ADVANCED ELECTRONIC MATERIALS Open Access
www.advelectronicmat.de

Quantized Neural Network via Synaptic Segregation Based on Ternary Charge-Trap Transistors

Yongmin Baek, Byungjoon Bae, Jeongyong Yang, Doeon Lee, Hee Sung Lee, Minseong Park, Taegeon Kim, Sihwan Kim, Bo-In Park, Geonwook Yoo, and Kyusang Lee**

Artificial neural networks (ANNs) are widely used in numerous artificial intelligence-based applications. However, the significant amount of data transferred between computing units and storage has limited the widespread deployment of ANN for the artificial intelligence of things (AIoT) and power-constrained device applications. Therefore, among various ANN algorithms, quantized neural networks (QNNs) have garnered considerable attention because they require fewer computational resources with minimal energy consumption. Herein, an oxide-based ternary charge-trap transistor (CTT) that provides three discrete states and non-volatile memory characteristics are introduced, which are desirable for QNN computing. By employing a differential pair of ternary CTTs, an artificial synaptic segregation with multilevel quantized values for QNNs is demostrated. The approach establishes a platform that combines the advantages of multiple states and robustness to noise for in-memory computing to achieve reliable QNN performance in hardware, thereby facilitating the development of energy-efficient AIoT.

1. Introduction

Bio-inspired artificial neural networks (ANNs) have been widely used in various applications, such as image classification, voice recognition, and natural language processing, with high accuracy via the utilization of powerful computing units and large datasets.[1–3] However, their widespread application to the artificial intelligence of things (AIoT) platforms has been hindered by their limited computing power and the considerable amount of energy consumed for operating the algorithm. To implement ANNs on energy-constrained hardware, cloud computing was introduced by remotely processing a computing-intensive ANN algorithm using abundant energy, large data storage, and computing power. However, as the demand for real-time computing and analysis increases, data processing on AIoT devices via ANN algorithms using cloud servers is affected by a time lag in the data transfer between AIoT devices and cloud servers. Hence, energy-efficient neural engines are proposed to execute ANN algorithms at the edges of AIoT devices. Analog in-memory computing and quantized neural network (QNN) algorithms are promising methods for constructing dedicated neural engines that utilize fewer computing resources.[4–9]

Y. Baek, B. Bae, D. Lee, H. S. Lee, M. Park, T. Kim, S. Kim, K. Lee
Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22904, USA
E-mail: kl6ut@virginia.edu

J. Yang, G. Yoo
School of Electronic Engineering
Soongsil University
06938 Seoul, South Korea
E-mail: gwyoo@ssu.ac.kr

B.-I. Park
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

B.-I. Park
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

G. Yoo
Department of Intelligent Semiconductors
Soongsil University
Seoul 06938, South Korea

K. Lee
Department of Material Science and Engineering
University of Virginia
Charlottesville, VA 22904, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aelm.202300303

DOI: 10.1002/aelm.202300303

Analog in-memory computing utilizes non-volatile memory devices, such as memristors and charge-trap transistors (CTTs), to perform multiply-accumulate computing (MAC) based on Ohm's and Kirchhoff's circuit laws using the programmed analog states of information.[8,10,11] In-memory computing consumes minimal energy by processing data using memory devices at the edge of the system, which reduces data transfer between the memory and computing units.[10,11] However, the accuracy of computation is limited by thermal, shot, and external noise, which is an inherent disadvantage of the analog signal processing approach.[12,13] It normally necessitates a large number of iterations to precisely program and verify the weight value in the memory.[14,15]

Alternatively, QNNs that use only a few digits (< 4 bits) to represent data have been introduced to minimize the memory usage of ANNs, whereas ANN algorithms on conventional 32-bit computing systems based on von Neumann architecture use 32 digits for single data. Because the QNN algorithm utilizes fewer resources compared with the conventional 32-bit full-precision weights in ANNs, it enhances the energy efficiency for MAC operations.[5,6,12,13,16] However, the implementation of this energy-efficient QNN algorithm is restricted owing to the absence of dedicated hardware platforms.[6]

Herein, we introduce an oxide semiconductor-based ternary CTT that can store ternary information. A pair of CTTs can be utilized for winner-takes-all-based quantization using non-linear summation and multiplication, which is suitable for implementing the QNN algorithm. The fabricated ternary CTT exhibits the ternary transfer characteristic with a stable intermediate saturated current region between the ON and OFF current states. In addition, it shows non-volatile memory characteristics similar to those of a conventional CTT, such as threshold voltage shifting via the application of external voltage pulses for programming and erasing operations. Based on these ternary and non-volatile memory characteristics, we develop artificial synaptic segregation, which provides quinary balanced weights (−1, −1/2, 0, 1/2, and 1) by combining two ternary CTTs in parallel to perform in-memory computing. Finally, we design a QNN algorithm for image classification to introduce the quantized neural engine based on our ternary CTTs into machine learning tasks, which consists of five balanced weight values and ternary activation function. The designed QNN algorithm using ternary CTTs demonstrates an accuracy level comparable to that of the full-precision (32 bits) method for image classification using the modified National Institute of Standards and Technology (MNIST) and fashion-MNIST image datasets.[17,18]

2. Results

2.1. A Tri-Layer Oxide Transistor Inspired by Synaptic Function

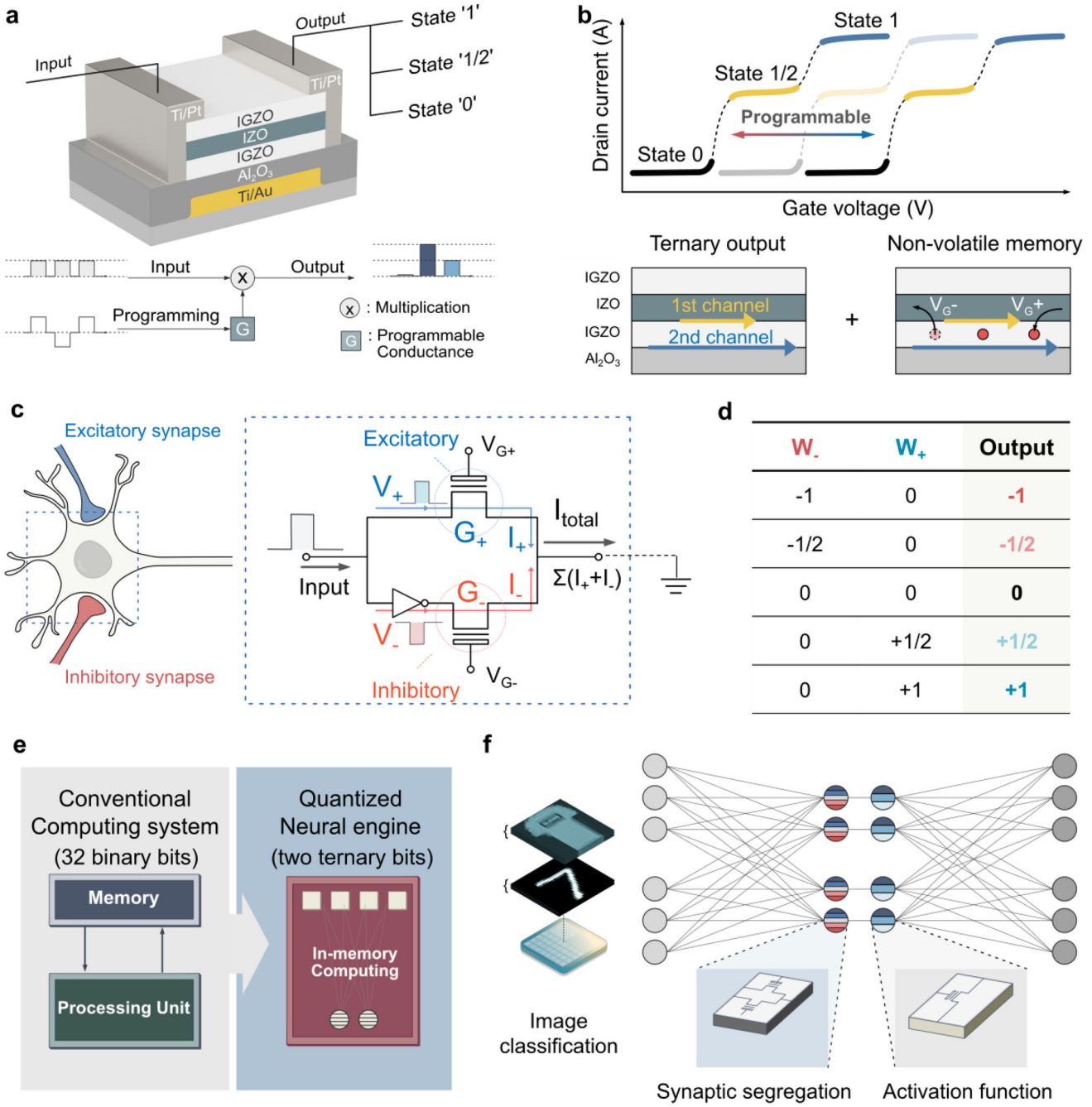
The widespread application of ANNs in power-constrained hardware fields is restricted because of their insufficient computing power, difficulty in managing significant amounts of data, and high energy consumption caused by data transfer between memory and computing units. Herein, we introduce an oxide-semiconductor-based ternary CTT that provides three discrete current levels and exhibits non-volatile memory characteristics. **Figure 1**a shows a schematic diagram of the fabricated ternary CTT featuring a three-layered oxide semiconductor stack. The channel conductance of a ternary CTT can be programmed to represent the three states. Subsequently, the multiplication of the input value by the programmed value generates three different output states (0, 1/2, and 1).

Figure 1b schematically illustrates the transfer characteristic of our ternary CTT that exhibits three distinct constant drain current levels as a function of the applied gate voltages. By applying programming voltages, the ternary transfer characteristic can be shifted in negative or positive directions along the gate voltage axis, showing programmability. This unique ternary transfer characteristic with non-volatile memory property is achieved by the presence of two separated conducting channels formed at the heterojunctions and charge trapping in the oxide semiconductor layer. Using the programmable non-volatile and ternary output characteristics, we demonstrated quantized synaptic plasticity, which was inspired by biological neurons.[6,13,19–21] Moreover, we showed synaptic segregation using a differential pair of ternary CTT (Figure 1c) that mimics the function of biological excitatory and inhibitory synapses. Each transistor generates a positive or negative output current depending on the polarity of its input voltage, similar to biological excitatory and inhibitory synapses that either enhance or suppress input signals. The differential pair circuit shares an output node that accumulates the current from each transistor, whereas each transistor employs separate input nodes. A positive input voltage is applied to the drain node of the excitatory transistor and a negative input voltage is applied to the source node of the inhibitory transistor.

Using a differential pair circuit, we employed the conductance difference between two ternary CTTs to realize five weight values, including positive and negative values, in the artificial synaptic segregation model. The application of voltage pulses to each gate node (V_{G+} and V_{G-}) allowed the channel conductance (G_+ and G_-) of each transistor to be tuned selectively. After tuning the channel conductance, we applied input voltages with identical amplitudes but opposite polarities to excitatory and inhibitory transistors using an inverter only connected to the inhibitory synapse. At each ternary CTT, the multiplication was performed using voltage inputs and channel conductance, resulting in current outputs. The total current flowing through the synapses was the sum of the positive and negative currents from each transistor, based on Kirchhoff's circuit law. Each transistor generated a computed output by multiplying the input voltage by the channel conductance, thus the differential pair of artificial synapses enabled multiplication and non-linear summation for the winner-take-all computing in synapses. This non-linear summation of the outputs from each ternary CTT is similar to the result of winner-take-all computing for generating five different outputs.[22,23]

To implement QNN computation using ternary CTTs, we utilized the channel conductance of the transistor as the weight of the QNN algorithm. We assigned three quantized weight values (W: 0, 1/2, and 1) to the channel conductance (G), which indicated three discrete states (OFF, intermediate, and ON). Consequently, the differential-pair circuit emulates the synaptic segregation with quantized weight values for QNNs, where the output signal is the product of the input data and the weight values, i.e.,

$$\text{Output} = I_+ + I_- = \text{Input} \times \left(W_+ + W_-\right) \quad (1)$$



W_-	W_+	Output
-1	0	-1
-1/2	0	-1/2
0	0	0
0	+1/2	+1/2
0	+1	+1

Figure 1. Structure of ternary charge-trap transistor, artificial synaptic segregation model, and activation function for quantized neural networks (QNN). a) Schematic illustration of ternary charge-trap transistor (CTT) with trilayer of oxide semiconductors, including programming, multiplication, and ternary output system diagrams. b) Schematic shows the programmable ternary transfer characteristic of the ternary CTT. The cross-section structure exhibits two separate conducting channels and charge-trap behaviors. c) Schematic geometry of the quantized output using synaptic segregation based on two ternary CTT pair (artificial inhibitory and exhibitory synapses) inspired by biological synapses, where each transistor is programmed by the application of gate voltage pulses. The positive and negative input voltages with the same amplitude induce the positive and negative currents to achieve five discrete states depending on the combinations of programmed channel conductance for each transistor. d) Synaptic segregation via two ternary CTT circuits, using combinations of G- and G+ to generate five quantized outputs (−1, −1/2, 0, +1/2, and +1). The utilized combinations are shown in the table. e) Illustration of quantized neural engine platform for in-memory computing via programmable ternary CTTs, compared with a conventional computing system. f) Schematic illustration of the QNNs-based neural network designed for image classification with the quantized output using the ternary CTT pair and ternary activation function.

Therefore, a segregated artificial synapse with the ternary CTTs can provide five different weight values (−1, −1/2, 0, 1/2, and 1) via the combination of W_+ (0, 1/2, and 1) and W_- (−1, −1/2, and 0). We utilized specific combinations of W+ and W- to produce five distinct outputs. The selected combinations (W_+, W_-, Output) are (−1, 0, −1), (−1/2, 0, −1/2), (0, 0, 0), (0, +1/2, +1/2), and (0, +1, +1), as illustrated in Figure 1d. The chosen combinations were intended to prevent the simultaneous activation of two ternary CTTs, such as (+1, −1), (+1, −1/2), (−1, +1), and (−1/2, +1), minimizing energy consumption during the quantization procedure. Additionally, Figure S1 (Supporting Information) shows the analog signal outputs obtained from all possible combinations using the segregated synapse.

Figure 1e shows a schematic illustration of the quantized neural engine platform based on our ternary CTT, which utilizes fewer memory cells (two ternary bits) to store the weight value compared with a conventional computing architecture (32 binary bits). Combining the advantages of multibit and in-memory computing characteristics, the ternary CTT allows us to achieve a QNN platform that can ameliorate the data transportation bottleneck issues of the von Neumann architecture occurring between the computing units and memory devices.[24]

To demonstrate the effectiveness of the neural engine platform, we designed a QNN model based on the characteristics of the fabricated CTT and implemented image classification tasks (Figure 1f). In the QNN model, we formulated the quantized weight values (−1, −1/2, 0, 1/2, and 1) and activation function (0, 1/2, and 1) based on the differential pair using two ternary CTTs and the single ternary CTT, respectively. We employed the MNIST and Fashion-MNIST datasets to evaluate the developed QNN model constructed using the characteristics of ternary CTTs, where the result shows that the image classification performance of the developed model was comparable to that of the conventional ANN algorithm.

2.2. Electrical Characteristics of the Ternary Transistor

Non-volatile memories storing multiple bits have been developed to enhance memory density and reduce the standby power of data-intensive applications such as machine learning. Recently, memristors and CTTs have been proposed as promising candidates for analog and neuromorphic computing applications owing to their analog synaptic behaviors. However, analog memory devices are inherently vulnerable to noise because of their indefinite states for storing data, which slightly perturbs the reading or programming results.[25] In this study, we demonstrated an oxide-semiconductor-based ternary CTT providing three discrete states with a large programming margin in the intermediate state. We fabricated a ternary CTT by stacking three layers of oxide semiconductors: Indium gallium zinc oxide (IGZO), indium zinc oxide (IZO), and IGZO. **Figure 2**a shows a schematic illustration of the fabricated transistor with a bottom-gate structure. A titanium/gold (Ti/Au) (5/50 nm) layer was used as the gate metal, and an aluminum oxide (Al_2O_3) (100 nm) dielectric layer was employed as the gate-insulating layer. The oxide semiconducting channel layers were deposited using a direct current (DC)/radio frequency (RF) magnetron system under argon (Ar) and oxygen (O_2) mixed atmosphere. The active area was determined by etching the IGZO/IZO/IGZO layers. Source and drain contacts were deposited using a Ti/platinum (Ti/Pt) (5/35 nm) layer with a channel width-to-length ratio of 20. Intermediate and ON-saturated conducting channels were observed at the interfaces of the Al_2O_3/IGZO and IGZO/IZO layers, respectively.

The stacked oxide semiconductors with sharp transitions between each oxide layer were confirmed by high-resolution transmission electron microscopy (HRTEM) (see Figure 2b). The hetero-interfaces between the oxide semiconductor layers allow charge carriers to be accumulated. Figure 2c shows a plane-view scanning electron microscopy (SEM) image of the ternary CTT with interdigitated source and drain electrodes and a mesa-defined active channel area. Figure 2d shows the transfer characteristic of a ternary CTT at the fixed drain voltage of 1 V while the gate voltage was swept from −15 to 15 V, and vice versa. The transfer curve was segmented into three distinct regions (labeled Regions I, II, and III) corresponding to the OFF, intermediate, and ON states of the device, respectively.

Region I corresponds to the OFF state when the gate voltage is applied below the threshold voltage (V_{th} = −5 V). The OFF current of ternary CTT was $\approx 10^{-13}$ A. As the gate voltage increased over the threshold voltage, the drain current showed a clear transition from the OFF to the intermediate states (i.e., Region II). In Region II, the drain current remained at $\approx 10^{-7}$ A when the gate voltage was between −3 and 3 V. This behavior is due to the potential barrier at the interface between the IZO and IGZO layers, which hinders carrier injection from the IZO layer to the bottom IGZO layer. As the applied gate voltage increased above V_{th} of −5 V, charge carriers accumulated at this potential barrier and formed the first conducting channel at the interface between the IZO and IGZO layers (Figure S2a,b, Supporting Information). However, the drain current was limited to $\approx 10^{-7}$ A owing to the relatively low density of the accumulated charge carriers at the potential barrier in this intermediate state. Region III corresponds to the ON state of the ternary CTT. As the gate voltage increased over 3 V, carrier injection to the IGZO/Al_2O_3 interface was induced by overcoming the potential barrier at the IZO/bottom IGZO interface. Furthermore, the increased gate voltage induced energy bands bending, lowering the conduction band of the bottom IGZO layer below the Fermi energy level. As a result, a conducting channel with a high carrier density is formed at the IGZO/Al_2O_3 interface (Figure S2c, Supporting Information). Thus, the ternary CTT was fully turned on and reached a saturation current level of $\approx 10^{-3}$ A.

Figure 2d confirms that the intermediate state of the transfer curve is maintained without hysteresis during the DC forward (black markers) and reverse sweep (red markers) measurement for gate voltages ranging from −15 to 15 V. Non-hysteresis behavior is also confirmed by using pulse *I*–*V* measurements. Figure S3 (Supporting Information) shows the negligible hysteresis of ternary CTT when the voltage pulses with a 500 µs width were applied to the gate electrode. This unique non-hysteresis characteristic distinguishes our multichannel structure from the hump-like behavior caused by defects in oxide semiconductors that exhibit large hysteresis during forward and reverse measurements.[26–28] The non-hysteresis characteristic allows us to avoid unintentional programming due to noise in the gate bias.

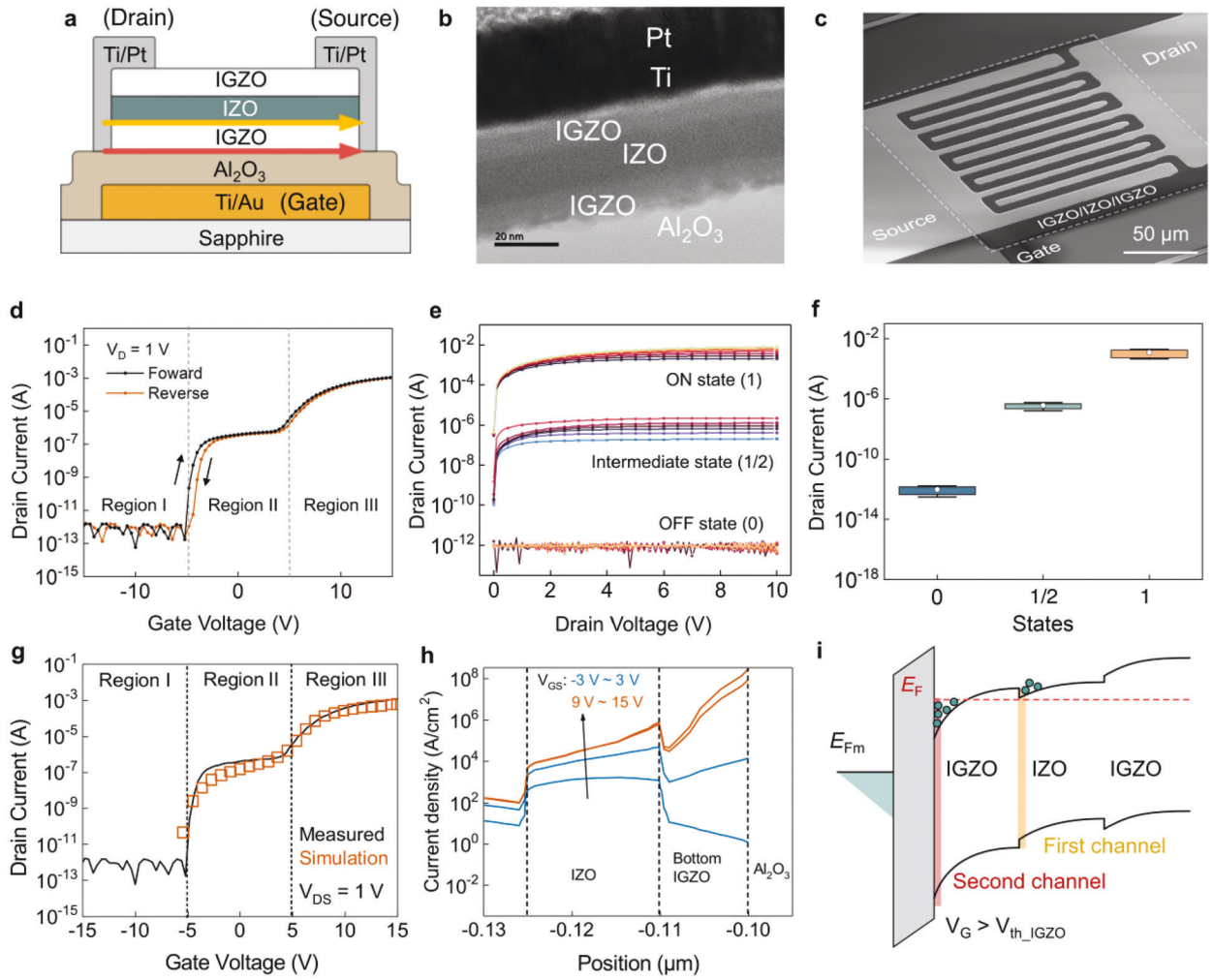


Figure 2. Ternary characteristics of the indium gallium zinc oxide(IGZO)/indium zinc oxide (IZO)/IGZO transistor. a) Schematic illustration of the fabricated bottom-gate transistor comprising three layers of oxide semiconductor which are IGZO, IZO, and IGZO with the two conducting channels. b) Cross-sectional transmission electron microscopic (TEM) images of trilayer oxide semiconductor transistors where the IGZO/IZO/IGZO three oxide semiconductor layers are deposited on the aluminum oxide (Al_2O_3) gate dielectric. c) Scanning electron microscope (SEM) image of the fabricated transistor having the interdigitated source/drain structure. d) Transfer characteristic of the ternary CTT in semi-log scale at V_D = 1 V by forward and reverse sweeps of the gate voltage from -15 V to 15 V. The gate voltage ranges are divided into three regions corresponding to the OFF, intermediate, and ON currents, where Region I, II, and III are labeled with guidelines, respectively. e) Output characteristic of the ternary CTT in a semi-log scale. The drain current is measured by applying the drain voltage from 0 V to 10 V with gate voltage applied in three different ranges: −15 V– −8 V, −3 V – 3 V, and 8 V–15 V. f) Distribution of the drain currents in each state, where box shows the data between 25 and 75%. The mean value and standard deviation of the data are shown as open circles and error bars. g) Transfer characteristic of the ternary CTT with a simulation model derived from technology computer-aided design (TCAD) simulation. h) TCAD simulation results show current density distributions of the ternary CTT with the gate voltage range of V_{GS} = −3 V–15 V. i) Schematic band diagram of the IGZO/IZO/IZO layer under a positive gate bias, showing the first and second channel formation at the interfaces.

Figure 2e shows an output characteristic of the ternary CTT over V_D ranging from 0 to 10 V for various gate voltage biases (V_G) from −15 to 15 V. The drain currents saturated and stabilized at ≈10^{-10}, 10^{-7}, and 10^{-3} A in the gate voltage ranging from −15 V to −8 V, −3 V to 3 V, and 8 V to 15 V (labeled as the OFF, intermediate, and ON states), respectively. The clustering of the output curves into three groups confirms that the fabricated ternary CTT comprises three saturation current levels (OFF, intermediate, and ON). Due to the absence of holes in IGZO and IZO materials, our ternary CTT maintains a low off current despite an increase in drain voltages. Additionally, Figure S4 (Supporting Information) shows the normalized drain currents based on the transfer and output characteristics, which were partitioned by the channel width for comparison. The stochastic data distribution of the measured drain currents for each state is shown in Figure 2f after 100 cycles of sweeping the gate voltage from −15 to 15 V. The box indicates the interquartile range comprising 50% of the data, and the error bars represent the standard deviation of each state. The mean and standard deviation (μ, σ) of the OFF, intermediate, and ON states were (3.70×10^{-7} A, 2.07×10^{-7} A), (9.79×10^{-13} A, 6.72×10^{-13} A), and (1.21×10^{-3} A, 7.59×10^{-4} A), respectively. This small standard deviation indicates

that the multichannel in the tri-layer oxide semiconductor is stable over a gate voltage ranging from −15 to 15 V. In addition, we investigated the reliability of the ternary CTT measured at fixed gate voltages of −10, 0, and +10 V, which correspond to the OFF, intermediate, and ON states, respectively (Figure S5, Supporting Information). In this measurement, the coefficients of variation for each state (OFF, intermediate, and ON) were 48.5%, 4.8%, and 2.0% after 100 cycles of measurement, respectively which confirms the consistent performance across multiple measurements.

Moreover, we performed physics-based technology computer-aided design (TCAD) simulation to understand the operation mechanism of our ternary CTT (Figure 2g). Figure 2h shows the current densities for the intermediate and ON states depending on various gate voltages. In the intermediate state, a current density of $\approx 10^5$ A cm^{-2} was observed at the IZO/bottom IGZO interface due to the accumulation of carriers at the potential barrier. Meanwhile, the current density at the interface of IGZO/Al_2O_3 is relatively low ($\approx 10^4$ A cm^{-2}) under a gate voltage of 3 V. It is explicitly shown that the first conductive channel was formed at the interface of IZO/bottom IGZO rather than that of IGZO/Al_2O_3.

As the gate voltage increased up to 15 V, the current density at the interface of the IGZO/Al_2O_3 significantly increased, reaching $\approx 10^8$ A cm^{-2}. This high current density is due to the carrier injection from the IZO layer and the conduction band bending. However, the current density at the interface of the IZO/bottom IGZO remained similar level of $\approx 10^6$ A cm^{-2} due to the small potential energy barrier. Therefore, the ternary characteristic of our device was the result of having two conducting channels; the first channel is formed due to the injected carriers and the potential barrier at the IZO/bottom IGZO, and the second channel is formed by the carrier injections from the IZO layer to the IGZO/Al_2O_3 interface and the band bending (Figure 2i). Further information about the TCAD simulation, including the material parameters of the top IGZO, IZO, and bottom IGZO layers used, are described in Note S1 (Supporting Information).

2.3. Non-Volatile Memory Characteristics of Ternary Charge-Trap Transistors (CTTs)

A key feature of our ternary CTT is that ternary behavior is embedded in the non-volatile memory characteristics. A schematic illustration of the charge-trapping mechanism in the ternary CTT is presented in **Figure 3**a, where the electrons are trapped in the deep-trap states of the bottom IGZO layer. The formation of a conducting channel at the IZO/bottom IGZO interface induces charge transport from the channel to the IGZO bulk traps.[29] Furthermore, the higher electron affinity of the IZO layer creates a potential-well structure between the bottom and top IGZO layers.[30,31] Consequently, the electrons in the channel are located at the potential barrier between the IZO and bottom IGZO layers. When positive voltage pulses are applied to the gate node, the electrons in the first conducting channel are easily transported into the deep trap states of the bottom IGZO channel owing to the potential-well structure (left, Figure 3a). By contrast, when negative voltage pulses are applied to the gate node, the electrons are trapped in the deep states of the IGZO layer. (right, Figure 3a). These trapped charges in the IGZO layer screen the electric field from the gate, thereby shifting the threshold voltage of the ternary CTT.[32] In addition, the potential well structure in the IGZO/IZO/IGZO layer stack enhances the retention of the ternary CTT. Although the charges trapped in the IGZO layer can escape from the traps when a negative voltage is applied, however, they encounter another potential barrier at the interface between the IZO and top IGZO layers. Consequently, the structure of the potential wells in the IGZO/IZO/IGZO layer stack enables the transistor to function as a nonvolatile memory.

Figure 3b shows a schematic illustration of voltage pulses applied to the gate and drain nodes during the programming and read operations, along with the equivalent circuit diagrams for each operation. For programming, positive or negative voltage pulses were applied to the gate node, while the source and drain nodes were grounded. The threshold voltage of the ternary CTT was controlled by adjusting the amplitude, width, and number of voltage pulses applied to the gate. As described above, positive or negative gate voltage pulses introduce or release charges in the bottom IGZO layer, thus shifting the threshold voltage of the ternary CTT. The shift in the threshold voltage results in a change in the drain current at a fixed drain voltage without gate bias. Therefore, we applied a read voltage (+1 V) to the drain node to measure the drain current, while the gate voltage was grounded (0 V). Subsequently, the measured drain current was converted to the conductance of the channel, which is defined as the programmable value of the ternary CTT.

To confirm the shift in the threshold voltage via pulse modulation, we investigated the transfer characteristics after programming. Figure 3c and d show the transfer curves under forward and reverse sweeps of the ternary CTT after the application of programming pulses to the gate electrode, respectively; the gate voltage was swept from −10 to +10 V at a fixed drain voltage of 1 V, and vice versa. Figure 3c shows the shifted transfer curves after the programming process via pulse modulation. Pulse modulation was conducted by gradually increasing voltage pulses, where the amplitudes ranged from 15 to 35 V in increments of 5 V with a fixed pulse width of 500 μs. A higher amplitude of pulse induced a more significant threshold voltage shift (green arrow in Figure 3c), indicating that more charges were trapped in the bottom IGZO layer.

The initial threshold voltage of ternary CTTs was -5 V. After the application of the voltage pulses with amplitudes from 15 to 35 V at intervals of 5 V, the threshold voltage shifted to −3.5, 0, 3.5, 6, and 10 V from −5 V. Meanwhile, the application of negative programming voltage pulses resulted in a negative shift of the threshold voltage. Figure 3d shows that the transfer curves shifted to the negative direction on the gate voltage axis (red arrow). When the negative gate voltage pulses were applied with various amplitudes from −15 to −35 V with a fixed pulse width of 500 μs, the threshold voltages were shifted to 0.5, 0.2, 0, −0.1, and −0.3 V from 1 V. Shifting the threshold voltage further toward the negative value requires a greater number and/or higher amplitude of pulses due to the absence of holes in the oxide semiconductor layer.[33–35] We conducted further investigations on negative voltage pulse modulation with different initial states of the ternary CTT, as the negative voltage pulses caused a relatively smaller shift in the transfer curve compared to the positive voltage pulses. The shift of transfer curves can be observed for any initial point on the transfer curves, without encountering any

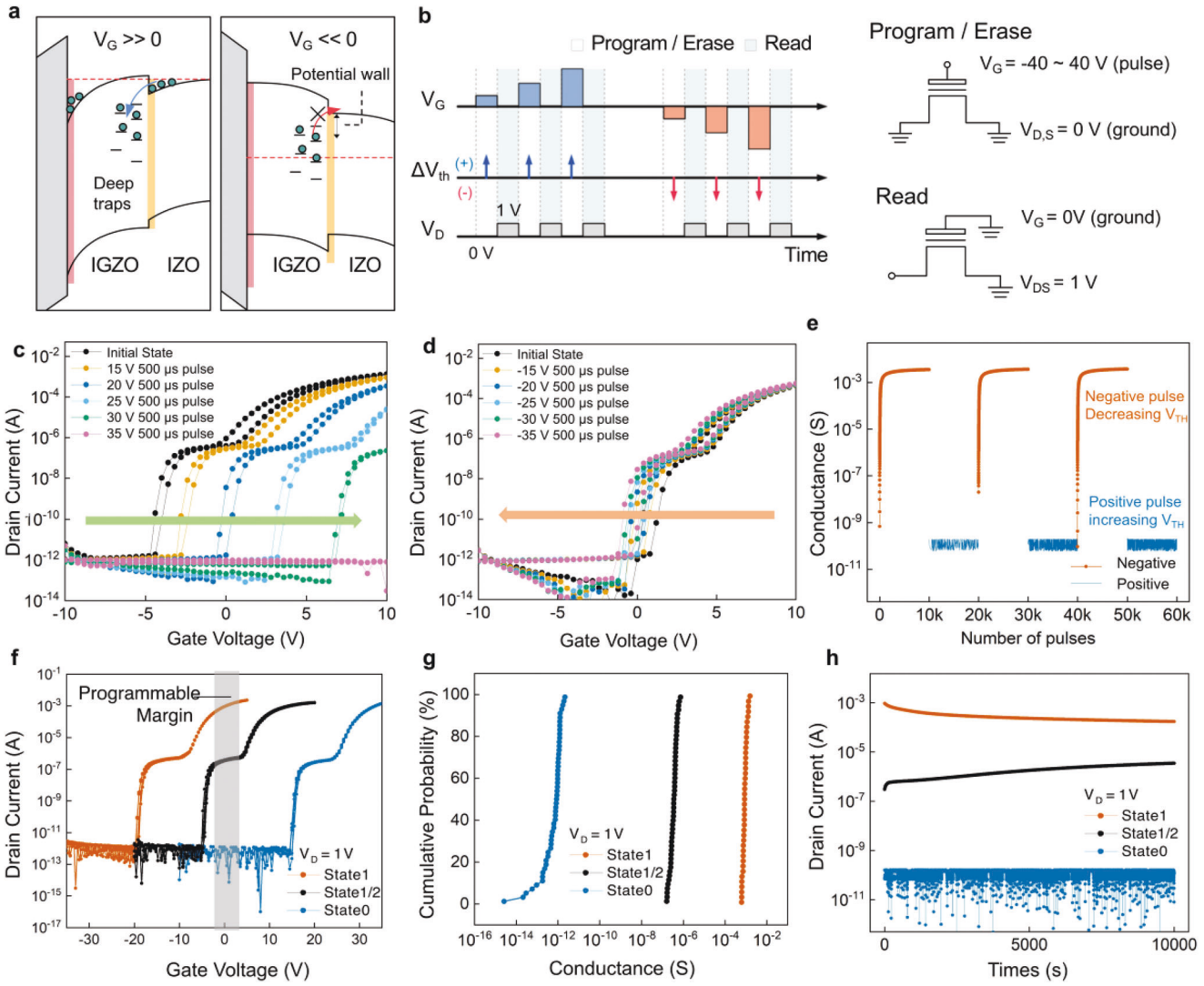


Figure 3. Non-volatile memory characteristics of the IGZO/IZO/IGZO transistor. a) Schematic illustration of charge trapping in the IGZO layer by application of the gate voltage pulses. b) Diagram illustrates the pulse trains at the gate and drain electrodes during the program/erase and read operations (left) and their respective circuit (right). The bar heights on V_G pulses represent the gradual increase in amplitude of input pulse during the process of programming and erasing. The arrow on V_{th} indicates the shift direction for V_{th}. c,d) Shifted transfer curves resulting from positive (c) and negative (d) voltage pulses to the gate node, respectively, where the amplitudes of the applied pulses are changed from ± 15 V to ± 35 V by ± 5 V. The shifted directions of the transfer curves are represented by positive (green) and negative (red) arrows, respectively. e) Programmable channel conductance measured at $V_G = 0$ V and $V_D = 1$ V as a function of the number of voltage pulses applied to the gate, showing the maximum and minimum conductance achievable via pulse modulation. $V_G = \pm 40$ V, depicted by the red and blue symbols, is applied for a complete erase and program. f) Transfer characteristics of programmed ternary CTTs with three distinct channel conductance states: state 0, 1/2, and 1, where the drain voltage was fixed at 1 V during the gate voltage sweeps. g) Cumulative probability of the drain current levels in each state under the repeated program and measurement of 50 cycles. h) Retention characteristic of the ternary CTT showing for each programmed state (Duration: 10000 s).

performance degradation or potential charge traps (Figure S6, Supporting Information). The ternary CTT maintained a nearly identical subthreshold swing after pulse modulation with varying gate voltage amplitudes that indicates the negligible density of deep-level trap sites in the film. The ternary characteristic of our ternary CTT remained unchanged even under extreme programming conditions (± 40 V) (Figure S7, Supporting Information). Although here we demonstrated threshold voltage shifting by employing high-gate voltage pulses, the required amplitude for the gate voltage pulse can be significantly reduced by employing a thin gate oxide (described in Note S2, Supporting Information).

To adopt the ternary CTT for neuromorphic computing applications, we investigated the range of programmable values. Figure 3e shows the measured channel conductance of the ternary CTT as a function of the number of voltage pulses applied to the gate nodes. To examine the maximum and minimum conductance range of the channel, positive (40 V) and negative (−40 V) programming voltage pulses with a pulse width of 500 µs were applied to the gate electrode (blue and red symbols,

respectively). After each programming pulse, a drain read-voltage pulse with a fixed amplitude (+1 V) and width (500 μs) was applied. The programmable channel conductance range of the ternary CTT was between 10^{-10} and 10^{-3} S. The channel conductance reached the fully turned-on state when ≈100 negative pulses (−40 V, 500 μs) were applied by shifting the threshold voltage to the negative side. By contrast, a single positive pulse (+40 V, 500 μs) reduced the channel conductance to the fully turned-off level by shifting the threshold voltage to the positive side. This programming behavior is shown in Figure 3c,d. The programming procedure and time on hardware can be minimized by employing a weight-transfer method that computes the weight values of the neural networks outside the hardware.[36,37]

Furthermore, we confirmed that the ternary characteristics of the fabricated ternary CTT were preserved during pulse modulation. Figure 3f shows that the pulse-modulated transistor comprises three distinct drain current levels at $V_G = 0$ V. Each programmed transfer curve (blue, black, and red symbols) exhibited threshold voltages of 15, −6, and -20 V, respectively. Unlike conventional CTTs, which feature a narrow intermediate programming margin as a multibit memory device, our ternary CTT exhibits a wide programming range from −3 to 3 V, as highlighted in the black shaded area in Figure 3f. This wide range of intermediate states allows us to minimize the repetitive programming and verification processes to achieve a target weight value after programming.

Figure 3g illustrates the cumulative probability distribution of the programmed channel conductance at $V_G = 0$ after 50 cycles of erasing and programming operations using pulse modulation. The transition from state 0 to +1 involved applying 100 negative pulses (−35 V, 500 μs), while the transitions from state +1 to state ½ and state 0 were accomplished by applying a single positive pulse (+15 V and +35 V, 500 μs), respectively. This pulse modulation demonstrates the effective distribution of channel conductance values achieved through iterative erasing and programming operations, even in the absence of comprehensive verification and high-precision programming. We also evaluated the device-to-device variation by fixing the drain voltage at 1 V and measuring the corresponding drain current from 28 devices (Figure S8, Supporting Information). The results reveal minimal device-to-device variation in the intermediate state, considering that the off and on currents are ≈10^{-12} and 10^{-4} A, respectively.

Additionally, we demonstrated that the ternary system can provide further well-distributed states, further suppressing possible interference between the states. We employed high-precision programming with verification to achieve accurate channel conductance for each state after programming, resulting in a smaller coefficient of variation for each state, as shown in Figure S9 (Supporting Information). This observation highlights the robustness, reliability, and capability of our ternary CTT in achieving the well-distributed three channel-conductance values.

We also examined the reliability of the intermediate region of the ternary CTT by implementing repetitive programming and erasing processes. The intermediate region was preserved after 1000 cycles of programming and erasing. However, its range was reduced from 8.5 to 2.2 V due to the charge trapping at deep-level trap sites in the IGZO layer, which can be prevented by applying a low operation voltage (Figure S10, Supporting Information). Despite the decrease in the voltage range of the intermediate state, the OFF and ON state regions were expanded, resulting in a 3.16 V expansion for each state. This expansion was attributed to the unchanged subthreshold swings even after 1000 cycles, ensuring that the OFF and ON states were not degraded. Therefore, the functionality and performance of the OFF and ON states remain intact despite the change in voltage range in the intermediate region.

Figure 3h presents the data retention characteristics of the ternary CTT obtained by measuring the channel conductance as a function of time. After programming the transistor, the channel conductance of each state was measured for 10 000 s under the application of 1 V drain bias voltages without applying the gate bias voltage. A conventional CTT typically exhibits different retention times for the intermediate and ON states, which may result in the failure of appropriate programming and read operations in non-volatile memory devices.[29,34] Unlike conventional ternary transistors, the intermediate state of our ternary CTT occurred in a wide constant-current region, which resulted in retention characteristics similar to those of the ON state.

Moreover, we estimated the energy consumption of our ternary CTT during the programming (see Note S3 for details, Supporting Information). We show that a single voltage pulse of 40 V with a 500 μs pulse width can change the state from +1 to 0, whereas 100 pulses of −40 V with a 500 μs width were required to update the state from 0 to +1. Thus, to completely update the states from +1 to 0 and 0 to +1 and only 3.79 and 775 pJ of electrical energy were required, respectively. This is considerably less energy compared to an average of 20 μJ consumed for writing and erasing in a commercial silicon-oxide-nitride-oxide-silicon (SONOS) transistor.[38]

2.4. In-Memory Computing and Software Simulation of Quantized Neural Network (QNN)

The implementation of ANN algorithms for power-constrained hardware, such as AIoT devices, has been limited by requirements such as large storage, powerful computing units, and sufficient energy for significant amounts of data transfers between memory and computing units.[4,12,39,40] Hence, QNN algorithms that utilize fewer bits to represent data have been introduced, thereby reducing the number of accesses to memory cells. Herein, we present a neural engine platform designed for energy-efficient QNN computing; it comprises the artificial segregated synapse and the ternary activation function that provide five and three quantized values, respectively, as shown in **Figure 4**a. Furthermore, we operated the QNN algorithm based on the characteristics of our ternary CTT to identify hardware-based simplified ANNs that can be adopted for various low-power device applications.

Because neural network computing typically employs both positive and negative weights, we first designed a current subtraction circuit with a differential pair of ternary CTTs to demonstrate artificial synaptic segregation that amplifies or suppresses the input signal. The designed artificial synaptic segregation structure comprises excitatory (W_+) and inhibitory (W_-) synapses that generate positive and negative current outputs, respectively, as shown in Figure 4b. The two ternary CTTs were programmed by individually applying pulse modulation via the

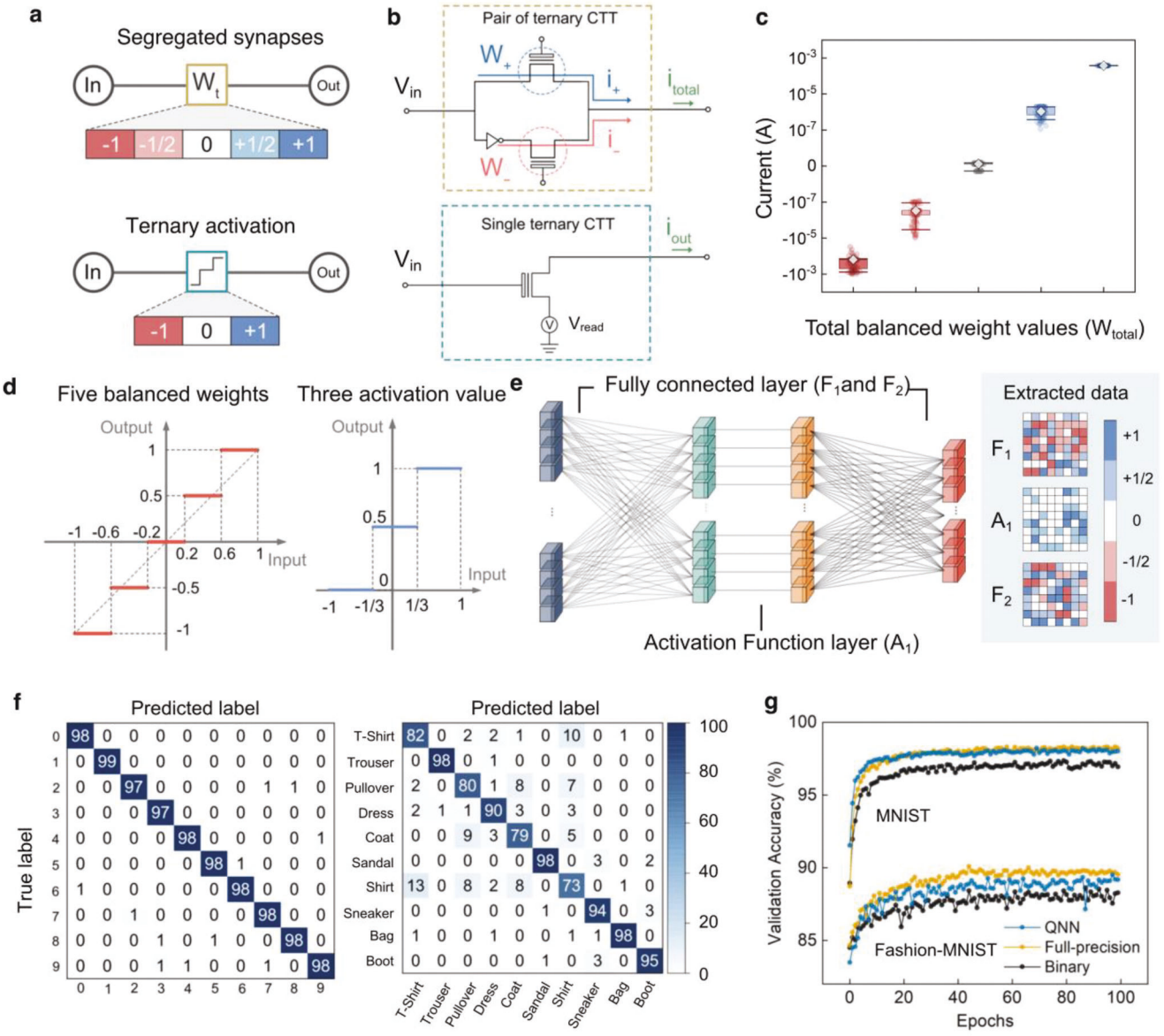


Figure 4. In-memory computing and software simulation of quantized neural network. a) Schematic diagram of the segregated synapses and ternary activation utilized in quantized neural network computing. The segregated synapses provide five quantized weight values while the ternary activation enables three quantized outputs. b) Circuit diagrams of segregated synapses and ternary activation, which consists of a pair of ternary CTT and a single ternary CTT, respectively. In a pair of ternary CTT, the currents flowing in opposite directions (i_+ and i_-) are accumulated at the shared output electrode (i_{total}). A single ternary CTT provides the output current (i_{out}) based on its ternary characteristic with the application of input (V_{in}) and reads voltage (V_{read}). c) Total output current measured at the common output node as a function of the total balanced weight values (W_{total}). The mean and standard deviation values of each W_{total} are depicted as diamond symbols and error bars, respectively. The box represents the interquartile range. d) Schematic diagram of quantized weights and activation function utilized in the software simulation based on the developed ternary CTT pair as the segregated synapses, and a single ternary CTT as the activation function. e) Composition of QNNs layers for image classification. Two fully-connected layers (F_1 and F_2) and one activation layer (A_1) are employed between the input and output layers (left). Extracted data from each layer (F_1, A1, and F2) show quantized weights and activation values after the training process, sampled as 8 by 8 images (right). f) Confusion matrix results of image classification for Modified National Institute of Standards and Technology (MNIST) and Fashion-MNIST image datasets. The highlighted diagonal values represent the percentage of correctly classified images. g) Validation accuracy of the binary neural networks (BNNs), the QNNs, and the full-precision deep neural networks (DNNs) models trained for MNIST and Fashion MNIST datasets. QNNs based on the ternary CTT show as high accuracy as DNNs provide in both MNIST and Fashion-MNIST classification after training Epoch of 100.

gate nodes, as described in the previous section. Subsequently, the positive and negative drain voltages with identical amplitudes (± 1 V) were applied to each transistor representing artificial excitatory and inhibitory synapses, separately. The applied drain voltage was multiplied by the weight values stored in each transistor as the channel conductance to compute the drain currents as the output using Ohm's law. Subsequently, the currents flowing through each transistor (i_+ and i_-) accumulated at the shared output node, based on Kirchhoff's law. At the shared output node, the output of each transistor accumulated to form the total output (i_{total}). In addition to the segregated synapse, we also demonstrated the ternary activation function using a single ternary CTT based on the ternary transfer characteristic.

Using the differential pair, we demonstrated the positive and negative weight values (−1, −1/2, 0, ½, and 1) for the QNN algorithm. Figure 4c shows the result of the winner-take-all computing in the segregated synapse using the fabricated two ternary CTTs. The accumulated total output current (i_{total}) from the segregated synapse is shown as a function of the total balanced weight (W_{total}). In the differential pair circuit, the polarity of the output current is determined by the dominant currents flowing through the excitatory or inhibitory synapses (W_+ and W). The total output current is positive when the current flowing through the excitatory synapse (i_+) is higher than that through the inhibitory synapse (i_-). By contrast, the total output current is negative when the current flowing through the inhibitory synapse (i_-) is higher than that through the excitatory synapse (i_+). Therefore, the output current (i_{total}) is determined by the higher state between W+ and W-. When only one of the two ternary CTTs is fully turned on while the other is turned off at $V_G = 0$ V, the output current (i_{total}) represents that W_{total} is either "+1" or "−1". When one of the two transistors produces intermediate currents and the other is turned off, the W_{total} is "+1/2" or "−1/2". Finally, the W_{total} of "0" can be achieved if both are turned off.

We repeated 100 cycles of the winner-take-all computing for various weight values using a differential pair of fabricated ternary CTTs. As shown in Figure 4c, the μ and σ of each weight value (W_{total}) from −1 to 1 are shown as the diamond symbol and error bar; the (μ, σ) values shown are (-3.28×10^{-4} A, 2.61×10^{-4} A), (-1.05×10^{-6}, 1.84×10^{-6} A), (6.24×10^{-10} A, 1.57×10^{-9} A), (1.26×10^{-6} A, 6.51×10^{-7} A), and (3.76×10^{-4} A, 5.18×10^{-7} A); the box indicates the interquartile range. The detailed descriptive statistics obtained from repeated winner-take-all computing are summarized in Table S1 (Supporting Information). The current levels were successfully quantized into five states that can be assigned weight values. The ratios of the mean values between weights (−1/2, −1), (0, −1/2), (1/2, 0), and (1, 1/2) were 3.12×10^2, 1.68×10^3, 2.02×10^3, and 2.98×10^2, respectively. The effects of standard deviations for each weight on ANN performance were negligible according to our simulation of image classification using the measured data. The standard deviations from the measured current values for each state were modeled in software simulation (Note S4, Supporting Information). Thus, the artificial synaptic segregation via a differential pair using two ternary CTTs allowed us to implement five quantized balanced weight values (−1, −1/2, 0, 1/2, and 1) for the QNN operation.

Using the aforementioned artificial synaptic segregation model and activation function, we developed a QNN model to evaluate its effectiveness by implementing image classification processes for the MNIST and Fashion-MNIST datasets. Figure 4d shows the processed quantized weights and activation values utilized in the software simulation. As the developed synaptic combination provides five balanced states, the weights are quantized into five values (−1, −1/2, 0, 1, 1/2, and 1). Similarly, the activation function is quantized into three different values (0, 1/2, and 1). Figure 4e (right) shows the neural network for image classification, which comprises four layers of QNNs: flattening, fully connected, activation, and output. We employed the quantized five values (−1, −1/2, 0, 1/2, and 1) and three values (0, 1/2, and 1) for the fully-connected and activation layers, respectively. The detailed parameters and layer information are provided in Note S5 (Supporting Information). For a comparative study, we performed identical image classification tasks using conventional full-precision deep neural networks (DNNs) (32 bits) and binary neural networks (BNNs) (1 bit). The data extracted from the weight values (F_1 and F_2) and activation functions (A_1) in the QNN model after training using the MNIST training image dataset is shown in Figure 4e (right). The sampled data, depicted as 8 × 8 images, represent the actual values used in the simulation.

Following training, we validated the trained model using quantized weight values and activation functions. Figure 4f presents the confusion matrix of the QNN model trained for image classification using the MNIST and Fashion-MNIST datasets. The predicted labels indicate that the machine classified the test images based on the trained model, whereas the true labels represent the actual image labels. The results of the classification tasks with both image datasets show that the QNN model using our ternary CTTs yielded superior performance, with a classification accuracy rate exceeding 90%.

Figure 4g shows the validation accuracy of the BNNs, full-precision DNNs, and developed QNN models for image classification tasks, where the validation accuracy was evaluated at each training epoch. In both tasks, the classification accuracy of the QNN model using ternary CTT was almost identical to that of the DNN model and superior to that of the BNN model. For the MNIST image classification task, the QNN and the full-precision DNN models achieved validation accuracies of 98.2% and 98.3%, respectively, whereas the BNN model achieved 97.4%. For the Fashion-MNIST dataset, the BNN, DNN, and QNN models achieved accuracies of 88.6%, 90.1%, and 89.5%, respectively. Thus, we confirmed that QNNs employing ternary CTTs require significantly fewer memory cells (two ternary bits) than full-precision DNNs (32 binary bits). Furthermore, the ternary-CTT-based QNN can achieve comparable classification accuracy for the MINST and Fashion-MNIST image classification tasks while minimizing energy consumption by reducing the number of accesses to the memory cells.

3. Conclusion

In summary, we developed a ternary CTT with unique multi-state non-volatile memory characteristics to achieve precise and energy-efficient data processing using a dedicated neural engine for power-constrained applications. Because the ternary CTT provides three programmable stable conductance states, we successfully demonstrated hardware-based QNN algorithm processing based on the ternary CTT characteristics. By mimicking

the biological excitatory and inhibitory synapse mechanisms, we demonstrated discrete balanced quinary current levels by combining two ternary CTTs as an artificial synaptic segregation model. The quantized quinary weights and ternary activation functions enabled by the fabricated CTTs were utilized to execute the QNN algorithm. This quantization technique allowed us to use fewer memory cells for the machine-learning algorithm compared with that for DNNs, which typically require a large memory space to store data. Furthermore, the software demonstration of the QNN algorithm indicated a high classification accuracy exceeding 90% for image processing using the MNIST and Fashion-MNIST datasets, which was comparable to the performance of full-precision DNNs. This unique approach to neural network computing using emerging devices and technologies will result in numerous novel in-memory computing and energy-efficient AIoT applications.

4. Experimental Section

Fabrication of Ternary CTT: For the fabrication of IGZO/IZO/IGZO ternary CTT, a Ti/Au gate metal (5/50 nm) was deposited by an e-beam evaporator on c-plane sapphire, patterned by photolithography. Then, an Al_2O_3 dielectric layer was deposited on top of the gate metal by a plasma-assisted atomic layer deposition system (ALD). For the active channel, the three layers of oxide semiconductors were deposited by DC/RF magnetron sputtering system with 3-inch IGZO and IZO sputtering targets. The first layer (IGZO, 10 nm) was sputtered by RF power of 100 W under 5 mTorr working pressure. The gas volume ratio of $[O_2]/[Ar]$ was fixed at 0.1 during the deposition, and the deposition rate was 0.06 nm s^{-1}. The second layer (IZO, 15 nm) was sputtered by a DC power of 100 W under 5 mTorr working pressure. The gas volume ratio of $[O_2]/[Ar]$ was fixed at 0.2 during the deposition, and the deposition rate was 0.2 nm s^{-1}. The third layer (IGZO, 5 nm) was sputtered by the same conditions as the first layer deposition. Then, the post-annealing process was implemented at 250 °C for 1 h in the ambient condition. The channel was defined via photolithography and wet etching process by using buffered oxide etchant (BOE). Following the etching process, Ti/Pt the source-drain metal (5/35 μm) was deposited by e-beam evaporation, and the metal contact was patterned by photolithography, where the W/L ratio of the channel was 7 μm/ 140 μm with 10 fingers of interdigitated design.

Electrical Measurement: The electrical performances of the ternary CTT were characterized by using a Keysight B1500A Semiconductor Analyzer. The pulse modulation was implemented by using MATLAB software to control Keysight B1500A Semiconductor Analyzer. During the programming, the voltage pulse train was applied to the gate node where the source and drain nodes were connected to the ground. *I–V* characteristics of the ternary CTT were investigated by using a Keysight B1500A Semiconductor Analyzer. A sampling rate of 1 ms, a hold time of 10 ms, and a delay time of 10 ms was employed as well as a 500-μs pulse width without the hold and delay times was used.

Each state in the ternary CTT was achieved by the pulse modulation using gate voltage pulses while drain and source nodes were grounded. State 0 was programmed from State 1 using a voltage amplitude of +35 V, a pulse width of 500 μs, and a single pulse. State ½ was achieved from State 1 with a voltage amplitude of +15 V, a pulse width of 500 μs, and a single pulse. State 1, on the other hand, was programmed using a voltage amplitude of −35 V, a pulse width of 500 μs, and 100 pulses. Additionally, when investigating the achievable maximum and minimum conductance values, the maximum voltage amplitude applied was ± 40 V.

Technology Computer-Aided Design Software Simulation: To confirm the operation mechanism of ternary IGZO/IZO/IGZO transistors, a Sentaurus TCAD (Synopsys corporation) simulation was performed. The simulation structure was constructed using the measured dimensions including the channel width, channel length, and layer thickness, as extracted from SEM and TEM images. The physical parameters, such as the effective density of electrons, band gap energy, and electron affinity, were considered based on reference values. To represent the sub-bandgap characteristics of the IGZO and IZO channels, a donor-like and acceptor-like trap model was employed, along with the trap concentration, energy mid of the trap, and distribution type. In addition, the constant and high field saturation mobility models were used to calculate the mobility in low and high electric fields. The device and material properties are described in detail in Note S1 (Supporting Information).

Layer Structure of the QNN: To demonstrate the effectiveness of QNN based on the ternary CTTs, software simulations were implemented in Python programming language and environment. The simulations were performed on the MNIST and Fashion MNIST datasets, which consist of 28×28 pixel images. The input layer of the network consisted of 784 neurons, representing each pixel of the flattened image data. A hidden layer with 512 neurons was employed, followed by batch normalization for stable and fast computing in the training process. The ternary sigmoid function served as the activation function in the hidden layer. After then, the output layer consisted of ten neurons, corresponding to the ten classes in the datasets, and the softmax activation function was applied for the final classification. The detailed parameters are described in Note S6 (Supporting Information).

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported by the U.S. National Science Foundation (NSF) under grant No. 1942868. This work was also supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2021R1A4A1033155).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

artificial intelligence, charge-trap transistors, oxide thin-film transistors, quantized neural networks, ternary transistors

Received: May 7, 2023
Revised: July 31, 2023
Published online: September 1, 2023

[1] A. M., G. H. Alex Graves, *Dep. Comput. Sci. Univ. Toronto* **2013**, *3*, 45.
[2] D. Bahdanau, K. H. Cho, Y. Bengio, *3rd Int. Conf. Learn. Represent. ICLR 2015 – Conf. Track Proc* **2015**, 1.
[3] L. Hertel, E. Barth, T. Kaster, T. Martinetz, *Proc. Int. Jt. Conf. Neural Networks* **2015**, 2015-Sept., 20.

[4] N. Mellempudi, A. Kundu, D. Mudigere, D. Das, B. Kaul, P. Dubey, arXiv preprint arXiv:1705.01462 **2017**.
[5] H. Alemdar, V. Leroy, A. Prost-Boucle, F. Petrot, *Proc. Int. Jt. Conf. Neural Networks* **2017**, May 2547.
[6] D. Wan, F. Shen, L. Liu, F. Zhu, J. Qin, L. Shao, H. T. Shen, in *Proceedings of the European Conference on Computer Vision (ECCV)*, **2018**, pp. 322.
[7] M. Di Marco, M. Forti, L. Pancioni, G. Innocenti, A. Tesi, *IEEE Trans. Cybern.* **2020**, *52*, 1822.
[8] Y. Halawani, B. Mohammad, M. Abu Lebdeh, M. Al-Qutayri, S. F. Al-Sarawi, *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2019**, *9*, 388.
[9] M. Le Gallo, A. Sebastian, G. Cherubini, H. Giefers, E. Eleftheriou, *IEEE Trans. Electron Devices* **2018**, *65*, 4304.
[10] H. S. Lee, Y. Baek, Q. Lin, J. Minsu Chen, M. Park, D. Lee, S. Kim, K. Lee, *Adv. Intell. Syst.* **2020**, *3*, 2000202.
[11] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, *Nat. Electron.* **2018**, *1*, 52.
[12] L. Deng, P. Jiao, J. Pei, Z. Wu, G. Li, *Neural Networks* **2018**, *100*, 49.
[13] S. Zheng, Y. Liu, S. Yin, L. Liu, S. Wei, *Proc. – Des. Autom. Conf.* **2018**, 137.
[14] J. Jang, S. Gi, I. Yeo, S. Choi, S. Jang, S. Ham, B. Lee, G. Wang, *Adv. Sci.* **2022**, *9*, 2201117.
[15] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, D. Strukov, *Nat. Commun.* **2018**, *9*, 2331.
[16] S. Jain, S. K. Gupta, A. Raghunathan, *IEEE Trans. Very Large Scale Integr. Syst.* **2020**, *28*, 1567.
[17] H. Xiao, K. Rasul, R. Vollgraf, *arXiv:1708.07747* **2017**.
[18] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, *Proc. IEEE* **1998**, *86*, 2278.
[19] G. Liu, *Nat. Neurosci.* **2004**, *7*, 373.
[20] J. W. Jeong, Y.-E. Choi, W.-S. Kim, J.-H. Park, S. Kim, S. Shin, K. Lee, J. Chang, S.-J. Kim, K. R. Kim, *Nat. Electron.* **2019**, *2*, 307.
[21] T. M. Bartol, C. Bromer, J. Kinney, M. A. Chirillo, J. N. Bourne, K. M. Harris, T. J. Sejnowski, *Elife* **2015**, *4*, 10778.
[22] H. M. El-Boghdadi, *J. Supercomput.* **2015**, *71*, 28.
[23] S. Brink, S. Nease, P. Hasler, *Neural Networks* **2013**, *45*, 39.
[24] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, E. Eleftheriou, *Nat. Nanotechnol.* **2020**, *15*, 529.
[25] S. Shin, H. Y. Jung, B. H. Juang, *Proc. Annu. Conf. Int Speech Commun. Assoc. Interspeech* **2011**, *102*, 1713.
[26] M. P. Hung, D. Wang, T. Toda, J. Jiang, M. Furuta, *ECS J. Solid State Sci. Technol.* **2014**, *3*, Q3023.
[27] M. P. Hung, D. Wang, J. Jiang, M. Furuta, *Proc. Int. Disp. Work.* **2013**, *1*, 286.
[28] M. P. Hung, D. Wang, J. Jiang, M. Furuta, *ECS Solid State Lett.* **2014**, *3*, Q13.
[29] J. Y. Bak, M.-K. Ryu, S. H. K. Park, C. S. Hwang, S. M. Yoon, *IEEE Electron Device Lett.* **2014**, *35*, 357.
[30] M. M. Billah, A. B. Siddik, J. B. Kim, D. K. Yim, S. Y. Choi, J. Liu, D. Severin, M. Hanika, M. Bender, J. Jang, *Adv. Electron. Mater.* **2021**, *7*, 2000896.
[31] S. Jeon, S.-E. Ahn, I. Song, C. J. Kim, U.-I. Chung, E. Lee, I. Yoo, A. Nathan, S. Lee, K. Ghaffarzadeh, J. Robertson, K. Kim, *Nat. Mater.* **2012**, *11*, 301.
[32] M. Mativenga, M. Seok, J. Jang, *Appl. Phys. Lett.* **2011**, *99*, 122107.
[33] Y. Yamauchi, Y. Kamakura, Y. Isagi, T. Matsuoka, S. Malotaux, *Jap. J. Appl. Phys.* **2013**, *52*, 094101.
[34] A. Suresh, S. Novak, P. Wellenius, V. Misra, J. F. Muth, *Appl. Phys. Lett.* **2009**,*94*, 123501.
[35] M.-F. Hung, Y.-C. Wu, J.-J. Chang, K.-S. Chang-Liao, *IEEE Electron Device Lett.* **2013**, *34*, 75.
[36] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, *Nature* **2020**, *577*, 641.
[37] M.-H. Kim, S.-H. Lee, S. Kim, B.-G. Park, *IEEE Access* **2022**, *10*, 37030.
[38] Y. Du, L. Du, W. Fan, Y. Xiao, M.-C. F. Chang, *IEEE J. Explor. Solid-State Comput. Devices Circuits* **2021**, *7*, 10.
[39] S. Venkataramani, K. Roy, A. Raghunathan, *Proc. Asia South Pacific Des. Autom. Conf. ASP-DAC* **2016**, 25–28 January, 308.
[40] P. Yao, H. Wu, B. Gao, S. B. Eryilmaz, X. Huang, W. Zhang, Q. Zhang, N. Deng, L. Shi, H.-S. P. Wong, H. Qian, *Nat. Commun.* **2017**, *8*, 15199.