses for both long term potentiation (LTP) long term depression (LTD). Transition from STP to LTP is achieved during potentiation operation. In order to observe these effects, we have applied a drain bias (V_{DS}) of 0.8 V, front gate voltage (V_{GS1}) of -1.0, and back gate voltage (V_{GS2}) of -0.5 V. The operation is performed with pulse time of 500 ns for potentiation and 100 ns for depression with gate voltage of 4 V and drain voltage of 0.8 V. Interval time for potentiation and depression is 500 ns.

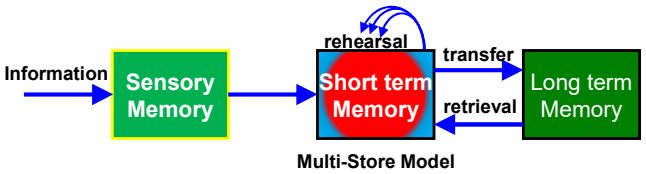

Fig. 2. Schematic of Atkinson's multi-store model of the human brain.

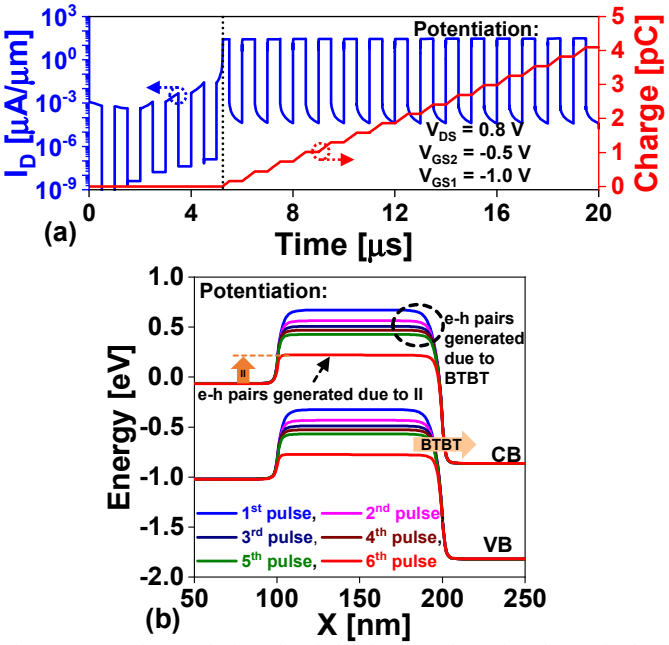

Fig. 3. (a) Transient analysis and variation in trapped positive charges in the nitride layer of the JL transistor during potentiation. The transition from STP to LTP is observed from the transient analysis. (b) Energy band diagram at different potentiation pulses (1st, 2nd, 3rd, 4th, 5th, and 6th). CB and VB indicate conduction and valence band, respectively.

During potentiation operation, at 1st pulse, the electron-hole (e-h) pairs are generated in the device due to the band-to-band tunneling (BTBT) mechanism as shown in the energy band diagram (in Fig. 3(b)). The energy band diagrams are extracted at 1 nm below the Gate 2 oxide. Further increase in repetitive pulse (2nd, 3rd, 4th, 5th pulse) generates more e-h pairs. The accumulated holes in the semiconductor film make a forward bias region between channel and source (the barrier between the source and channel is lower for the 6th pulse as shown in Fig. 3(b)), which triggers the floating body effect (impact ionization (II)) in the device. The generated holes due to II in the device near the drain junction gain sufficient energy to tunnel through tunneling oxide and start trapping into the nitride layer. Fig. 3(a) shows the same, from the 6th pulse, trapped charges are increasing linearly with an increase in pulse number, which shows that the device is in the LTP state, and a further increase in hole concentration in the device is due to floating body effect. These results confirm that the device can achieve LTP at a lower drain bias. The depression operation is performed with hot electron injection mechanism. Fig. 4 shows the contour plots of electron concentration in the silicon film and trapped electron concentration in the nitride layer. The contour plots at different pulses (1st, 11th, 21st, 31st, 41st, 61st) are extracted

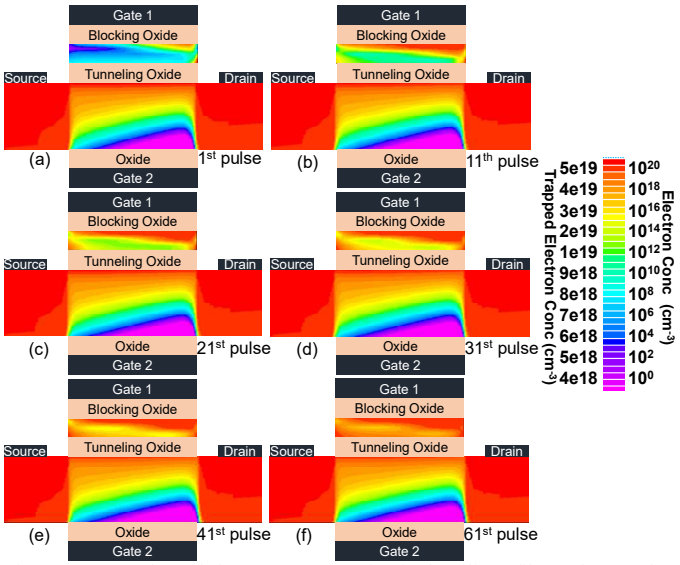

Fig. 4. Contour plots of electron concentration in the silicon film and trapped electron concentration in the nitride layer during depression operation at different pulses (a) 1st, (b) 11th, (c) 21st, (d) 31st, (e) 41st, and (f) 61st pulse. Conc represents concentration.

after performing the operations. The contour plots indicate that applying a positive bias at the gate and drain electrodes generates e-h pairs near the drain junction. The generated electrons tunnel from the tunneling oxide and start trapping into the nitride layer due to higher electron energy and positive bias at the gate electrode.

B. Device conductance, Non-linearity calculation, and Power consumption

For hardware neural networks, the device conductance value should be linear and low energy consumption during inference operation [5], [23]. Inference in a biological system is similar to the read operation of the memory. The conductance values of the synaptic device are estimated during the inference operation by applying a lower drain bias of 0.1 V to minimize the conductance disturbance. Fig. 5(a) and (b) show the normalized conductance values for 64 and 32 pulse operations, respectively. The non-linearity (NL) of the conductance value is estimated by normalizing the conductance value and pulse number as shown in Fig. 5(b). The MATLAB algorithm is used to calculate the NL of conductance value, as demonstrated in [5], [23]. It is evident from the figure that the JL device achieves almost linear conductance for LTP operation, while for LTD, NL is 2.7 for both 64 and 32 pulses operation. These results convey that the JL device is a promising candidate for an artificial synaptic to achieve linear conductance values. Table 2 demonstrates the power consumption during synaptic operations. The device achieves low power during synaptic operations, which results in low energy consumption.

C. Convolution Neural Network Algorithm with JL Device

In order to measure the learning abilities of the JL device, we conduct the software simulation based on JL device at a representative CNN, a brief variant of the VGG-NET [24]. The network structure is illustrated in Fig. 6. Our network consists of 6 convolutional layers for feature extraction and three fully connected layers for image classification. For every two convolutional layers, we adopt a max pooling layer behind to sub-sample and aggregate the feature. The model is evaluated in the training and testing on the CIFAR-10 dataset, an open-source dataset containing 10

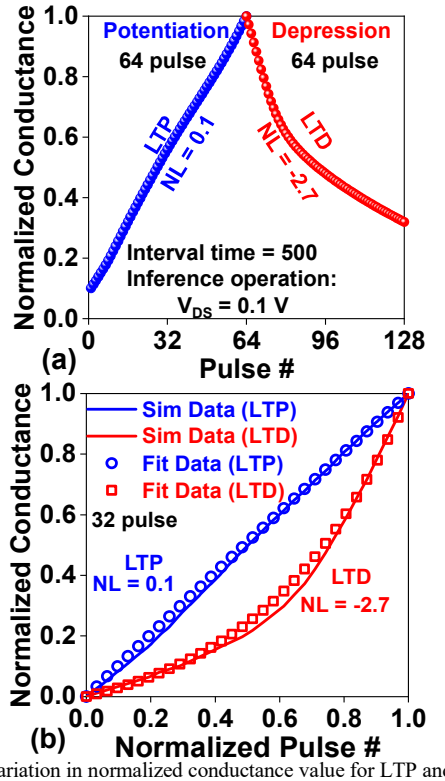

Fig. 5. (a) Variation in normalized conductance value for LTP and LTD with pulse number for 64 pulses. (b) Variation in normalized conductance with normalized pulse numbers for LTP and LTD with pulse numbers for 32 pulses.

Table 2. Power consumption during synaptic operations.

Synaptic operations			V_{DS}	Power (μW)
Potentiation	STP		0.8	5.15×10^{-4}
	LTP			20.8
Depression			0.8	400
Inference	Initial		0.1	0.2142
	LTP	1st pulse		0.5089
		64th pulse		5.695
	LTD	1st pulse		5.539
		64th pulse		1.683

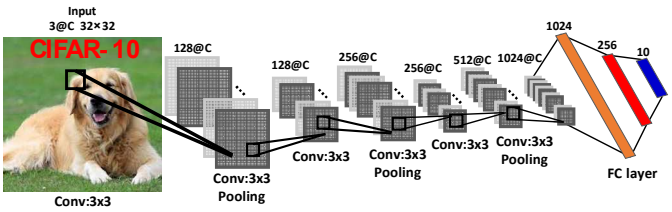

Fig. 6. The neural network structure for the CNN model (a variant of VGG-NET) consists of 6 convolutional layers (CONV), 3 max-pooling layers (Pooling), and 3 fully connected layers (FC).

class images of 32x32 size. To simulate with JL based synapse, the conductance values extracted from the JL device are exploited for every weight at convolutional layers and linear layers. JL device has shown suitable properties on both LTP and LTD with 128 and 64 states. We conceived the idea to compare 6bit and 5bit quantized parameter networks separately with our device synapse. The comparison results are shown in Fig. 7. During the training process, we conduct quantization aware training method on the model, which means scaling the weight into a low bit parameter or JL device conductance in the forward pass while remaining the same for the backward pass. However, the device conductance in a usual memristor array only

holds positive values, while the weight of the neural network at the algorithm level could be neither positive nor negative. We acquire the method mentioned in [23] and map our device conductance from 0∼1 to -1∼1. For each cell, -1 and 1 represent the minimum and maximum conductance of the device, respectively.

The simulation results show that a neural network with JL based synaptic device achieves comparable results with software-based CNN algorithm. The algorithm simulation results are demonstrated in Fig. 7. The full precision is to train on 32-bit floating-point (FP32) arithmetic by default, achieving the highest accuracy of 88.72% with CIFAR-10 dataset. During the quantization process on software, high accuracies of 86.9 3% and 86.63% are achieved for both 6 and 5-bits. When we adopt device conductance as the quantization scheme for 5 and 6 bits, it receives a slight decline of 1.89% and 1.78%, respectively, from the ideal 5 and 6-bit quantized accuracy. These results convey that the same NL conductance value achieves approximately the same accuracy for 64 and 32 pulses.

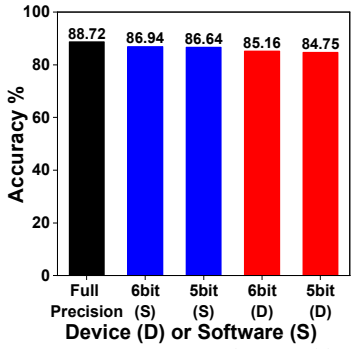

Fig. 7. Training accuracy on CIFAR-10 dataset with different weight precisions and device conductance. Bit (S): quantization with software simulation. Bit (D): quantization with device conductance.

IV. Conclusion

In this work, we have demonstrated the junctionless transistor as an artificial synapse with high and linear conductance values. Also, the transition from STP to LTP is achieved at the 6th pulse due to the generation of e-h pairs. We have used the device conductance values for the training/inference accuracy of a CNN with the CIFAR-10 dataset. Results demonstrate that accuracy depends on the linearity of conductance value rather than the total number of pulses applied for LTP and LTD. The junctionless-based device is CMOS compatible, operates at a lower drain bias, consumes low power, and achieves high image recognition accuracy. These advantages make the JL device a promising candidate for implementing a hardware neural network.

V. References

[1] D. V. Christensen *et al.*, "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Comput. Eng.*, vol. 2, no. 1, pp. 0–31, Jan. 2022.

[2] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990.

[3] N. K. Upadhyay, H. Jiang, Z. Wang, S. Asapu, Q. Xia, and J. Joshua Yang, "Emerging Memory Devices for Neuromorphic Computing," *Adv. Mater. Technol.*, vol. 4, no. 4, pp. 1–13, 2019.

[4] D. Marković, A. Mizrahi, D. Querlioz, and J. Grollier, "Physics for neuromorphic computing," *Nat. Rev. Phys.*, vol. 2, no. 9, pp. 499–510, 2020.

[5] S. Yu, "Neuro-Inspired Computing with Emerging Nonvolatile Memorys," *Proc. IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[6] V. M. Ho, J. A. Lee, and K. C. Martin, "The cell biology of synaptic plasticity," *Science (80-.).*, vol. 334, no. 6056, pp. 623–628, 2011.

[7] D. Kuzum, S. Yu, and H. S. Philip Wong, "Synaptic electronics: Materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, 2013.

[8] D. Ielmini, "Brain-inspired computing with resistive switching memory (RRAM): Devices, synapses and neural networks," *Microelectron. Eng.*, vol. 190, pp. 44–53, 2018.

[9] M. H. R. Ansari, S. Cho, J.-H. Lee, and B.-G. Park, "Core-Shell Dual-Gate Nanowire Memory as a Synaptic Device for Neuromorphic Application," *IEEE J. Electron Devices Soc.*, vol. 9, no. June, pp. 1282–1289, 2021.

[10] M.-K. Kim and J.-S. Lee, "Ferroelectric Analog Synaptic Transistors," *Nano Lett.*, vol. 19, no. 3, pp. 2044–2050, 2019.

[11] Yu, Cho, and Park, "A Silicon-Compatible Synaptic Transistor Capable of Multiple Synaptic Weights toward Energy-Efficient Neuromorphic Systems," *Electronics*, vol. 8, no. 10, p. 1102, Sep. 2019.

[12] S. Gundapaneni, M. Bajaj, R. K. Pandey, K. V. R. M. Murali, S. Ganguly, and A. Kottantharayil, "Effect of band-to-band tunneling on junctionless transistors," *IEEE Trans. Electron Devices*, vol. 59, no. 4, pp. 1023–1029, 2012.

[13] C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, and J.-P. Colinge, "Junctionless multigate field-effect transistor," *Appl. Phys. Lett.*, vol. 94, no. 5, p. 053511, Feb. 2009.

[14] S. J. Choi *et al.*, "Nonvolatile memory by all-around-gate junctionless transistor composed of silicon nanowire on bulk substrate," *IEEE Electron Device Lett.*, vol. 32, no. 5, pp. 602–604, 2011.

[15] J. P. Colinge *et al.*, "Junctionless Transistors: Physics and Properties," no. January, pp. 187–200, 2011.

[16] S. Sahay and M. J. Kumar, "Insight into Lateral Band-to-Band-Tunneling in Nanowire Junctionless FETs," *IEEE Trans. Electron Devices*, vol. 63, no. 10, pp. 4138–4142, 2016.

[17] D.-Y. Jeon, S. J. Park, M. Mouis, S. Barraud, G.-T. Kim, and G. Ghibaudo, "New method for the extraction of bulk channel mobility and flat-band voltage in junctionless transistors," *Solid. State. Electron.*, vol. 89, pp. 139–141, Nov. 2013.

[18] M. Vinet *et al.*, "Bonded planar double-metal-gate NMOS transistors down to 10 nm," *IEEE Electron Device Lett.*, vol. 26, no. 5, pp. 317–319, May 2005.

[19] R. Yu *et al.*, "Impact ionization induced dynamic floating body effect in junctionless transistors," *Solid. State. Electron.*, vol. 90, pp. 28–33, Dec. 2013.

[20] M. H. Raza Ansari, D. Kim, S. Cho, J. H. Lee, and B. G. Park, "Core-Shell Dual-Gate Nanowire Synaptic Transistor with Short/Long-Term Plasticity," *2021 5th IEEE Electron Devices Technol. Manuf. Conf. EDTM 2021*, vol. 2, pp. 4–6, 2021.

[21] Atlas User's Manual, Silvaco Int., Santa Clara, CA, USA, 2015

[22] R. C. Atkinson and R. M. Shiffrin, "Human Memory: A Proposed System and its Control Processes," in *Journal of Physics A: Mathematical and Theoretical*, vol. 44, no. 8, 1968, pp. 89–195.

[23] 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," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 40, no. 11, pp. 2306–2319, Nov. 2021.

[24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, Sep. 2014.