Ph.D. DISSERTATION

A Novel Design of Overpass Channel Synapse Array with Ultra-Low Power Operation for Neuromorphic Systems

초저전력 뉴로모픽 시스템을 위한
육교형 시냅스 어레이

BY

TAE JIN JANG

AUGUST 2023

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

A Novel Design of Overpass Channel Synapse Array with Ultra-Low Power Operation for Neuromorphic Systems

초저전력 뉴로모픽 시스템을 위한
육교형 시냅스 어레이

지도교수 최 우 영

이 논문을 공학박사 학위논문으로 제출함

2023 년 8 월

서울대학교 대학원

전기 · 정보공학부

장 태 진

장태진의 공학박사 학위논문을 인준함

2023 년 8 월

위 원 장 : 김 재 준 (인)
부위원장 : 최 우 영 (인)
위 원 : 조 성 재 (인)
위 원 : 김 윤 (인)
위 원 : 이 수 연 (인)

Abstract

Von Neumann-based deep neural networks have achieved excellent performance with the rapid development of computing power. However, moving data consumes a significant amount of time and energy due to the serial connection of the central processing unit (CPU) and memory. As a result, neuromorphic systems have emerged as a promising approach to address the computational challenges of artificial neural networks while maintaining extremely-low-power operation. In particular, synapse, one of the crucial building blocks of neuromorphic systems, stores the weights between neurons and transmits signals. According to Kirchhoff's current law, the artificial synapse can operate quickly with low power consumption because vector-matrix multiplication (VMM) is expressed as a current sum.

Therefore, various synaptic devices have been proposed, including memristor-based two-terminal devices and flash memory-based synaptic devices. The two-terminal devices have a simple structure and are advantageous for high-density and large-capacity integration. However, these face challenges such as device variation, reliability, and sneak current. On the contrary, flash memory-based synaptic devices are a mature field with a long history of research, offering stable and multi-bit operation. NAND-, NOR-, and AND-type arrays are typically implemented as

synapse arrays. NAND-type arrays require an additional circuit because the weight values must be read sequentially. NOR- and AND-type arrays can perform VMM operations by sensing current from the source line. However, these arrays suffer a limited degree of integration.

This dissertation proposes a poly-Si overpass channel synaptic (OCS) transistor with scaled cell size for extremely-low-power operation. The OCS transistor exhibits two key structural advantages. Firstly, the on-current is decreased to sub-100 nA, maintaining a high on/off ratio due to the channel wrapping around the fin-shaped bottom gate. Secondly, the weights of the OCS transistors can be finely adjusted by augmenting the volume of the charge storage layer. We experimentally demonstrate the inference and weight update of the fabricated NOR-type OCS array.

It is verified that the fabricated diode-connected (D-C) OCS array is suitable for VMM, exhibiting a weighted-sum error of less than 1% during inference. The synaptic weights in the D-C OCS array are adjusted to sub-nA resolution using Fowler-Nordheim (FN) tunneling with asymmetric gates. Furthermore, stable operation is demonstrated by verifying process-voltage-temperature variations. Finally, the classification accuracy for the Fashion MNIST dataset reaches 91.29% after a year, with four-bit quantization of spiking neural networks (SNNs).

Keywords: poly-Si based synaptic device, NOR-type array, flash memory, low-power operation, spiking neural network (SNN), neuromorphic system.

Student Number: 2016-20970

Contents

List of Table

List of Figures

Chapter 1

Introduction

1.1 Neuromorphic Systems

Computing devices based on the Von Neumann architecture have experienced a dramatic increase in computing performance, showing notable advantages in arithmetic operations. Furthermore, artificial neural networks (ANNs) have made tremendous advancements in image classification, defect detection, and autonomous driving using vast data and advanced learning algorithms [1-3]. However, the computational workload grows as ANNs become more complex and require massive data. Consequently, the significant time and power consumption present the most considerable challenges ANNs must address. The performance gap between the central processing unit (CPU) and memory is growing, which results in the Von

Neumann bottleneck during the data transfer process, as illustrated in **Figure 1.1** and **Figure 1.2** [4, 5]. Therefore, numerous studies on innovative data processing have been actively conducted to overcome the limitations of conventional computing architectures [6-8].

The human brain consumes approximately less than 20 W, capable of processing various and complex data in real-time. Therefore, research on neuromorphic systems, which emulate human neural systems in hardware, has been actively explored. The primary goal of these neuromorphic systems is the efficient integration of critical elements, such as neural circuits and synaptic arrays, which are massively interconnected in parallel.

The role of a neuron is to receive and integrate signals from numerous synapses through dendrites. Additionally, when the membrane potential exceeds a specific threshold voltage, the signals are produced. Subsequently, the membrane potential undergoes a refractory period before entering the resting state, preparing to receive other signals. Neurons consume energy only when they fire, resulting in low power consumption. Therefore, research has been conducted on integrate-and-fire neuron circuits to emulate these characteristics.

A prominent example is a neuron circuit composed of inverters and capacitors [9-12]. However, there exists a limitation in the integration density due to the large area occupied by capacitors. Thus, much research has been focused on creating neuron

circuits without capacitors [13-17].

Synapses facilitate signal transmission from the pre-neuron to the post-neuron. Simultaneously, they store the synaptic weight between neurons, with excitatory synapses increasing and inhibitory synapses decreasing the membrane potential. Those synaptic weights are updated to deal with the various complex situations. Therefore, synapses perform two core functions, memory and signal transmission, representing one of the most critical components in neuromorphic systems.

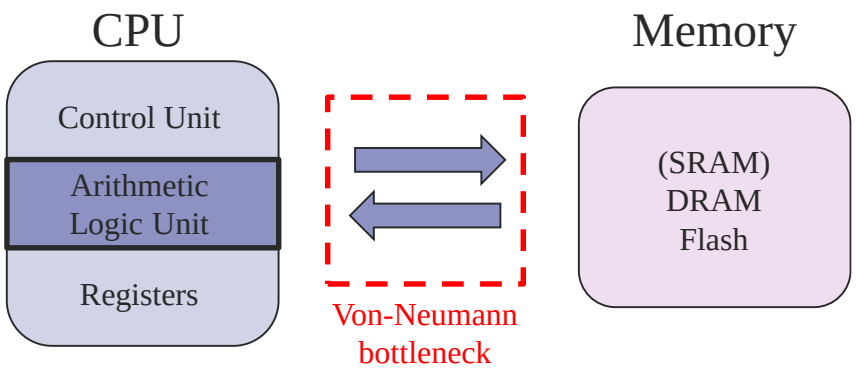


Figure 1.1. Schematic of Von Neumann architecture.

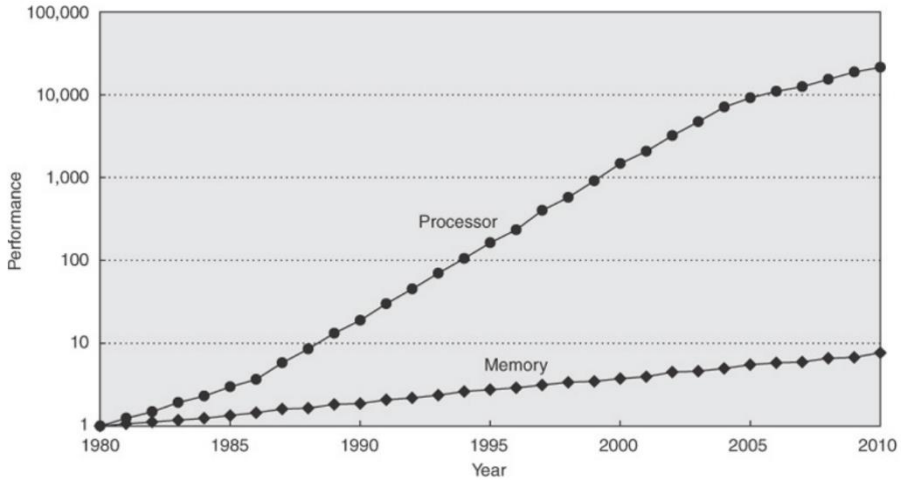


Figure 1.2. Annual performance comparison between CPU and memory [4].

1.2 Candidates for Synaptic Device

1.2.1 SRAM & Memristor

The human brain has approximately 100 billion neurons and 150 trillion synapses. As a result, artificial synapses and neurons must enable low power, high density, and stable operation. Additionally, a synaptic device requires memory operation to store multiple weights. Many synaptic devices have been proposed since the absence of a synaptic device that meets all these requirements.

The country and industry have extensively investigated static random access memory (SRAM) based digital-analog mixed-type neuromorphic chips. Representatively, state-of-the-art process technologies such as Spinnaker, TrueNorth, and Loihi integrated hundreds of millions of synapses and hundreds of thousands of neurons, as illustrated in **Figure 1.3** [18-23]. Implementing these technologies is crucial in enabling energy efficient and scalable designs, both of which are vital for advancing neuromorphic systems. However, there are significant limitations in terms of integration due to the substantial number of transistors required for SRAM-based neuromorphic chips.

Two terminal memristor devices have attracted attention as synaptic devices due to their simple structures and ease of integration, as shown in **Figure 1.4**.

Representatively, memristor devices such as resistive random access memory (RRAM), phase-change memory (PCM), and ferroelectric random access memory (FeRAM) can perform vector matrix multiplication (VMM) in a cross point array [24-31]. Memristors capable of gradual switching has been proposed to represent various states, and a neuromorphic system has been verified by implementing a large synapse array. However, these memristors still require verification of stable operating characteristics such as retention and device variation. Additionally, due to sneak current, a fundamental limitation of two terminal devices, errors may occur in VMM [32-34]. Consequently, additional devices, such as extra transistors or selectors, must address this problem.

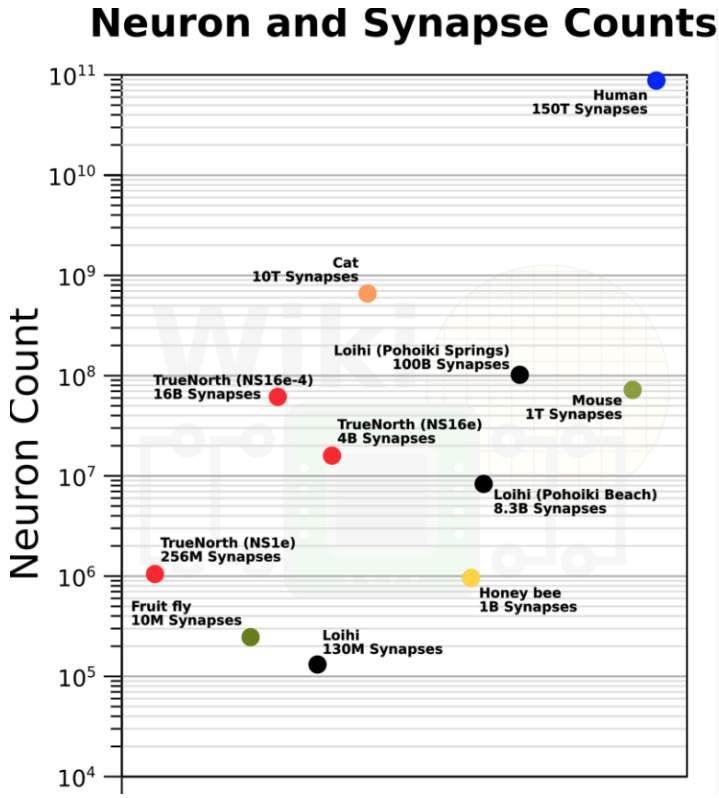


Figure 1.3. Recent neuromorphic chips [35].

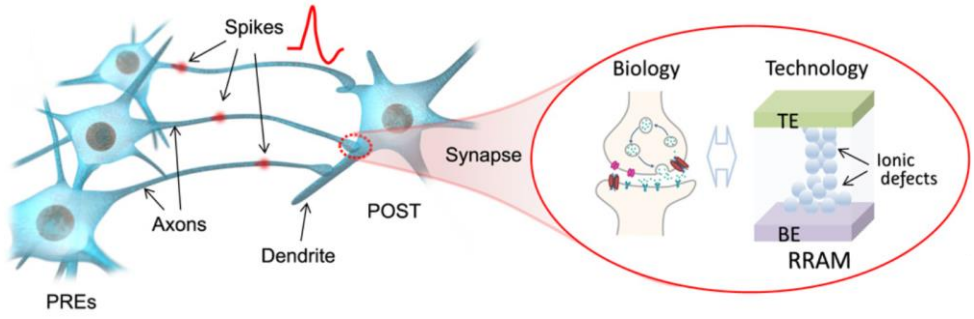


Figure 1.4. Schematic of artificial synapse with resistive random access memory [29].

1.2.2 Flash memory

Since flash memory has been commercialized for decades, it is a highly mature field. It possesses significant advantages as a synaptic device due to its cycle endurance and retention stability compared to other memristors. Flash memory-based synaptic devices include NAND-, NOR-, and AND-type arrays with unique structures and operation mechanisms.

Table 1.1 shows the main characteristics of flash memory-based synaptic arrays. The vertical integration capability of the NAND-type array offers a significant advantage in terms of integration compared to other devices. However, sensing the conductance of an individual cell necessitates applying a pass voltage to other transistors within the string due to the series connection. Consequently, the conductance of the target cell is distorted, depending on the weights assigned to other cells in the exact string. Moreover, VMM cannot be executed directly through current sensing owing to the series structure. Thus, an external analog-to-digital converter or a circuit is required for performing calculations [36, 37].

Both NOR- and AND-type arrays have a structure similar to a neural network because they can produce parallel output signals when a voltage input is applied to the word line (WL) [38-40]. Therefore, these arrays perform VMM by sensing the current, which is the product of the conductance and voltage input, on the source line

(SL). The two structures exhibit different SL and drain line (DL) arrangements. SL and DL are perpendicular in the NOR-type array, whereas they are arranged parallelly in the AND-type array, as shown in **Figure 1.5**.

During the weight adjustment process, the NOR-type array modulates the conductance of individual cells through hot carrier injection (HCI). In this process, hot carriers damage the tunneling or gate oxide, resulting in reliability problems. Moreover, energy consumption is high since the current continuously flows, and program efficiency decreases. Conversely, the AND-type array uses Fowler-Nordheim (FN) tunneling for program and erase, resulting in lower energy consumption and precise conductance control using incremental step pulse program (ISPP) and incremental step pulse erase (ISPE) methods.

In inference, the NOR-type array is advantageous for low-power event-driven operation since the same voltage input is applied to both the word line (WL) and DL. Therefore, the array consistently operates in the saturation region, enabling stable current output and complete suppression of off-current. Otherwise, the voltage must always be applied to the DL in the AND-type array. Therefore, the leakage current accumulates on the SL even if no signal is received from any neurons. An additional peripheral circuit can be integrated to alleviate this issue. This configuration enables DL to receive only voltage when a signal is applied to the WL, eliminating the leakage current without a voltage input. Nevertheless, when the voltage input is applied to the

WL, the leakage current is still summed in the SL. This leakage current subsequently leads to VMM errors.

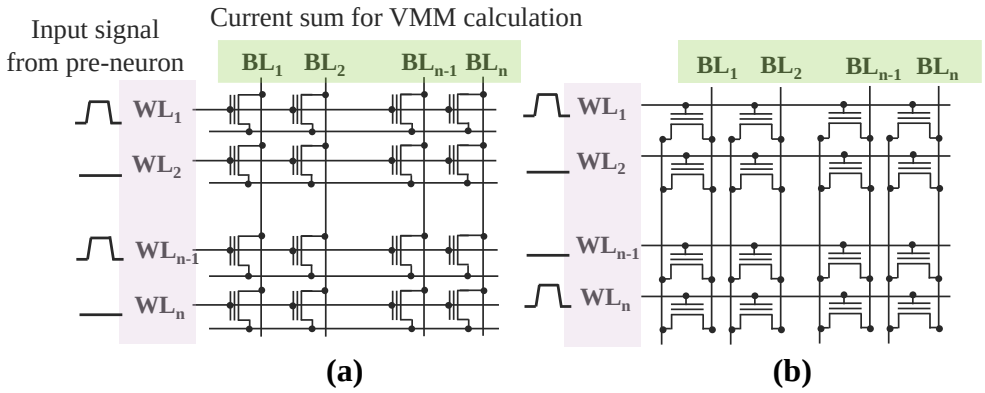


Figure 1.5. Schematic of (a) NOR-type array, (b) AND-type array.

Table 1.1. Comparison of critical characteristics among the several arrays.

	NAND-type array	NOR-type array	AND-type array	This work
VMM calculation	Current sum (serial connection)	Current sum at source line		
Weight update (method)	Fowler-Nordheim tunneling	Hot carrier injection	Fowler-Nordheim tunneling	
Weight update (power consumption)	Low	High	Low	Low
Inference mode	Gate (clock base)	Gate & drain or gate	Gate	Gate & drain
Leakage current	Low	Low	High	Low

1.3 Outline of Dissertation

This dissertation presents a flash memory-based ultra-low power overpass channel synaptic transistor and operating scheme. We also verify its performance through high-level simulation that accounts for the characteristics of the fabricated synapse array.

In Chapter 2, the concept and operation mechanism of the proposed device are described, and compatibility with neurons is verified. In addition, the current level is optimized through TCAD simulation. The low-power program/erase scheme is proposed, and the adjustment of individual cell conductance is demonstrated.

Chapter 3 explains the entire process flow of the scaled device and array, and the electrical characteristics of the fabricated device are analyzed with a process-voltage-temperature variation. Synaptic weight control of the target cell as well as vector multiplication, is shown in this chapter. Furthermore, the diode-connected synapse array that shares a gate and drain is proposed for high integration. This array has advantages in minimizing the leakage current.

In the final chapter, we perform high-level simulations on the Fashion MNIST dataset based on the retention and weight-tuning characteristics. Finally, we validate

the feasibility of the proposed overpass channel synaptic transistor for ultra-low power neuromorphic systems.

Chapter 2

Overpass Channel Synaptic Transistor

2.1 4-terminal Flash Memory-based Synaptic Device

2.1.1 Previous 4-terminal Devices

Previously, several flash-based four-terminal transistors are proposed for synaptic devices. The NOR-type synaptic array of the proposed device controls the conductance of individual cells by using two asymmetric gates [41-43]. The characteristics of these devices are summarized in **Table 2.1**, and they implemented both short- and long-term memory of synaptic characteristics. Additionally, the gate is located far from the charge storage layer (CSL), providing significantly improved resistance to read disturb during inference. However, there are some limitations to

ultra-low power operation due to the adjustment of weights through hot carrier injection or the current level at the uA or higher. Specifically, as scaling down the cell size, degradation occurs due to various short channel effects.

Table 2.1. Characteristics of the 4-terminal synaptic device.

	[42]	**[41]**	**This work**
Channel material	Single crystalline silicon	Polycrystalline silicon	Polycrystalline silicon
Program/Erase	Hot carrier injection	FN tunneling	FN tunneling
On current	> 1 μA	> 1μA	< 100nA

2.1.2 Mechanism of Overpass Channel Synaptic Transistor

As described in section 2.1.1, when the critical dimension (CD) of a flash memory device decreases, the device characteristics deteriorate due to the short channel effects, and the amount of charge that can be stored in the CSL drastically decreases. As shown in **Figure 2.1**, the trap density of Si_3N_4 is approximately $10^{19}/cm^3$, and the maximum trapped charge is only about 100-200 in devices with tens of nm of CD.

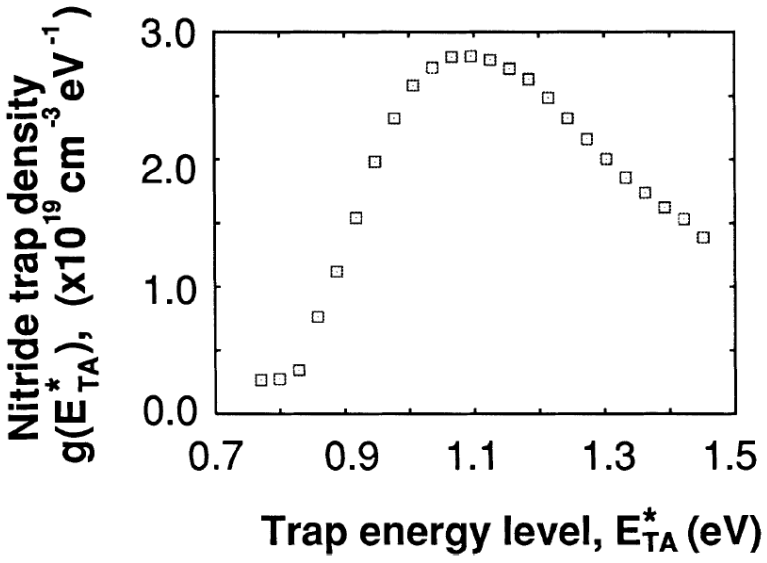


Figure 2.1. Trap density of Si_3N_4 as a function of trap energy level [44].

Thus, there are limitations in implementing various states, and retention problems arise due to charge loss. Increasing the gate length and CSL is necessary to solve these problems fundamentally. In the case of NAND flash memory, these led to the development of 3D NAND flash structures, which stack cells vertically.

Similar to the reasons mentioned above, the overpass channel synaptic (OCS) transistor is proposed to form a vertical structure for a longer effective gate length (EGL) and CSL, as illustrated in **Figure 2.2**(a). EGL and CSL can be formed larger depending on the height of the fin and can be expressed by the following eq. (2.1).

$$\text{Effective gate length} = \text{gate length} + 2 \times \text{fin height} \tag{2.1}$$

Thus, the various side effects can be fundamentally resolved during scaling down. Furthermore, the increased EGL along with the poly-Si channel, significantly reduces the on-current of the device. It offers substantial benefits for ultra-low-power operation for neuromorphic systems.

Additionally, by forming an electric field capable of FN tunneling in the tunneling oxide with asymmetric gates, the conductance of the OCS transistor can be adjusted by injecting electrons or holes into the CSL. As shown in **Figure 2.2**(b), the trapped charge influences the threshold voltage of the top gate, similar to the body effect. Moreover, the impact of the trapped charge on the threshold voltage of the top gate

can be enhanced by forming a thin active layer.

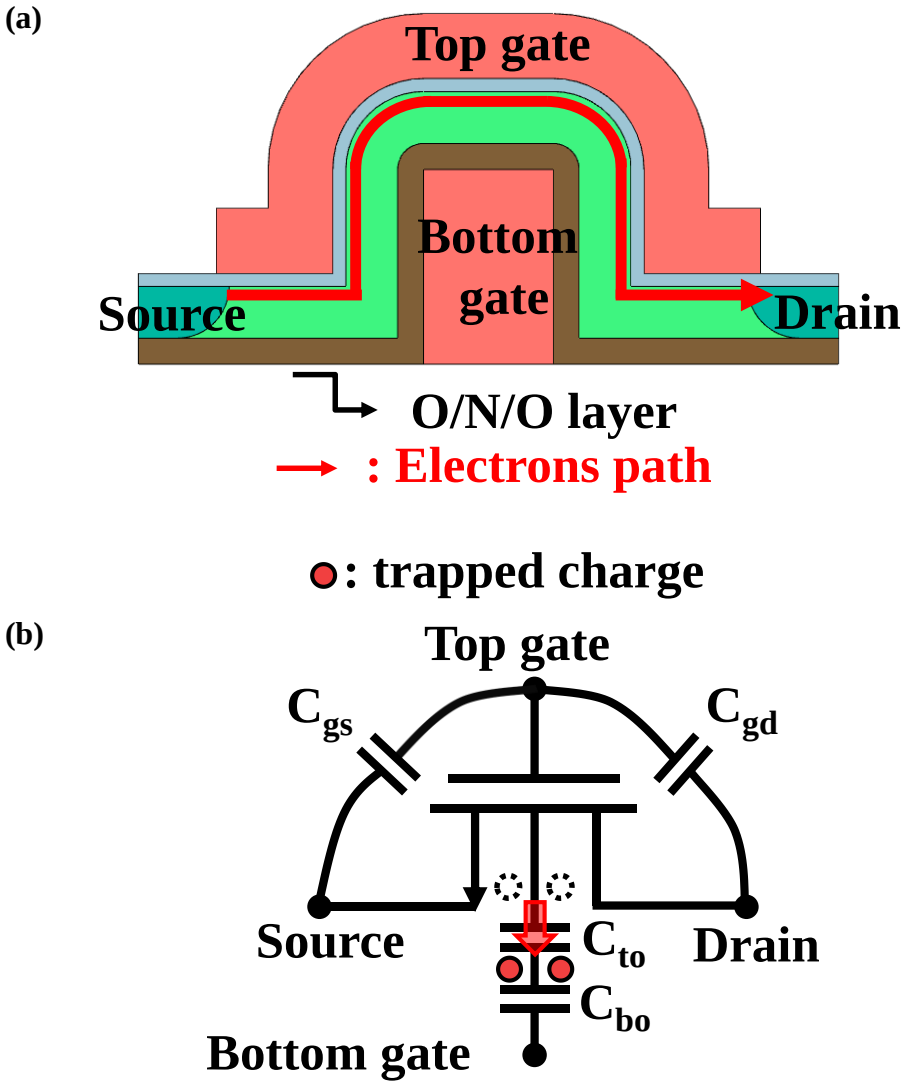


Figure 2.2. (a) Schematic of overpass channel synaptic (OCS) transistor. (b) Capacitance network of OCS transistor. The capacitance between the floating body and the bottom gate is composed of the serial connection of the capacitances with the tunneling oxide (C_{to}), Si_3N_4 and blocking oxide (C_{bo}).

2.2 TCAD Simulation Results

2.2.1 Neuron Circuit Simulation

Synapse arrays and neuron circuits are fundamental building blocks of neuromorphic systems, interacting and exchanging signals. Therefore, when designing a synapse array, it is essential to consider compatibility with the neuron circuits. As shown in **Figure 2.3**, current flowing through excitatory and inhibitory synaptic arrays, enters the post-neurons via the current mirror. Consequently, setting an appropriate current level in the synaptic array is crucial, as it determines the energy consumption and operating speed of the systems.

The current mirror at the front of the neuron circuits transmits current to the neuron, regardless of the charge accumulated on the membrane capacitance, by fixing the source voltage of the synaptic transistor. Thus, the current mirror in the post-neuron section is simulated with the Silvaco SmartSpice tool. The transistor model used in this thesis is BSIM-CMG model version 105.0 and the 45 nm FinFET technology node. The synaptic current is optimized by analyzing the output current (I_{out}) according to the input current (I_{in}) with a channel length of 45 nm and a fin pitch of 80nm.

If I_{in} exceeds the maximum level that the M1 transistor can drive, not only is the

current not accurately copied, but a substantial amount of energy is consumed, as shown in **Figure 2.4**. If I_{in} is too small, the RC delay dramatically increases due to parasitic capacitance components. For instance, when I_{in} is 224 pA, the propagation delay (t_{pd}) for charging the gate capacitances of the current mirror is 225 ns, as illustrated in **Figure 2.5**. Consequently, when a voltage input with a pulse width of several μs is applied, a significant signal loss of tens of percent occurs, leading to VMM errors.

Due to sub-nA I_{in} causing large t_{pd} of hundreds of ns, the current transmitted to the post-neuron becomes distorted, as shown in **Figure 2.6**. Therefore, by setting the current in the synapse to tens of nA, t_{pd} and power consumption are minimized.

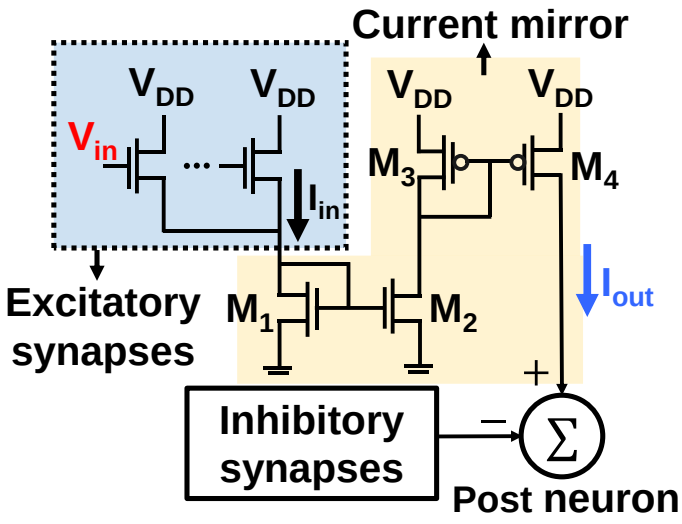


Figure 2.3. Schematic of neuron circuit and synaptic arrays performing VMM.

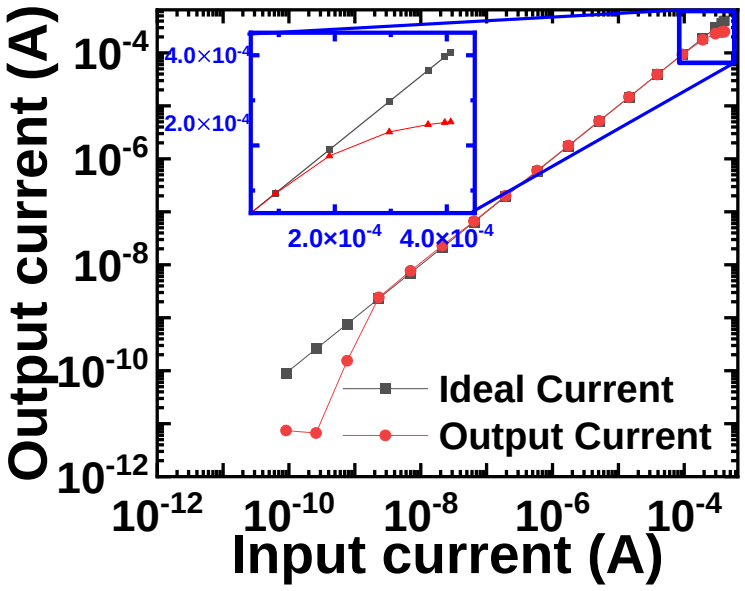


Figure 2.4. The M1 transistor's capability to manage the current and the RC delay determines the current propagation.

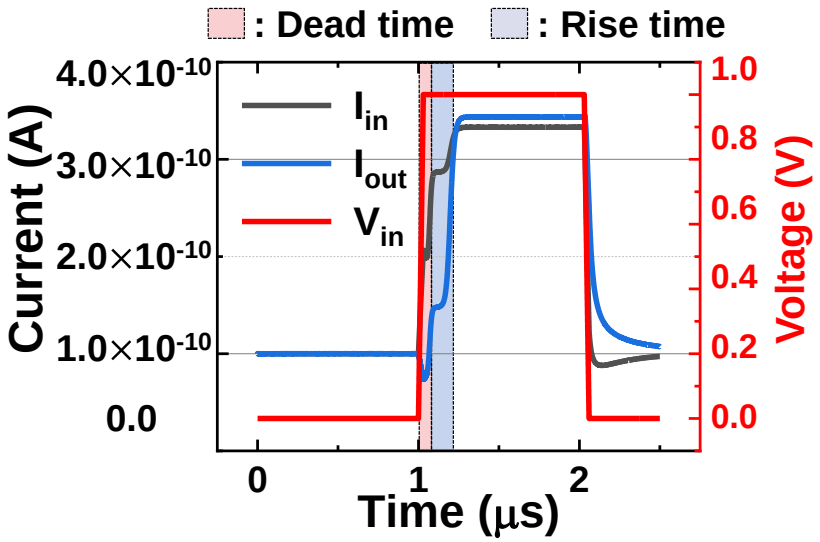


Figure 2.5. The product of the dead time (t_{dead}) until the output current occurs and the rise time (t_{rise}) until 90% of the maximum current is the propagation delay (t_{pd}).

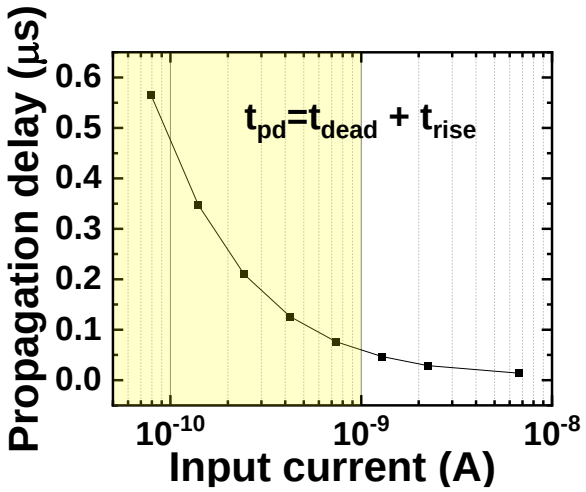


Figure 2.6. propagation delay vs. input current.

2.2.2 Program/Erase Scheme (unit cell)

The hot carrier injection, commonly used in NOR-type arrays, causes damage to the gate oxide and diminishes program/erase efficiency. In contrast, the proposed 4-terminal OCS transistor utilized two asymmetric gates to trap charge in the CSL through FN tunneling. As depicted in **Figure 2.7**, electrons from the top gate are injected into the CSL through FN tunneling while erase injects holes. Simultaneously, the source and drain remain floating and do not directly participate in the program and erase operations. The program/erase conditions are summarized in **Table 2.2**, and a sufficient electric field is applied to the tunneling oxide by the two gates to inject charges.

Figure 2.8(a) illustrates the key parameters of the OCS transistor: fin height (H_{fin}), the tail part where the top gate covers the bottom active (T_{tail}), and active thickness (t_{si}). Because H_{fin} causes an increase in the effective gate length and CSL volume, the maximum trapped charge increases while the current is reduced, as depicted in **Figure 2.8**(b). Furthermore, the thinner t_{si} results in a larger electric field between the two gates, which enhances the program efficiency. Consequently, increasing t_{si} from 7 nm to 40 nm augments the amount of trapped charge, as shown in **Figure 2.9**(a). Additionally, if t_{si} is thin enough, the trapped charge greatly influences the threshold voltage shift of the top gate, as depicted in **Figure 2.9**(b).

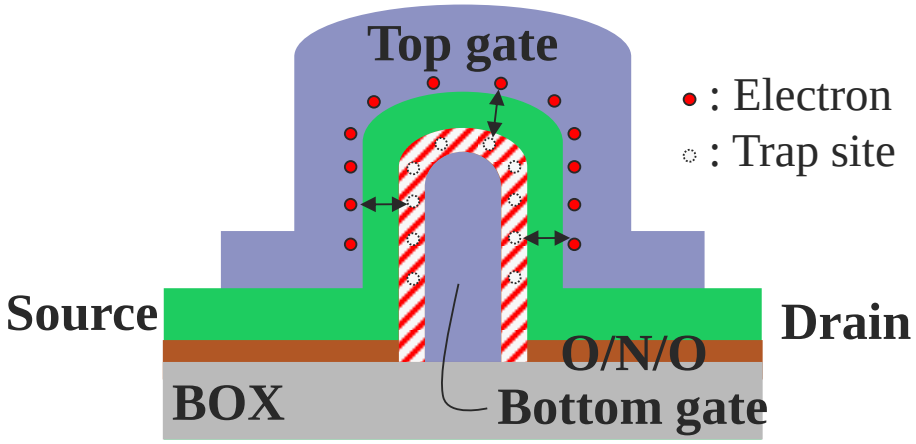


Figure 2.7. Schematic of program/erase operation with FN tunneling.

Table 2.2. Voltage conditions of program/erase operation with FN tunneling.

	Program	**Erase**
Top gate voltage (V)	0	V_{ers}
Bottom gate voltage (V)	V_{pgm}	0
Source/Drain voltage (V)	Floating	Floating

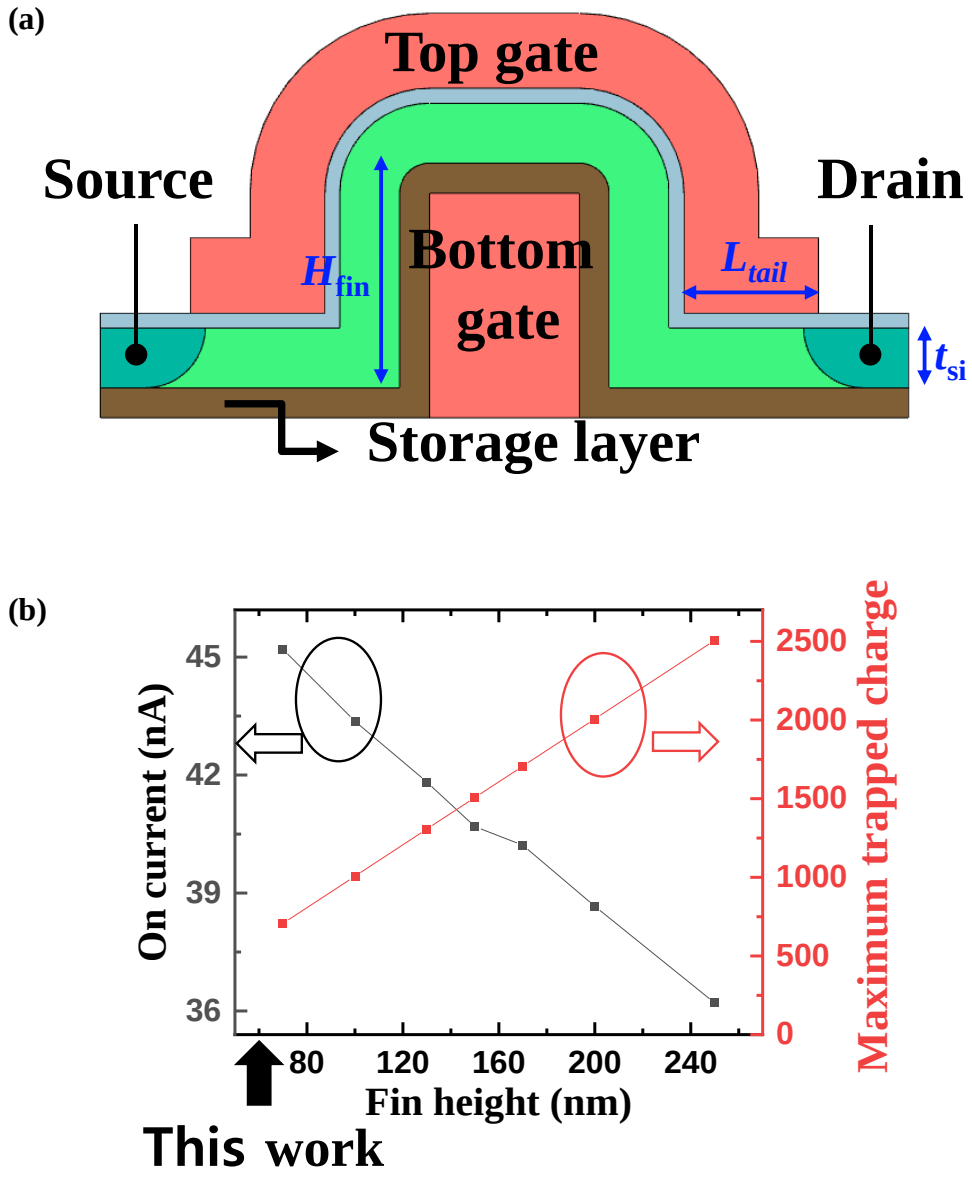
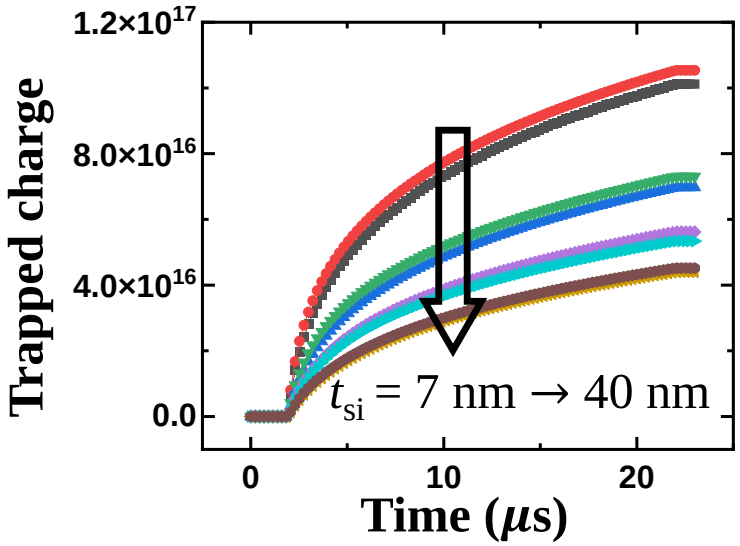


Figure 2.8. (a) Core parameters of OCS transistor. (b) On current and maximum trapped charge of OCS transistor as a function of the fin height.

(a)



(b)

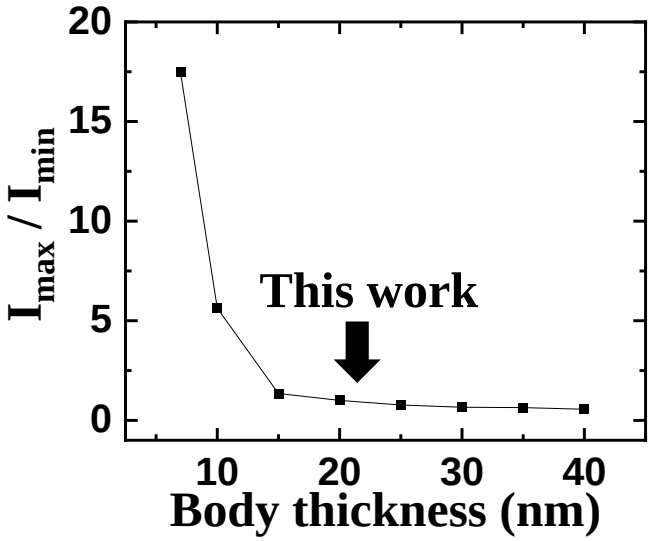


Figure 2.9. (a) Trapped charge in the CSL of OCS transistor during the program with different active thicknesses (t_{si}). (b) Program efficiency as a function of the t_{si}.

2.2.3 Program/Erase Scheme (NOR-type Array)

Accurate weights must be established in the target cell within the NOR-type array for practical implementation. Therefore, optimizing the program (PGM)/erase (ERS) scheme is essential while considering the PGM/ERS disturb on neighboring cells. The primary 2×2 NOR-type array is depicted in **Figure 2.10**, with the voltage condition of the target cell presented in **Table 2.2**. The conductance change in adjacent cells can be prevented by applying 1/3 V_{pgm} and 1/2 V_{ers} as the PGM/ERS disturbance voltage.

Since the SL and DL are floating in the NOR-type array, the voltage of SL and DL is not fixed. However, due to the inherent characteristics of the proposed OCS transistor structure, the T_{tail} can effectively block the SL/DL voltage. In other words, as depicted in **Figure 2.11**, the SL and DL voltages vary depending on the top gate voltage conditions in the capacitance network of the array. However, when 0 V is applied to the top gate, the channel is cut off at $T_{tail,}$ and the V_{pgm} of the bottom gate reduces the conductance of the target cell.

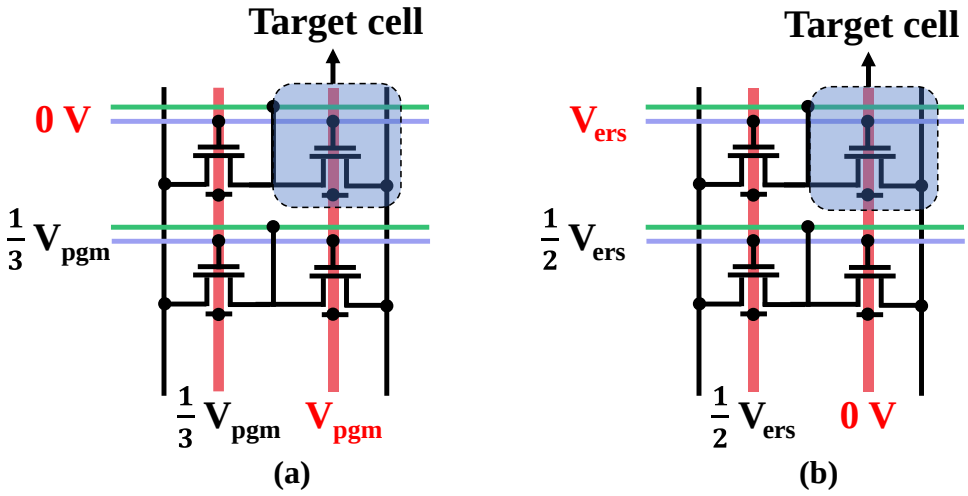


Figure 2.10. Schematic of (a) program (b) erase operation with FN tunneling.

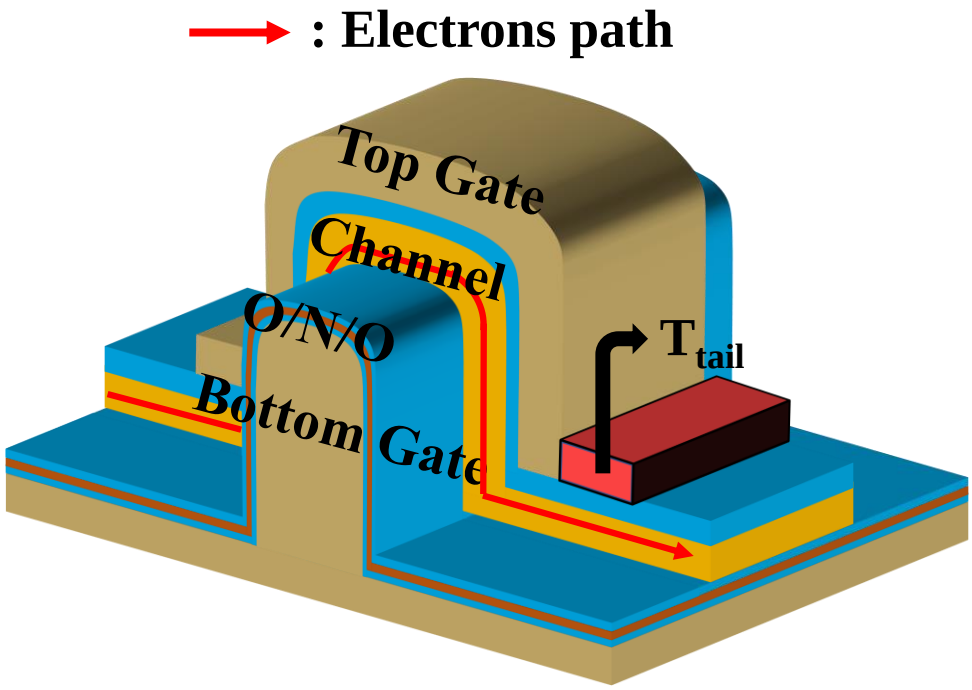


Figure 2.11. Schematic of the OCS transistor.

In the synapse array, the length of T_{tail} plays a crucial role in inhibiting neighboring cells by blocking the SL and DL voltage. When programming the target cell, the SL voltage in **Figure 2.10**(a) is determined by the capacitance network of the top and bottom gates, as it is shared with the SL voltage of the cell below. In the worst case, since the capacitance value of the top gate is larger than the bottom gate capacitance value, the SL voltage is formed at approximately 1/2 V_{pgm}. As shown in **Figure 2.12**, the top gate blocks the channel when the length (L_{tail}) of T_{tail} is 100 nm. Therefore, the voltage at SL has no impact on the channel.

On the other hand, when the L_{tail} is 30 nm, the voltage of SL affects the channel because the top gate cannot completely block the channel. In this case, the source side of the channel is not programmed by the gates but rather by the voltage transferred from the SL and the bottom gate voltage, resulting in lower efficiency. **Figure 2.13** illustrates the trapped charge in CSL for the two situations described above. When L_{tail} is long enough, the trapped charge is evenly distributed with a high concentration of $10^{19}/cm^3$ on the source and drain side of CSL. However, when L_{tail} is short, the voltage near the source of the channel is boosted, causing the charge near the source side to be distributed at approximately $10^{17}/cm^3$, which is significantly lower than that near the drain.

Figure 2.14(a) shows the program efficiency according to L_{tail}. When L_{tail} is 50 nm or longer, the channel is cut off, and the program efficiency improves rapidly. In

addition, the program efficiency of the target cell varies depending on the inhibit voltage. The influence of the source voltage on the channel is further reduced by applying 1/3 V_{pgm}. **Figure 2.14**(b) demonstrates that as the L_{tail} increases, the trapped charge increases. Therefore, there is a trade-off between program efficiency and the degree of integration for L_{tail}. The L_{tail} should be set to a minimum to design a more compact synapse array, while for programming/erasing at a low voltage, L_{tail} should be set to 50 nm or longer.

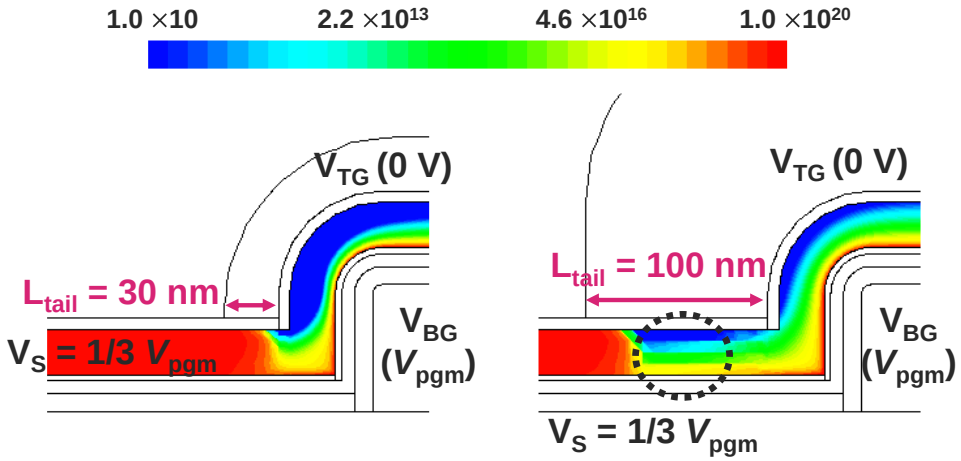


Figure 2.12. Channel electron density according to different L_{tail}.

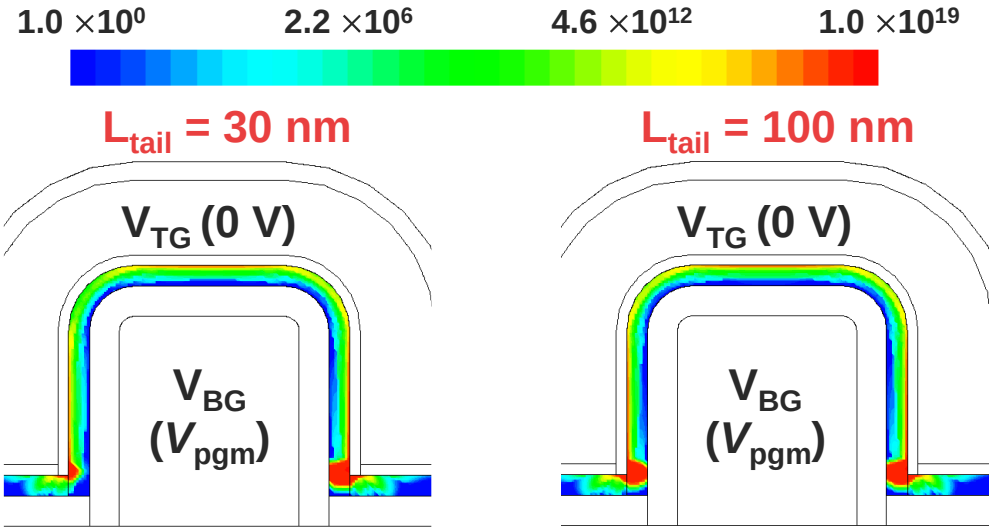


Figure 2.13. Trapped charges with different L_{tail} during the program.

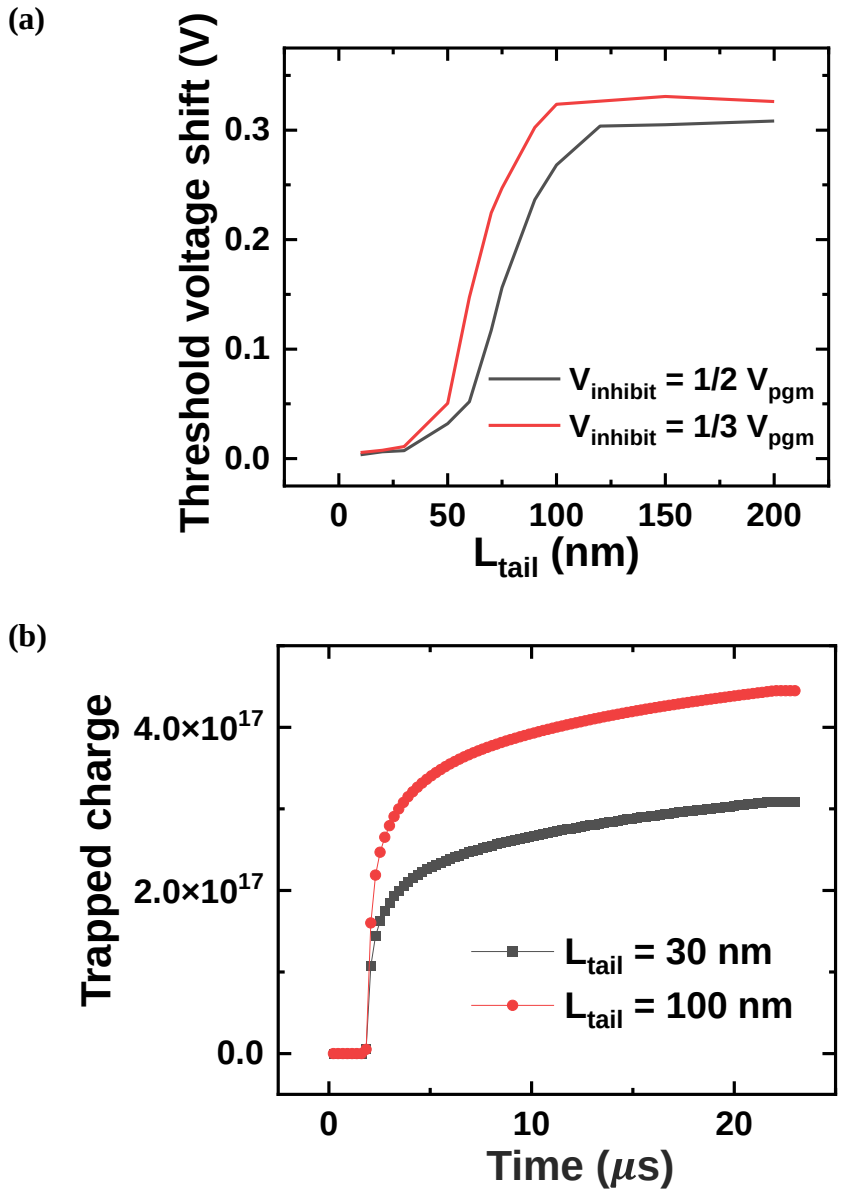


Figure 2.14. (a) Threshold voltage shift as a function of L_{tail} with different inhibit voltages. (b) Trapped charge vs. time.

Chapter 3

Fabricated Device Characteristics

3.1 Process Flow of Overpass Channel Synaptic Transistor

Figure 3.1 presents the complete process flow of the OCS transistor. Initially, a thick buried oxide layer is formed on the bare wafer using wet oxidation, followed by the deposition of 70 nm in-situ doped poly-Si serves as the bottom gate. A significant advantage of the OCS transistor is to achieve a large effective gate length and CSL area within a small footprint. Thus, a mix-and-match process is used with electron-beam lithography (EBL) and photolithography for sub-100 nm patterns. Bottom gate fins with widths ranging from 30 nm to 300 nm are patterned. **Figure 3.2** shows scanning electron microscope (SEM) images of bottom gate fins with 40 nm and 50

nm widths.

Subsequently, $SiO_2/Si_3N_4/SiO_2$ are deposited using low-pressure chemical vapor deposition (LPCVD) to serve as the CSL. During the initial stage of blocking oxide deposition, the corner regions of the fin are oxidized, resulting in rounded edges, as illustrated in **Figure 3.3**. Next, a thin amorphous silicon (a-Si) channel is deposited to reduce the current level and amplify the impact of trapped charges in CSL on the top gate's threshold voltage. The poly-Si active is also formed using mix-and-match lithography, as shown in **Figure 3.4**.

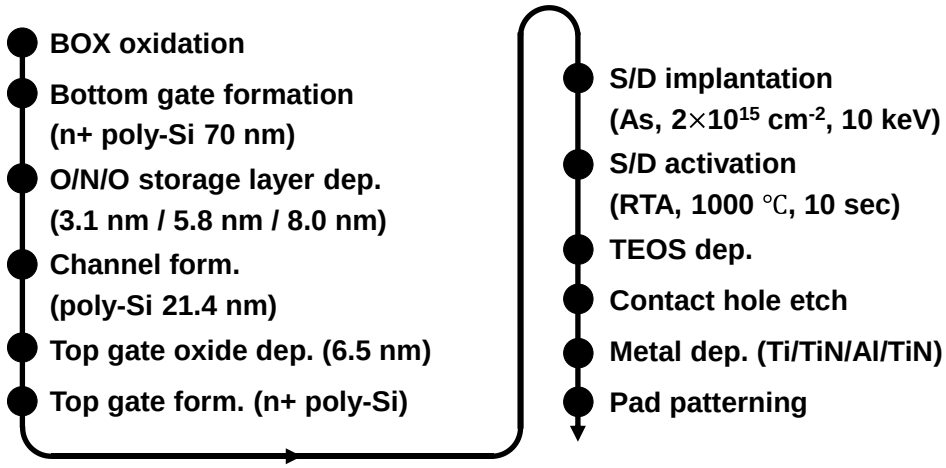


Figure 3.1. Entire process flow of OCS transistor with specific thickness.

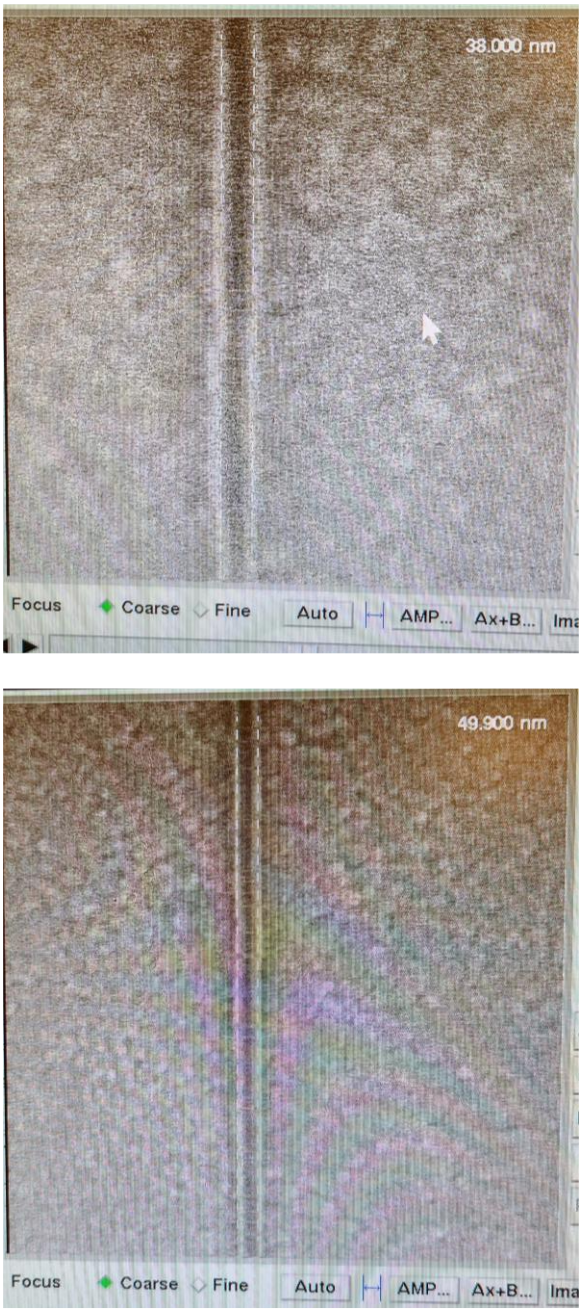


Figure 3.2. Scanning electron microscope (SEM) image of bottom gate fin.

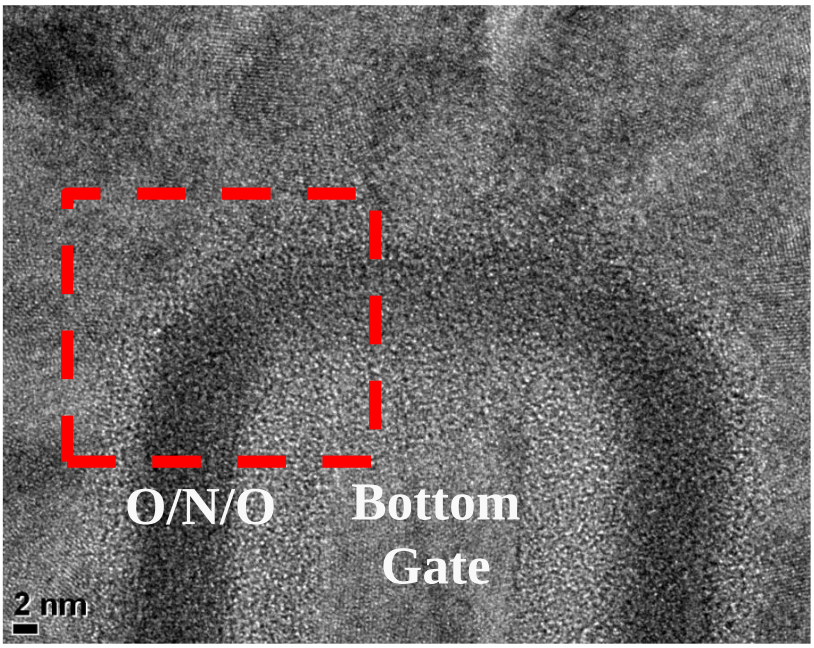


Figure 3.3. Transmission electron microscope (TEM) image of CSL layer.

Subsequently, gate oxide is formed using two different methods. The first method involves thermal oxidation at 1000 ℃ for 4 minutes, while the second method involves LPCVD to deposit 7 nm SiO_2. Typically, gate oxide is formed through oxidation for single crystalline silicon and through CVD for poly-Si. This is because SiO_2 formed by oxidation has the best film quality. However, when poly-Si is oxidized, the SiO_2 near the grain boundaries has poor film quality, which can cause a leakage current. Since the poly-Si is crystallized and gate oxide is formed simultaneously, the SiO_2 on poly-Si could be formed more stably.

The gate oxide thickness in this process is formed thicker than in conventional metal-oxide-semiconductor field-effect transistors (MOSFETs). One of the most critical aspects of fabricating the OCS transistor is that top gate etching must account for the step height caused by the bottom gate fin. When patterning the top gate through dry etch, it is essential to over-etch beyond the fin height, and this process can cause two problems. First, excessive etching causes the top gate oxide to be punctured, damaging the thin poly-Si active layer, as shown in **Figure 3.5** (a).

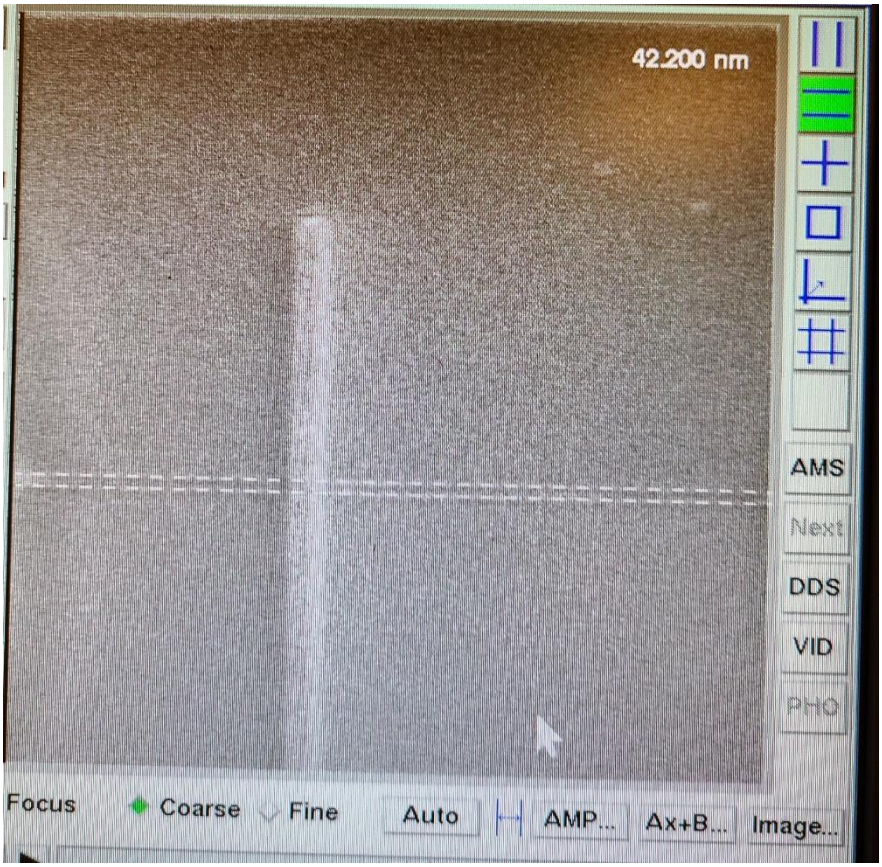


Figure 3.4. SEM image of poly-Si active.

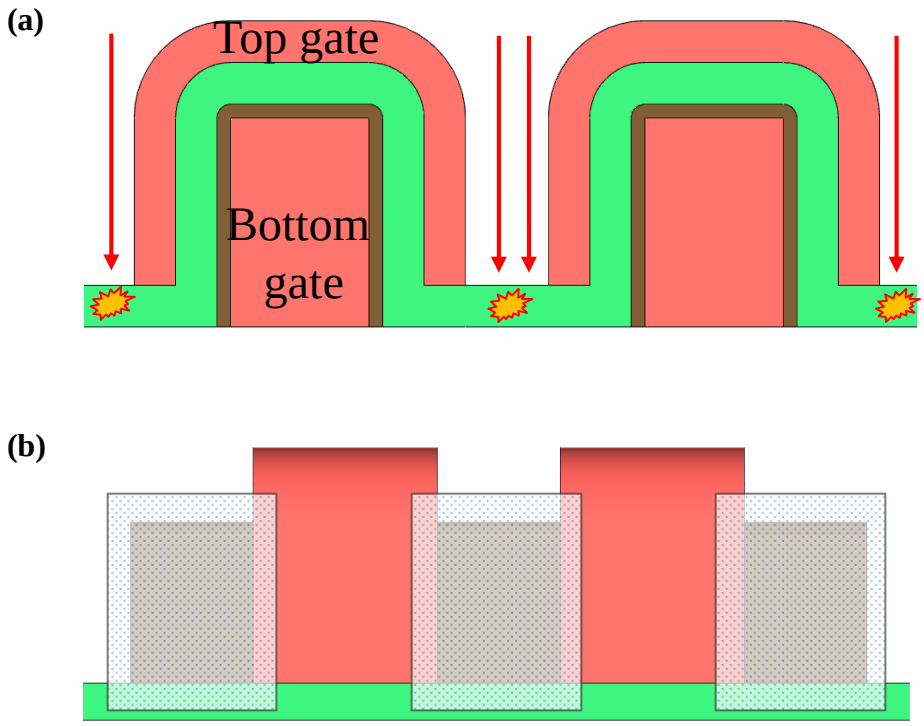


Figure 3.5. Two situations can arise during gate formation: (a) over-etch and (b) under-etch.

Secondly, if the over etch is insufficient, as shown in **Figure 3.5**(b), sidewalls are formed among the gate lines, which can cause a short circuit. As a result, signals from the pre-neuron are transmitted to all synaptic transistors through the top gate sidewalls. **Figure 3.6** shows the TEM image of the two situations described above during fabrication.

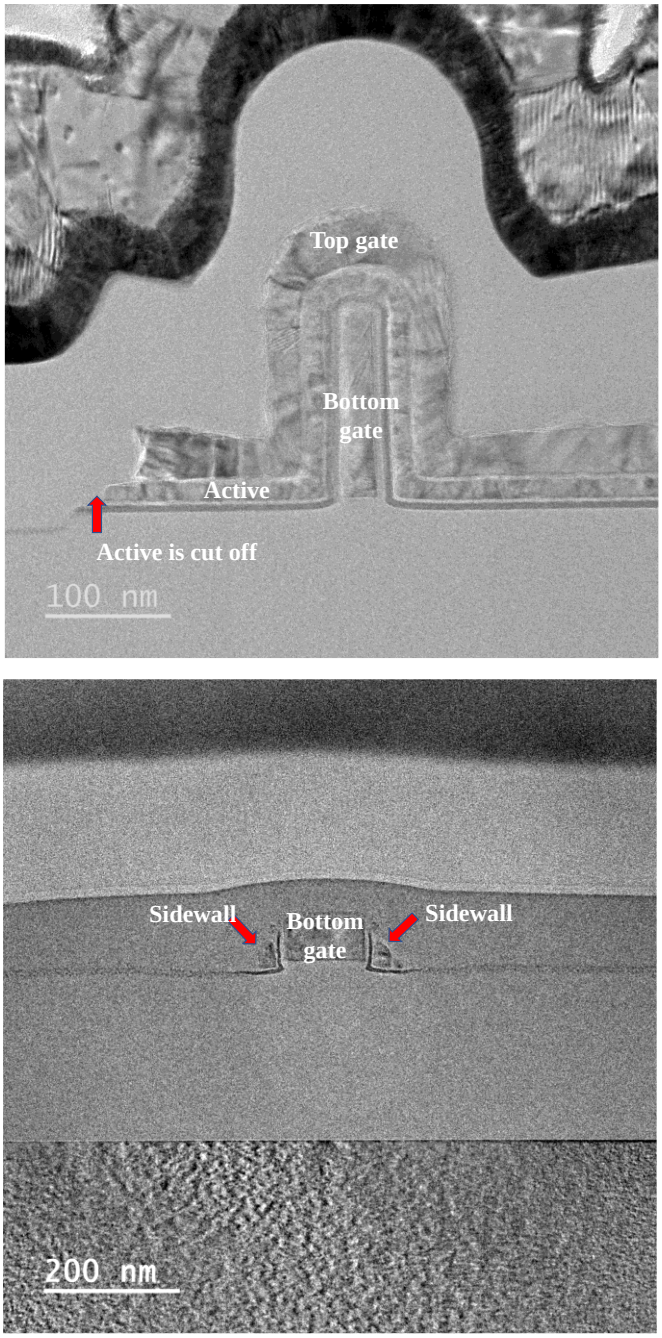
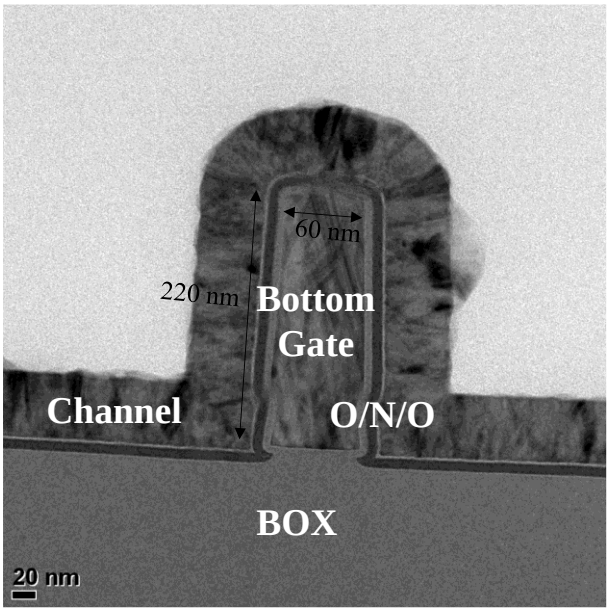


Figure 3.6. Two cases of TEM image during gate formation (a) active open (b) top gate line short.

The unit process is carried out with a chemical dry etcher (CDE), an isotropic etching technique to prevent the two situations mentioned above. As illustrated in **Figure 3.7**, because the etch rate is different depending on the grain of poly-Si and the top gate length is too short, isotropic etching poses multiple risks when fabricating scaled devices. Therefore, the gate oxide is formed with 7 nm, and the top gate with excellent selectivity is formed using a poly-Si etcher with HBr as the etching gas. This poly-etcher offers over 20 times the etch selectivity between poly-Si and SiO_2. Furthermore, since the OCS transistor has a structure in which the top gate surrounds the active region, precise alignment of the bottom gate fin and the top gate is essential. Misalignment within tens of nanometers is achieved by employing two global markers and four chip markers during e-beam patterning, as shown in **Figure 3.8**.

Subsequently, the source and drain are formed through arsenic ion implantation at 10 keV with a dose of 2×10^{15} cm^{-2}, followed by source and drain activation using rapid thermal annealing (RTA) at 1000 ℃ for 10 seconds. Finally, the Ti/TiN/Al/TiN layers are deposited using an Endura sputter, and the pads are patterned. **Figure 3.9** displays the fabricated OCS transistors and arrays under an optical microscope. The process flow for OCS transistors and arrays is identical, allowing the procedure to be carried out within a single wafer.

(a)



(b)

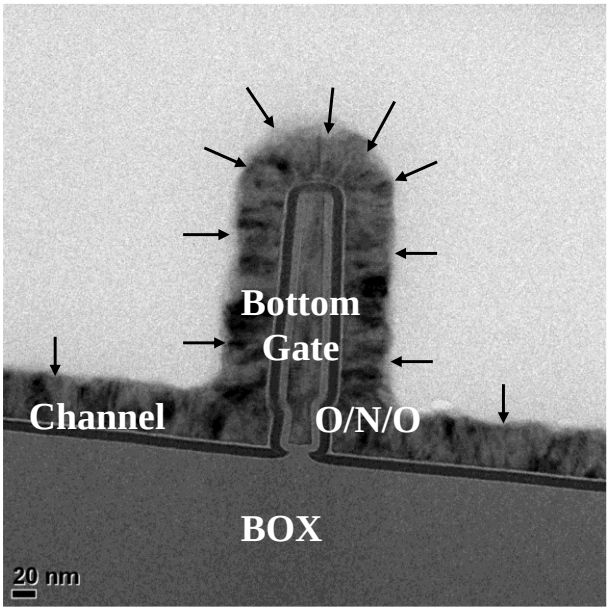


Figure 3.7. Front view of TEM image (a) before chemical dry etch (CDE) (b) after CDE.

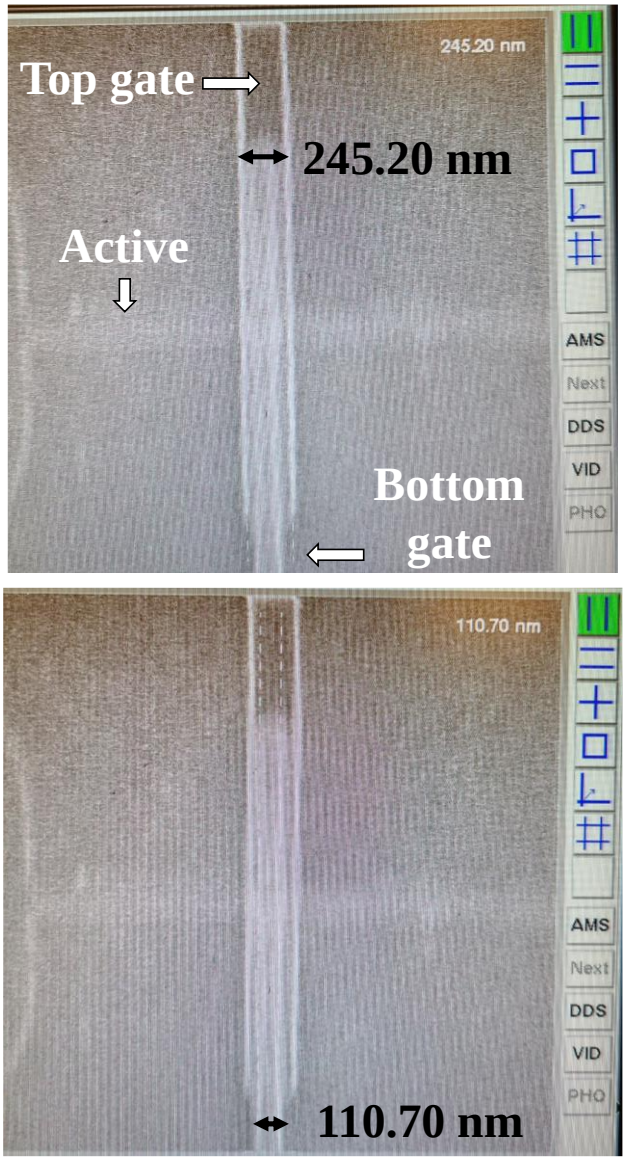
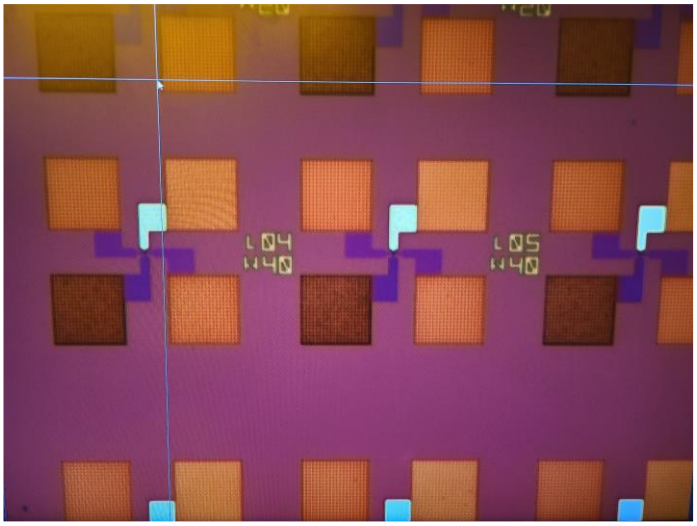


Figure 3.8. SEM image after top gate patterning. The top and bottom gates are well aligned, and the thin poly-Si active region is preserved after the formation of the top gate.

(a)



(b)

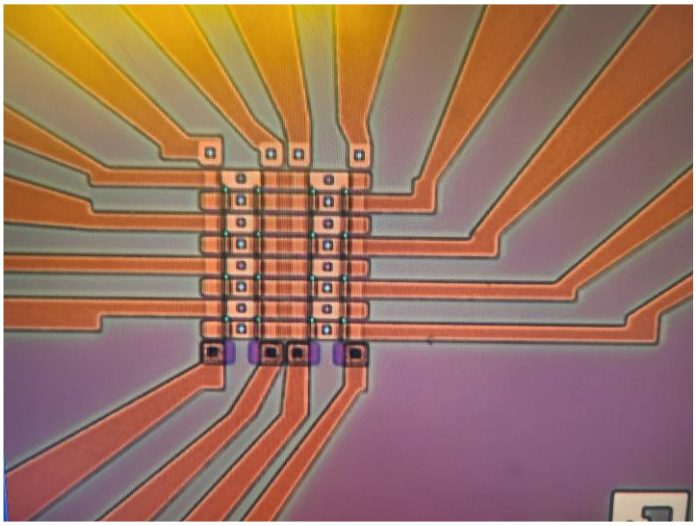


Figure 3.9. Bird's eye view of fabricated (a) OCS transistor, (b) OCS transistor array.

3.2 Unit Cell Characteristics

3.2.1 Electrical Characteristics

In **Chapter 2.2.1**, the synaptic current level is optimized considering the integrated system of synapses and neurons for ultra-low power operation. Therefore, the OCS transistor is fabricated with the goal of an on-current of several tens of nA, enabling low-power operation while minimizing propagation delay. Furthermore, as depicted in **Figure 3.2** and **Figure 3.10**, the electric field is reduced by over 20% due to the rounding of the fin corner.

Figure 3.11 shows the fabricated OCS transistor with the bottom gate length (L_{bg}) of 44.92 nm, top gate length (L_{tg}) of 289.30 nm, and active thickness (t_{act}) of 21.42 nm. The transfer curve of OCS transistors with L_{bg} = 80 nm, W = 40, 50, and 80 nm are shown in **Figure 3.12**(a). The on-current is several tens of nA, suitable for ultra-low power operation at a V_{read} of 3V. A positive bottom gate voltage corresponds to the erase state, where holes are trapped in the CSL, while a negative bottom gate voltage corresponds to the programmed state, where electrons are trapped, as depicted in **Figure 3.12**(b).

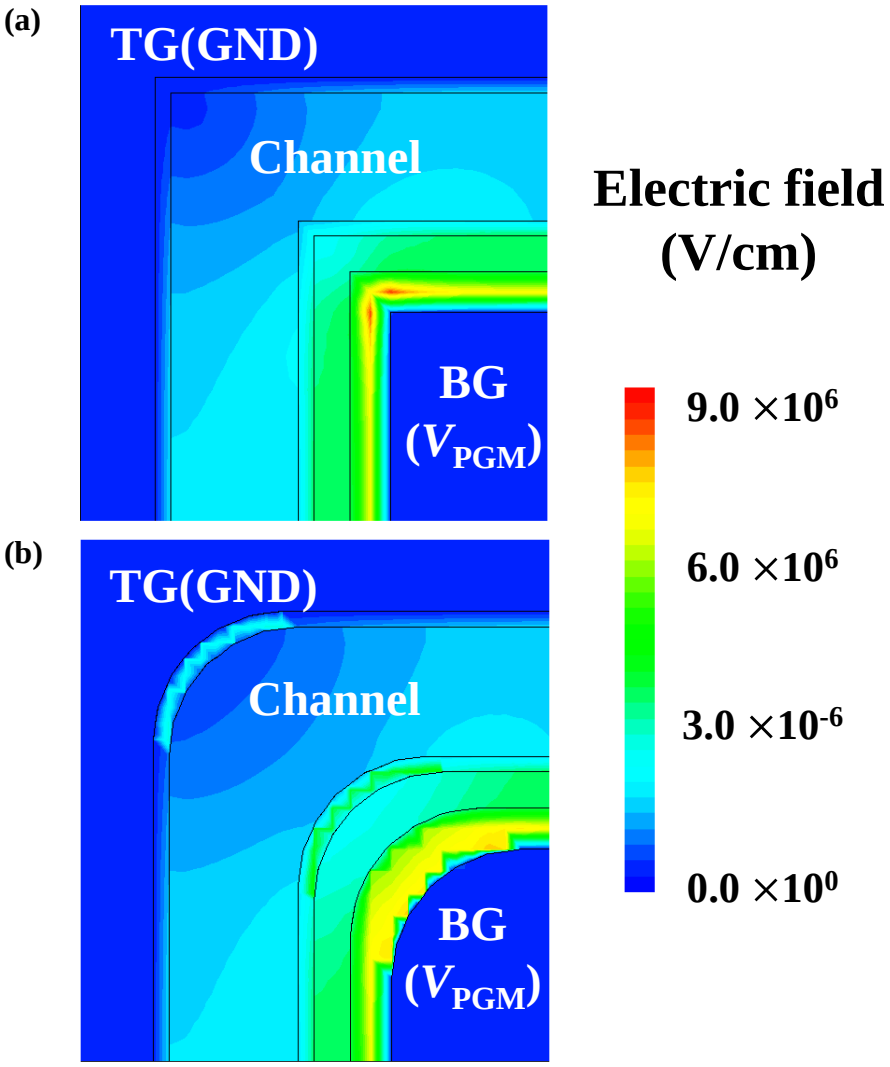


Figure 3.10. The electric field of (a) sharp corner and (b) rounded corner during the program.

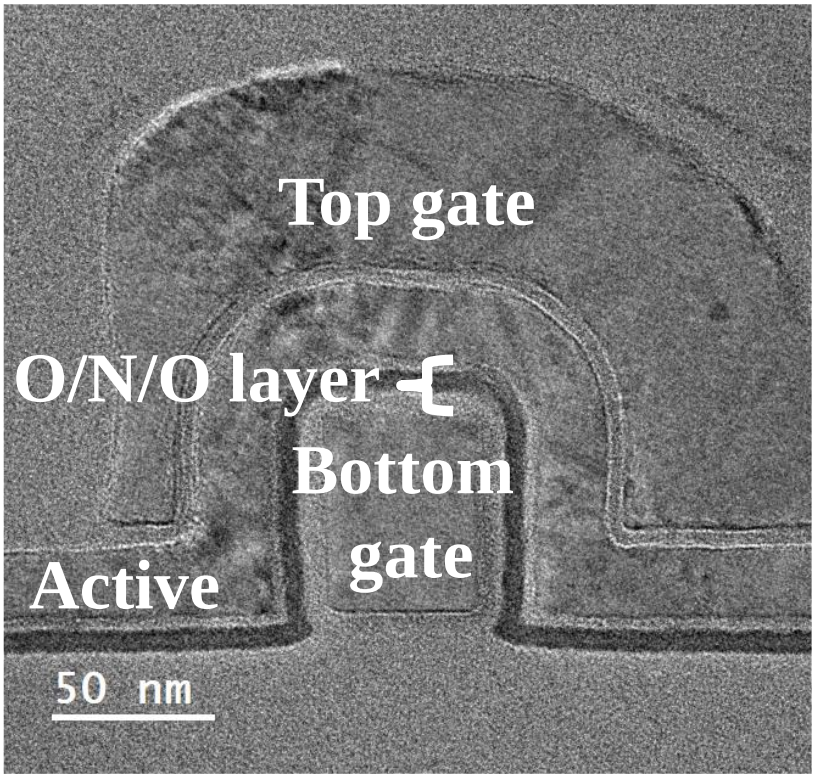


Figure 3.11. TEM image of OCS transistor. The height of the bottom gate fin can increase the effective gate length and the volume of the CSL.

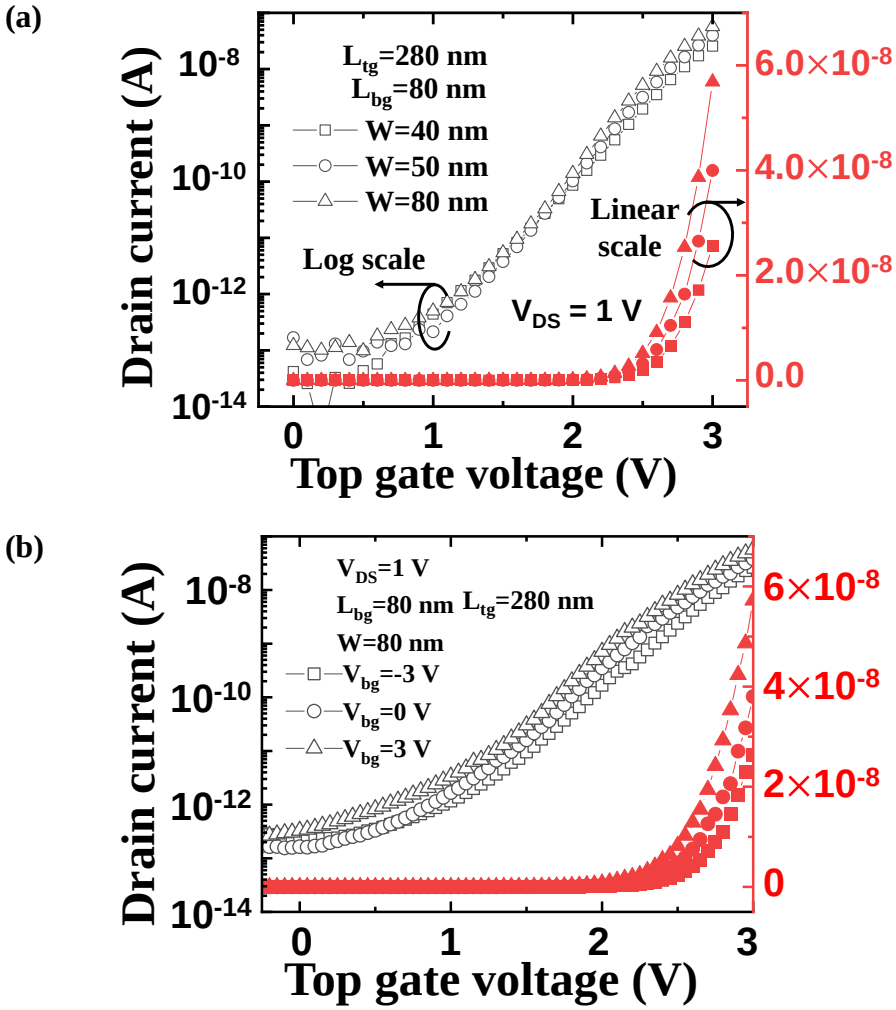


Figure 3.12. Transfer curve of OCS transistor with (a) different active width and (b) bottom gate voltage.

Figure 3.13(a) shows the output curve of the OCS transistor with specific parameters. When the drain voltage exceeds the top gate voltage (V_{TG}), impact ionization occurs due to the strong electric field on the drain side. During this process, hot electron-hole pairs are generated, and these electrons move toward the drain. At the same time, generated holes are trapped at the grain boundary, and these holes cause an increase in current, which is referred to as the kink effect. Thus, when the impact ionization occurs, the current rapidly increases, and the hot carriers can damage the gate oxide film. Therefore, a large drain voltage is not applied to this system.

The bottom gate affects only a part of the channel, preventing the current flow as shown in **Figure 3.13**(b). Therefore, the signal from the pre-neuron must be transferred to the top gate during the inference operation, while the bottom gate is involved in the PGM/ERS operation. **Figure 3.14** illustrates the extremely-low-power operation of the fabricated OCS transistor and the comparison data for the power consumption of various synaptic devices [45-52].

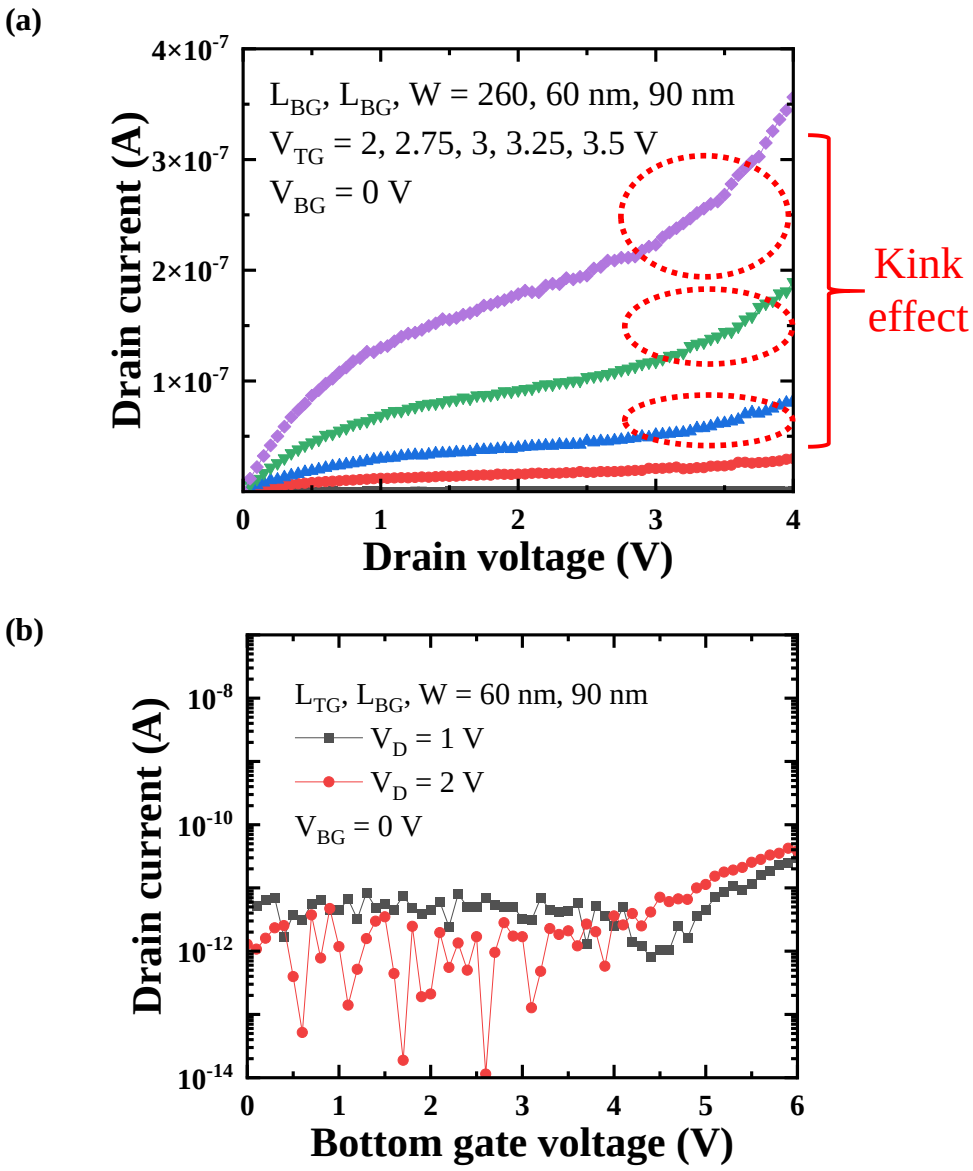


Figure 3.13. (a) Output curve of OCS transistor according to different top gate voltage. (b) Current as a function of bottom gate voltage.

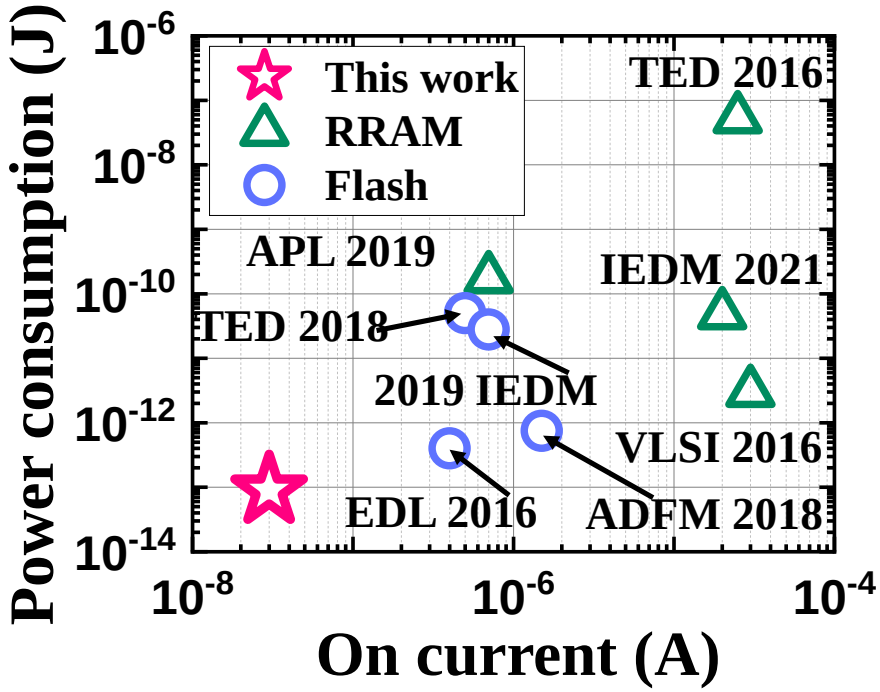


Figure 3.14. Benchmark of power consumption as a function of on current.

3.2.2 Weight Modulation

The PGM/ERS scheme of the OCS transistor is conducted as described in **Table 2.2**. **Figure 3.15** shows the transfer curve of the smallest OCS transistor after applying the PGM/ERS pulses. Since the source and drain are floating, the top gate and bottom gate voltages boost the channel voltage. Threshold voltage changes from 3.320 V to 3.431 V while applying PGM pulses and from 3.432 V to 3.312 V while applying ERS pulses. Consequently, the boosted channel voltage and bottom gate voltage create the FN tunneling condition in the tunneling oxide, allowing the conductance of the target cell to be changed.

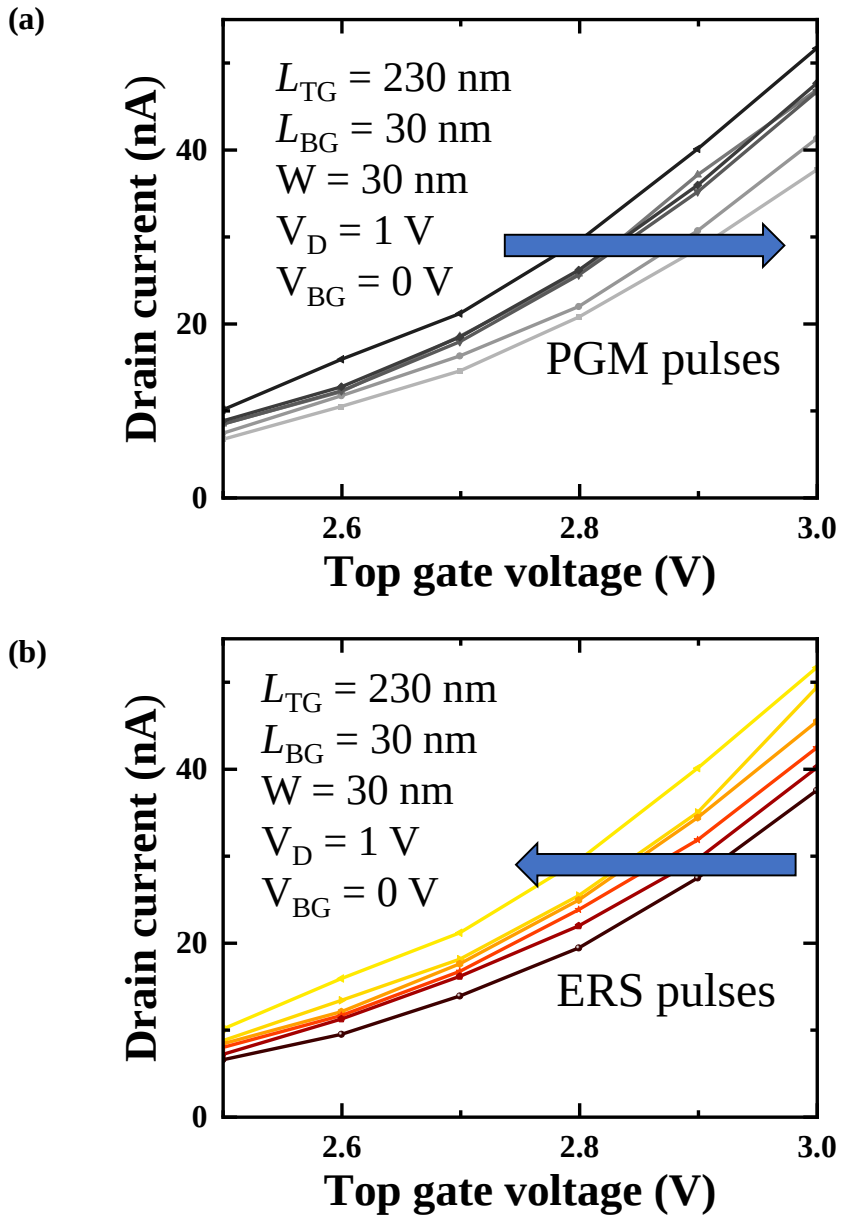


Figure 3.15. Transfer curve of aggressively scaled OCS transistor with (a) program state and (b) erase state.

3.3 Synapse Array Characteristics

3.3.1 Inference

In the synapse array, the current sum is the product of the VMM operation, which is crucial in neuromorphic systems. To verify the VMM operation, two cases are assumed: the turn-on of all transistors and a single transistor. The current sum difference is measured, and the current sum error (CSE) is calculated below Eq. (3.1):

$$\mathrm{CSE} = (\Sigma I_{\mathrm{i}} - I_{\mathrm{all\_select}})/I_{\mathrm{all\_select}} \times 100\ (\%) \tag{3.1}$$

where I_{i} and $I_{\mathrm{all\_select}}$ indicate the turn-on current of one and all transistors, respectively. **Figure 3.16** shows the individual I_{i} and $I_{\mathrm{all\_select}}$ for each of the four SLs in the fabricated 4×4 OCS array. The closer the CSE is to 0, the VMM is ideally performed. By comparing the CSE at various read voltages (V_{read}), the V_{read} can be optimized. The synapse array is measured using the Keysight B1500A instrument and the medium power source/monitor unit (MPSMU).

As the V_{read} increases, a reduction in the device's channel resistance is noted, which amplifies the error originating from the line resistance, as shown in **Figure 3.17**. For V_{read} values exceeding 3.5 V, the CSE is greater than 2%. As a result, the V_{read} is adjusted to 3 V, where the average CSE is 0.79%, and it can minimize the

degradation of VMM operation in the neural network. Finally, **Figure 3.18** shows the accurate VMM operation with no difference between ΣI_i and I_{all_select}.

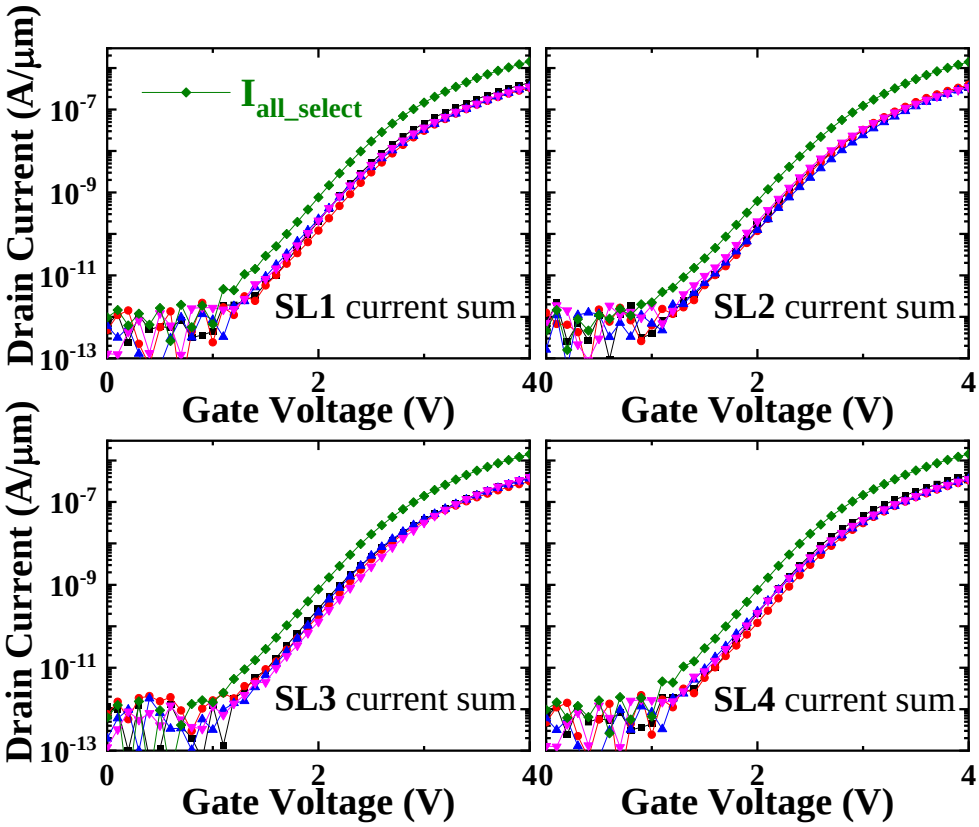


Figure 3.16. Current sum for each of the 4 source lines according to read voltage.

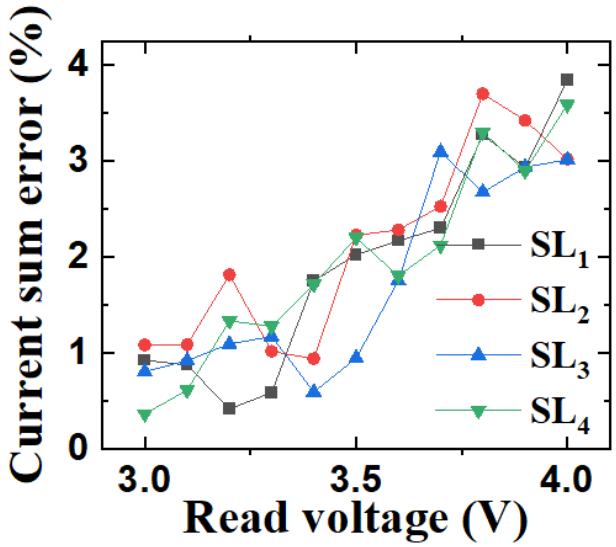


Figure 3.17. Current sum error (CSE) for each of the 4 SLs as a function of the read voltage.

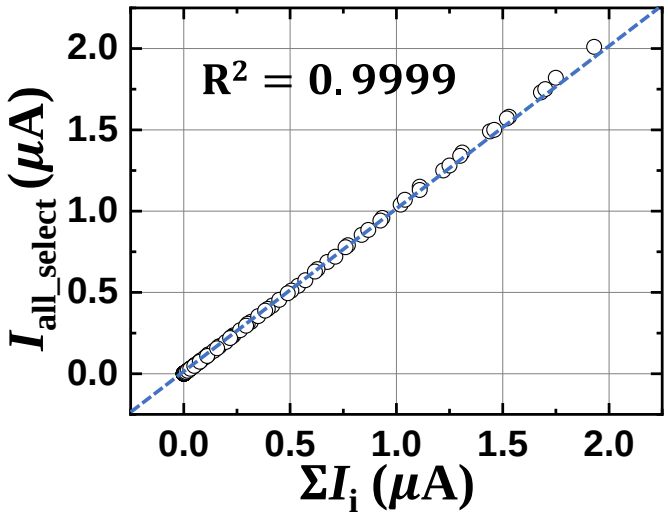


Figure 3.18. VMM operation with respect to the input current. The dotted line represents a perfect VMM operation without any errors.

3.4 Diode-Connected Synapse Array

3.4.1 Fabrication Flow

The 4-terminal NOR-type asymmetric dual gate synaptic array has more terminals than a conventional NOR-type array, which provides an advantage in controlling conductance. However, it also has the disadvantage of a larger cell size. In the 4-terminal NOR-type array, the top gate line (TGL) and DL are parallel, resulting in a vertical dimension of 6 minimum feature sizes (F) and a horizontal dimension of 5 F as shown in **Figure 3.19**(a). Therefore, it is limited to fabricate a high-density synapse array because the cell size is three times as large as that of a conventional NOR-type array.

For this reason, we propose a diode-connected synapse array that shares two parallel TL and DL, as shown in **Figure 3.19**(b). This allows for a much more compact integration with horizontal and vertical feature sizes of 4 F and 3 F, respectively. In other words, it is possible to scale down the cell size to 12 F^2 by forming a single contact hole for the four terminals of the top gate and drain contact of two adjacent cells, as illustrated in **Figure 3.19**(c).

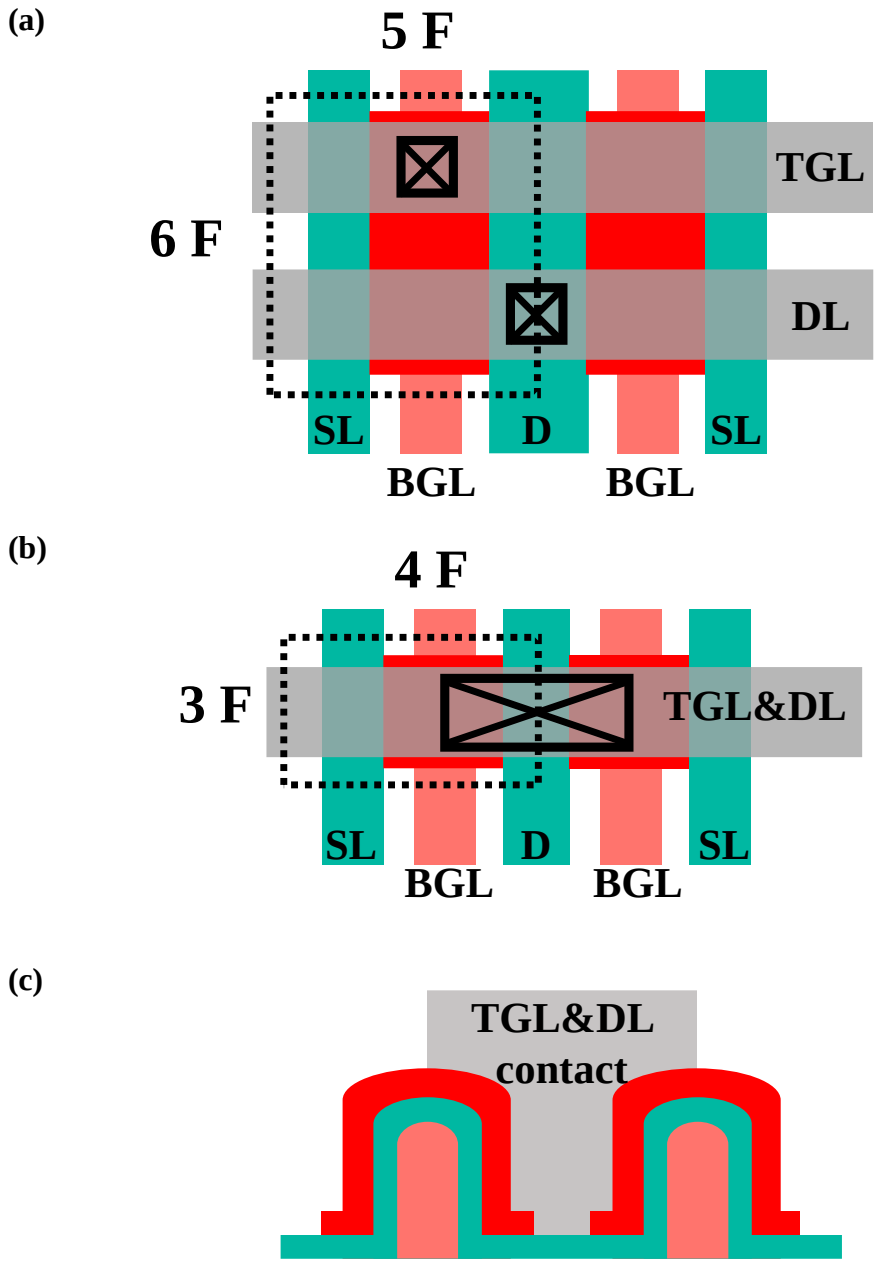
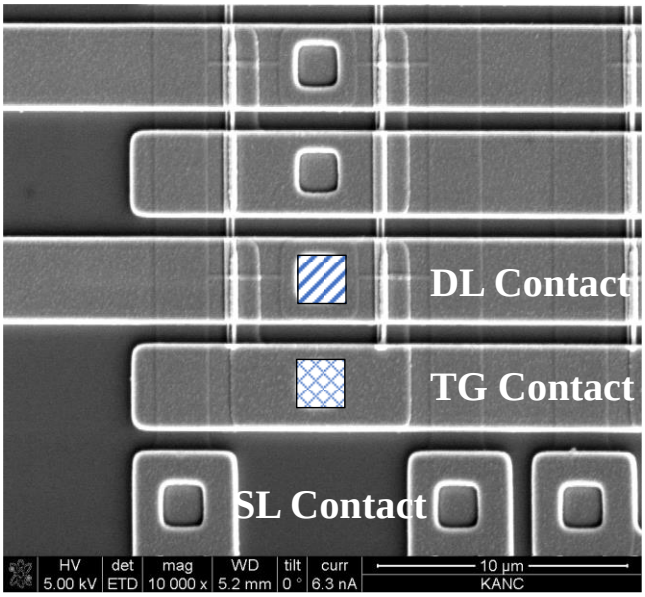


Figure 3.19. Top view of (a) conventional 4-terminal NOR-type synapse array and (b) diode-connected(D-C) synapse array. (c) Front view of D-C synapse array.

The SEM images of the two types of fabricated synapse arrays show that the cell size is scaled down to less than half by integrating TGL and DL, as shown in **Figure 3. 20**.

(a)



(b)

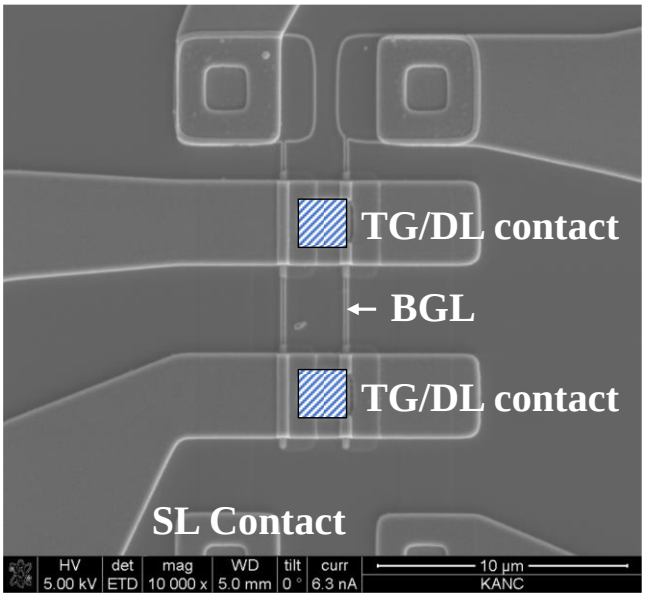


Figure 3. 20. SEM image of (a) conventional 4-terminal NOR-type synapse array and (b) D-C synapse array.

3.4.2 Inference

V_{DD} is always applied to the DL in the AND-type array for the event-driven operation. Therefore, even in the standby state when no signal is input from the pre-neuron, leakage current constantly flows to the SL, causing errors in the VMM operation. However, TGL and DL are set to 0 V in the standby state in the proposed D-C array, enabling double suppression of leakage current. **Figure 3. 21** shows that when V_{DD} is constantly applied to the drain, the off-current flows at hundreds of pA due to gate-induced drain leakage (GIDL). In contrast, the D-C array has an extremely low current level at the sub-pA level. Therefore, there is no impact from leakage current, considering the measurement equipment's error. In other words, the D-C array is suitable for low-power and accurate VMM operation, with the on-current remaining the same while reducing the off-current by more than three orders of magnitude.

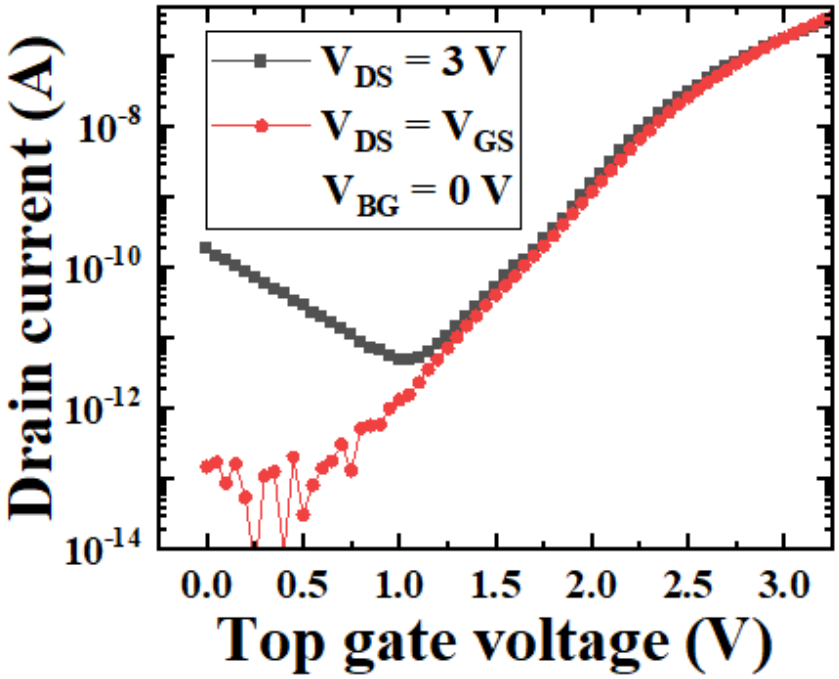


Figure 3. 21. Drain current vs. top gate voltage. Both of two cases allow for event-driven operation. However, the power consumption in the standby state is significantly lower for the D-C array.

3.4.3 Weight Modulation

As described in section 3.4.2, the D-C array has several advantages during inference, such as reduced cell size and leakage current. **Figure 3. 22** shows the PGM/ERS voltage scheme in the simple 2×2 array. The difference from the scheme described in the previous section 2.2.3, is that the DL and TGL are integrated, and the same voltage is applied. Since the L_{tail} separates the SL/DL voltage from the channel voltage, the overall PGM/ERS mechanism does not change significantly. Moreover, PGM/ERS disturb can be prevented by applying PGM/ERS inhibit voltage to the TGL, DL, and BGL.

Figure 3. 23(a) shows the scheme for controlling the conductance of the target cell in the fabricated OCS array, where $t_{r,f}$ and t_{pw} represent the rising, falling time, and pulse width, respectively. 40 ISPP and ISPE pulses are applied at intervals of 0.1 V from 5 V to 9 V in the PGM cycle and intervals of 0.2 V from 14 V to 22 V in the ERS cycle. Three PGM/ERS cycles are performed, allowing the conductance of the target cell to be adjusted at sub-nA intervals. Total 240 pulses are applied to verify the PGM/ERS operation.

To verify the inhibition of neighboring cells when applying PGM/ERS pulses, the percentage of cell inhibition (PCI) is defined as follows:

$$\text{Percentage of cell inhibition} = \left| \frac{I_{surrounding}}{I_{target}} \times 100 \right| \quad (3.1)$$

The lower the PCI value in each PGM/ERS cycle, the more precise conductance of the target cell is controlled without changing the conductance of neighboring cells. The PCI in the first PGM cycle is 14.285%, but it stabilizes at 4.424% and 3.633 % in the second and third cycles, respectively. **Figure 3.25**(a) shows the voltage conditions during the program, with the PGM inhibit voltage applied to the top gate and drain of the PGM disturb cell. According to Gauss's law, the electric field is concentrated at the circular region indicated by the dashed line, as illustrated in **Figure 3.25**(b). Charges are trapped at this location through tunneling, as shown in **Figure 3.25**(c). In subsequent cycles, some of these charges act as fixed charges, leading to improved PCI.

Figure 3. 26 shows the distribution of pristine current values measured from 144 cells with nine different dies. While 96.52% of cell current is distributed within the range of 20 nA to 50 nA, the charge in the CSL has little effect on the top gate. Therefore, as analyzed in **Figure 2.9**(b), a reduction in body thickness to 10 nm increases the weight modulation range by 5.63 times, and all cells can be adjusted to specific weights.

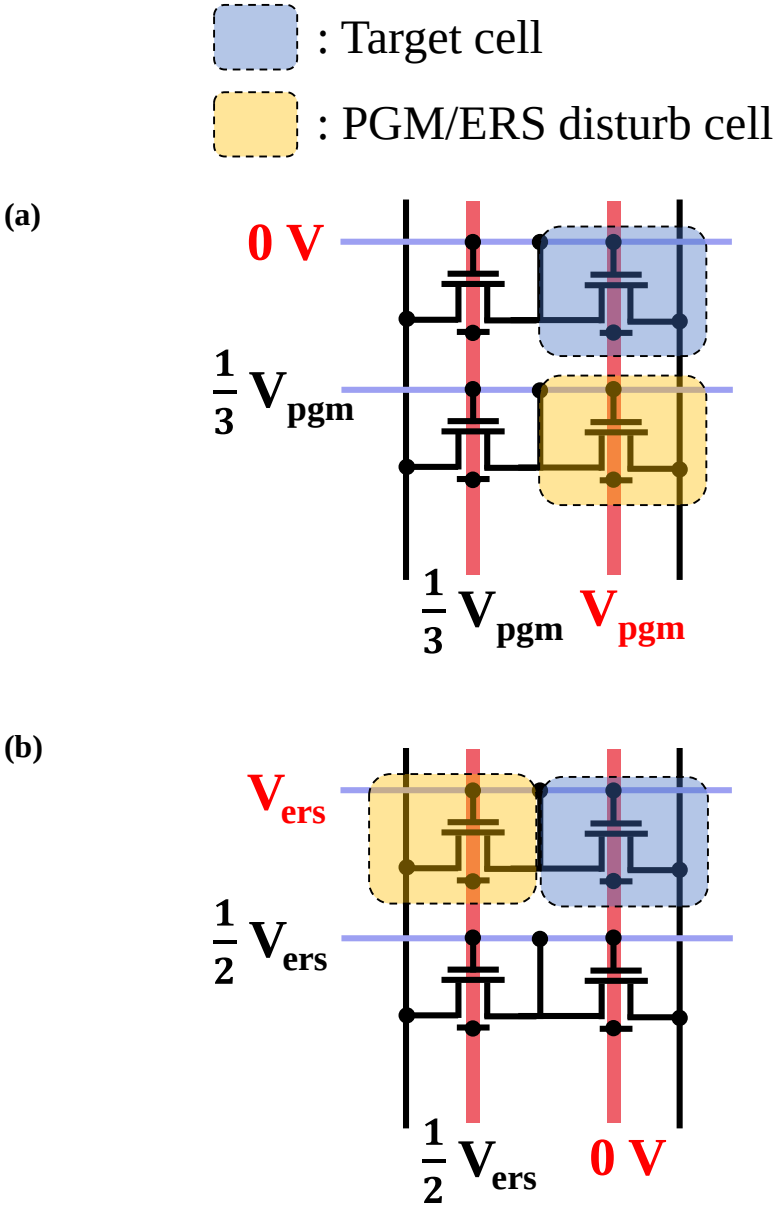
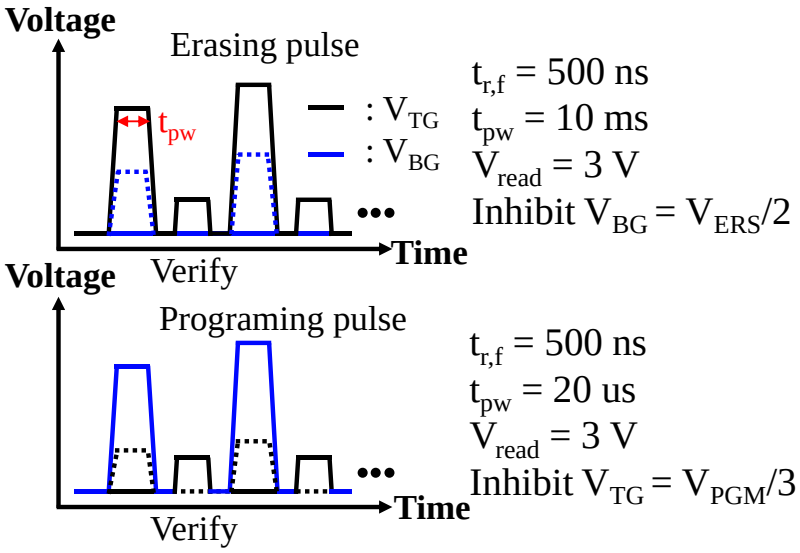


Figure 3. 22. Equivalent circuit of 2×2 diode-connected NOR-type array and voltage conditions of the target and surrounding cells.

(a)



(b)

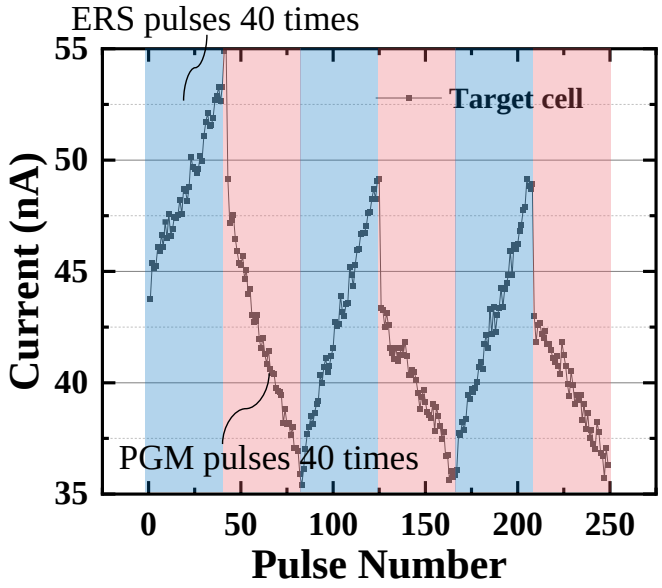


Figure 3. 23. (a) Utilizing the ISPP/ISPE scheme, the current is verified after each pulse. (b) The conductance change of the target cell is observed in each of the three PGM/ERS cycles.

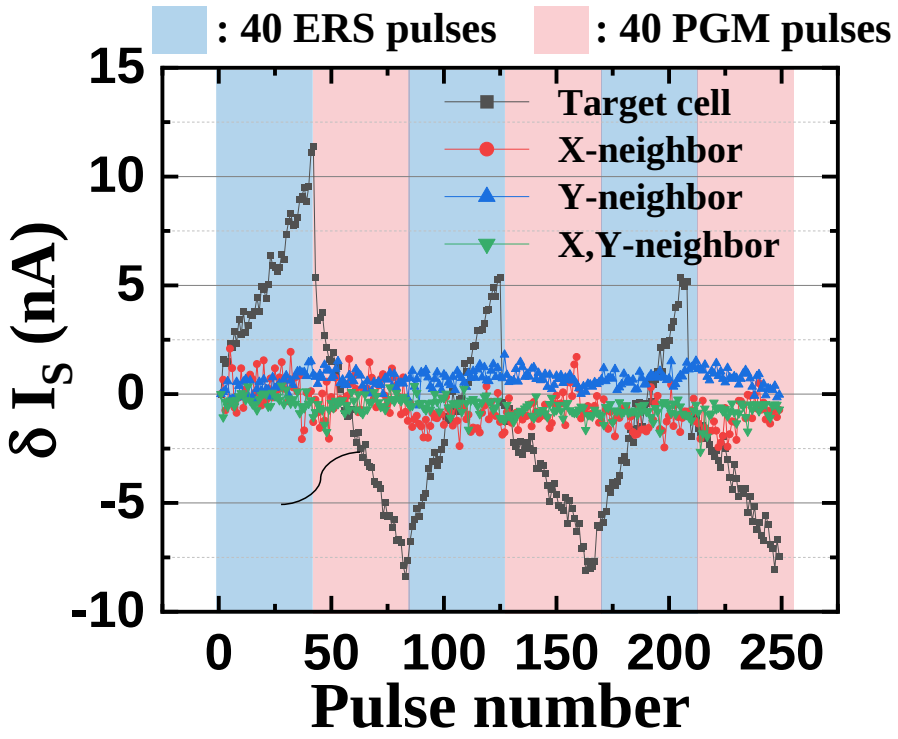


Figure 3. 24. The synaptic weight change of the target and neighboring cells as a function of PGM/ERS pulses.

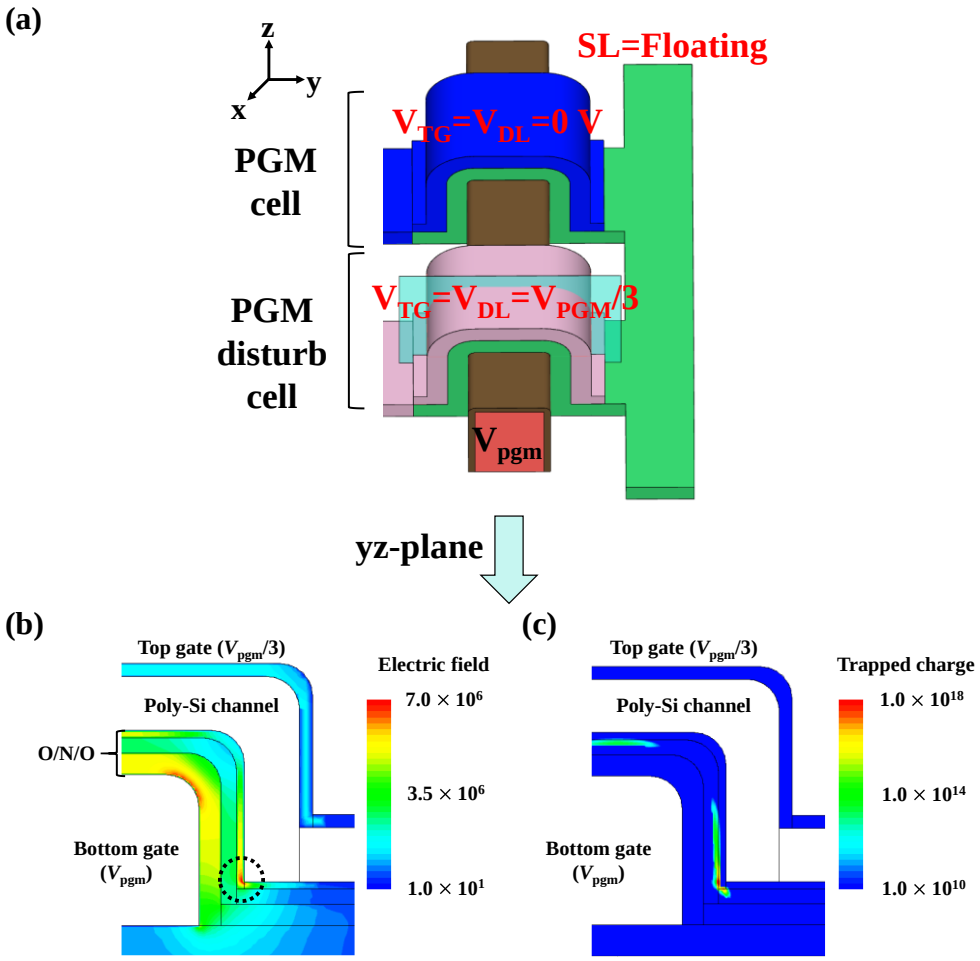


Figure 3.25. (a) Schematic of OCS array during the program. (b) The electric field and (c) trapped charge of the PGM disturb cell.

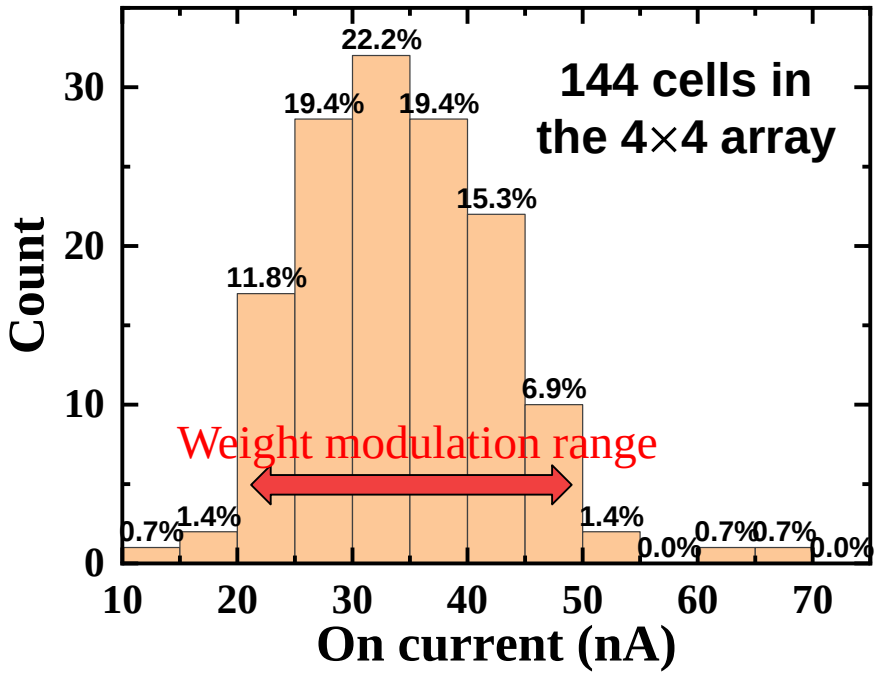


Figure 3. 26. Current distribution of 144 cells.

3.4.4 PVT Variations

It is essential to verify process-voltage-temperature (PVT) variations to ensure the stable operation of the synapse array. The proposed OCS transistor can reduce the impact of process variation compared to other scaled devices because of the increased effective gate length and volume of the CSL. Furthermore, even if the cell current differs due to the process variation, it can be adjusted by the fine-tuning of weights, as demonstrated in **Figure 3. 24**.

Subsequently, to analyze the influence exerted on the OCS transistor by input voltage variations, the voltage accuracy of the Keysight B1500A and the high voltage semiconductor pulse generator unit (HV-SPGU) are considered. The amplitude accuracy of HV-SPGU is $\pm$(0.5% + 50 mV). Thus, given that V_{read} is 3 V, it can vary up to 65 mV. Therefore, the current variation due to the voltage for various synaptic weights is measured as depicted in **Figure 3. 27**. It changes linearly from 2.9 V to 3.1 V with an average $\triangle I/\triangle V=1.152\times10^{-7}$ A/V. The effect of a 65 mV variation on current is illustrated in **Figure 3. 28** with different current levels. Then, high-level simulations are conducted in Chapter 4.2 with these results.

Finally, the influence of temperature variations on synaptic current is measured in **Figure 3. 29**. The temperature is varied from 27 ℃ to 95 ℃ with 12 different weights. The current of the poly-Si channel device increases linearly as the

temperature increases as electrons in the grain boundary are emitted [53]. Therefore, if a device that discharges the membrane capacitance is configured as a synaptic device, the effect of the temperature variation can be minimized, as shown in **Figure 3. 30** [54]. As the current of the synapse array increases at high temperatures, the current of the synaptic discharge also increases. Therefore, this parallel increase facilitates the maintenance of the firing rate as depicted in **Figure 3. 31**.

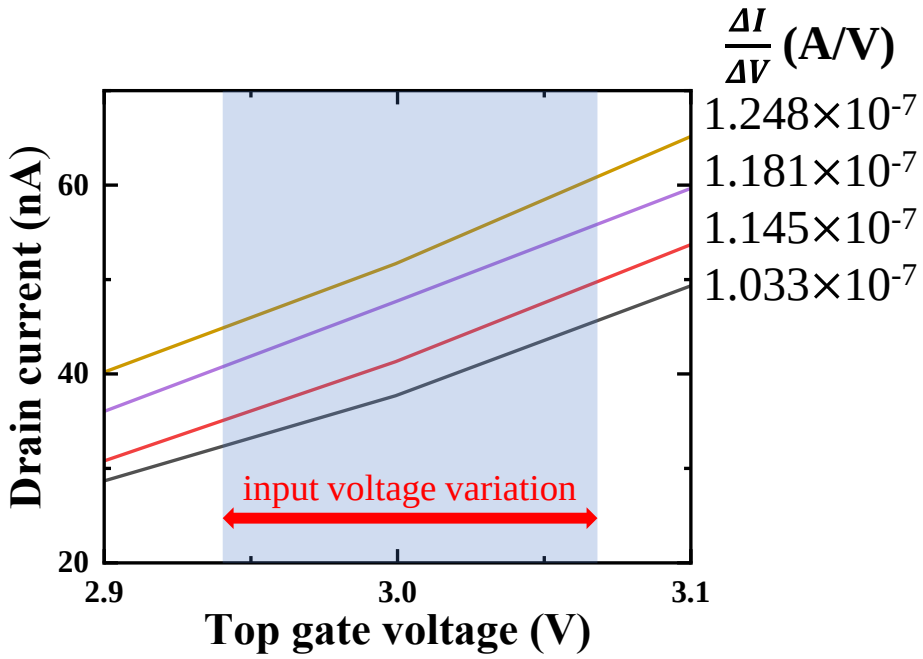


Figure 3. 27. Drain current as a function of top gate voltage.

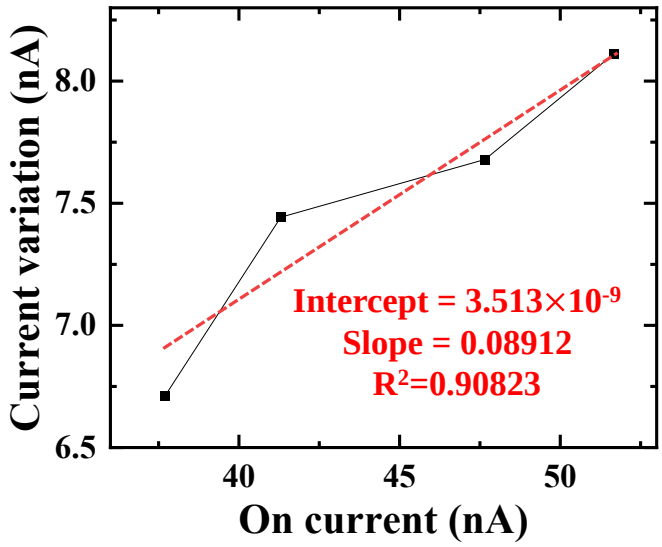


Figure 3. 28. Current variation as a function of the on current.

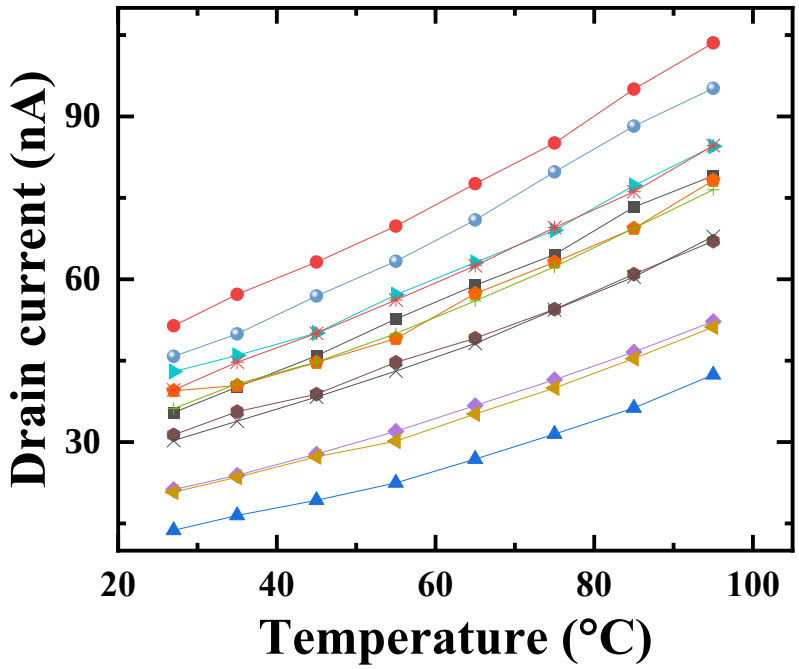


Figure 3. 29. Drain current vs. temperature. As the temperature increases, the current increases linearly.

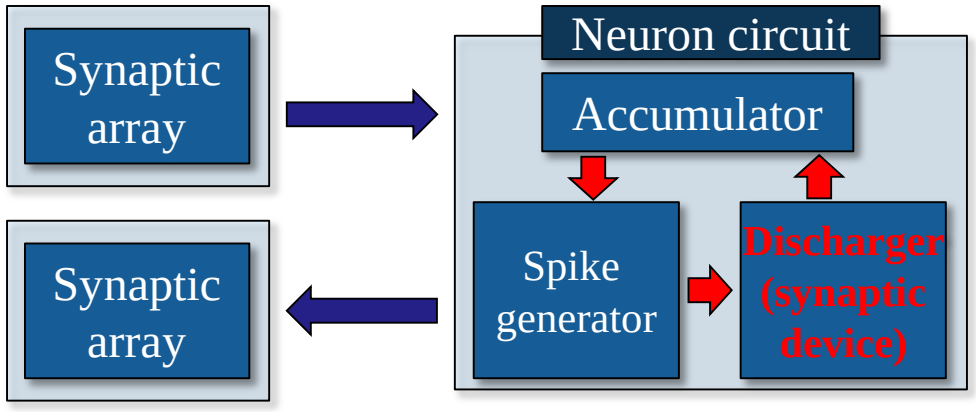


Figure 3. 30. Schematic of neuron circuit for compensating for changing synaptic properties.

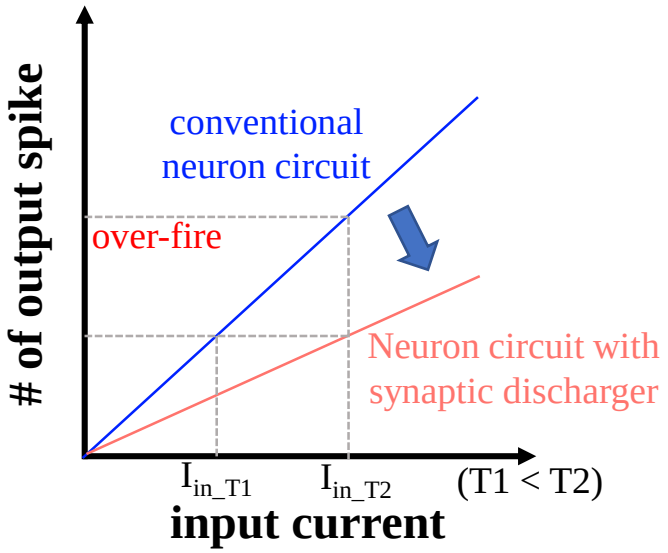


Figure 3. 31. A number of spikes does not differ with different temperatures through the utilization of the synaptic discharger.

Chapter 4
Hardware Demonstration of Artificial Synaptic Array

This chapter discuss the potential VMM errors that can arise in synapse arrays. The main factors that can lead to VMM errors include leakage current, multi-bit error, conductance loss due to retention and etc. These problems can be addressed at both the synaptic array and system levels.

4.1 Revised Bias Scheme

In the fabricated synapse array, ion-implanted poly-Si is used as the SL instead of a metal line, which results in a high line resistance. Moreover, the body thickness is thin to increase the PGM/ERS efficiency, increasing the line resistance. Therefore, this situation can lead to errors in VMM operation. As shown in **Figure 4.1**, the

resistance network of the synapse array can be represented by the line resistance (R_L) and the channel resistance (R_x) of the individual OCS transistor. Kirchoff's Current Law derives the equations for individually turning on device and turning on two devices and the equations are as follows [55]:

$$I_a = \frac{1}{R_a+4R_L} = 244 \text{ nA} \quad (4.1)$$

$$I_b = \frac{1}{R_b+3R_L} = 358 \text{ nA} \quad (4.2)$$

$$I_c = \frac{1}{R_c+2R_L} = 325 \text{ nA} \quad (4.3)$$

$$I_d = \frac{1}{R_d+R_L} = 327 \text{ nA} \quad (4.4)$$

$$I_{a+b} = \frac{1}{(R_a+R_L)//(R_b+3R_L)} = 637 \text{ nA} \quad (4.5)$$

When V_{read} is 3 V or lower, the channel resistance is too large to extract the resistance using the above equations. Therefore, V_{read} is set to 4 V to reduce the channel resistance further. R_L is 38.654 kΩ, and the respective resistance values are R_a=4.306 MΩ, R_b=2.678 MΩ, R_c=2.998 MΩ and R_d=3.084 MΩ, which are roughly 100 times R_L. Therefore, the influence of R_L is reduced by setting V_{read} to 3 V, allowing for a reduction in the VMM error.

Another cause of VMM error is the leakage current in the synaptic array, which is always summed at the post-neuron even when there is no input signal. Therefore,

the device-level solution to this problem is to use a synaptic device with a large on/off ratio or D-C array, as described in Chapter 3.4 to minimize the leakage current.

The learning algorithm and related computations are often carried out on a separate, more powerful computer or processing unit in off-chip learning. The neuromorphic system receives the parameters, such as synaptic weights, from the external device and transfers them to the synaptic devices. Thus, when transferring the synaptic weights, both the on- and off-current of each synaptic device are determined. The current sum is calculated as the sum of the product of the voltage input and synaptic weight plus a bias, as shown in the following equation (4.6):

$$\Sigma I = \Sigma V_i W_i + \mathrm{b} \tag{4.6}$$

$$\Sigma I = \Sigma V_i W_i + \mathrm{L} + (\mathrm{b}\text{-}\mathrm{L}) \tag{4.7}$$

$$\Sigma I = \Sigma V_i W_i + \mathrm{L} + \mathrm{b}` \tag{4.8}$$

As shown in

Figure 4.2, when V_{DD} is always applied to the drain, a constant leakage current (L) flows to SL, and it can be expressed by equation (4.7). Therefore, errors due to leakage current are eliminated by transferring the revised bias (b`), obtained by subtracting the leakage current sum value from the bias, rather than transferring the bias value calculated by the external processing unit. However, even if the value of

the leakage current increases, the effect of the leakage current can be removed by adjusting the value of b`. Nevertheless, there is no benefit in power consumption.

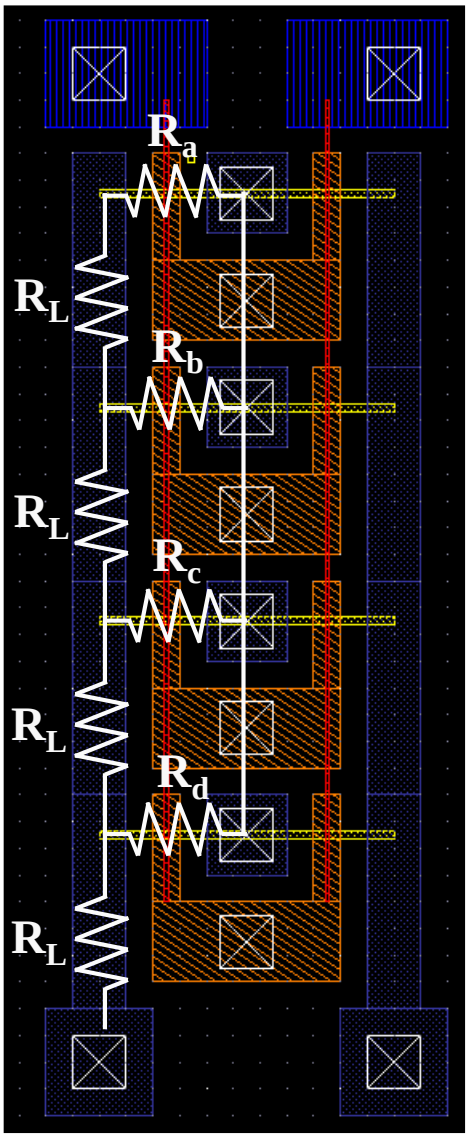


Figure 4.1. Equivalent circuit of 4×4 synapse array with layout.

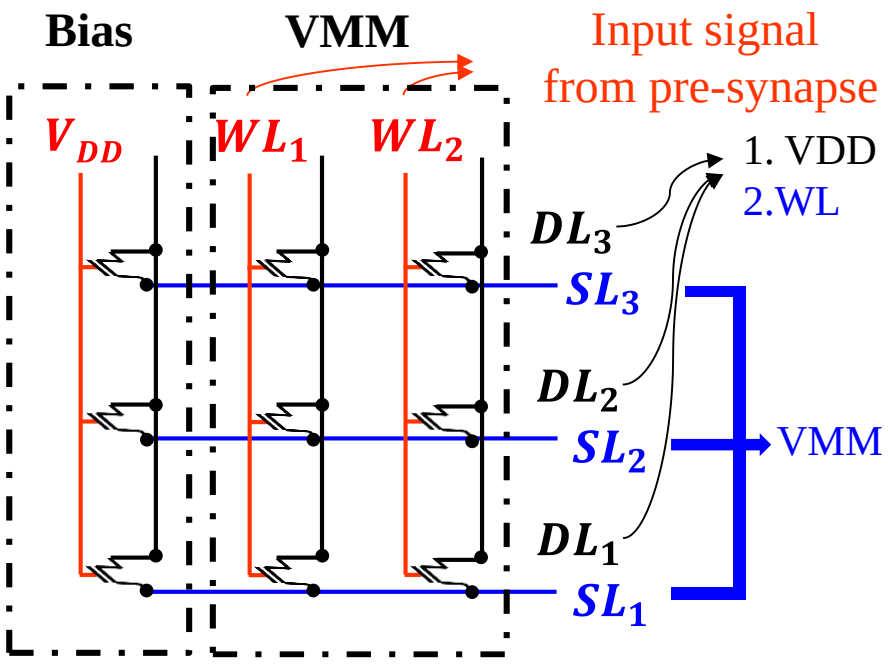


Figure 4.2. Schematic of VMM operation in NOR-type array.

4.2 High-Level Simulation with Retention Characteristics and Voltage Variations

Finely adjusting the weights of the synaptic device is demonstrated in **Figure 3. 24**, and these weights should be well preserved over time. Hot temperature retention characteristics are measured in OCS transistors to evaluate the synaptic weight performance. Trapped charges can leak over time, causing data loss and significantly degrading the performance of neuromorphic systems. By measuring the hot temperature retention characteristics, it is possible to extrapolate the retention behavior at average operating temperatures.

The Arrhenius equation determines the acceleration in charge de-trapping in the flash memory [56]. The acceleration factor (AF) is calculated below:

$$\text{Rate of reaction} = Ae^{-\frac{Ea}{RT}} \tag{4.9}$$

$$\text{AF=Rate of change at T2/ Rate of change at T1} = (Ae^{-\frac{Ea}{RT2}})/(Ae^{-\frac{Ea}{RT2}}) \tag{4.10}$$

$$= e^{-\frac{Ea}{R}(T2-T1)} \tag{4.11}$$

$$= e^{-K(1/T2-1/T1)} \tag{4.12}$$

According to the Arrhenius equation, the obtained AF for the OCS transistor is 647.5 from 30 ℃ to 85 ℃. **Figure 4.3** displays the measurement results of hot

temperature retention characteristics, and the weight values after several years are inferred by applying the calculated AF. Based on this data, the accuracy of convolutional neural networks (CNNs) is analyzed by high-level simulation.

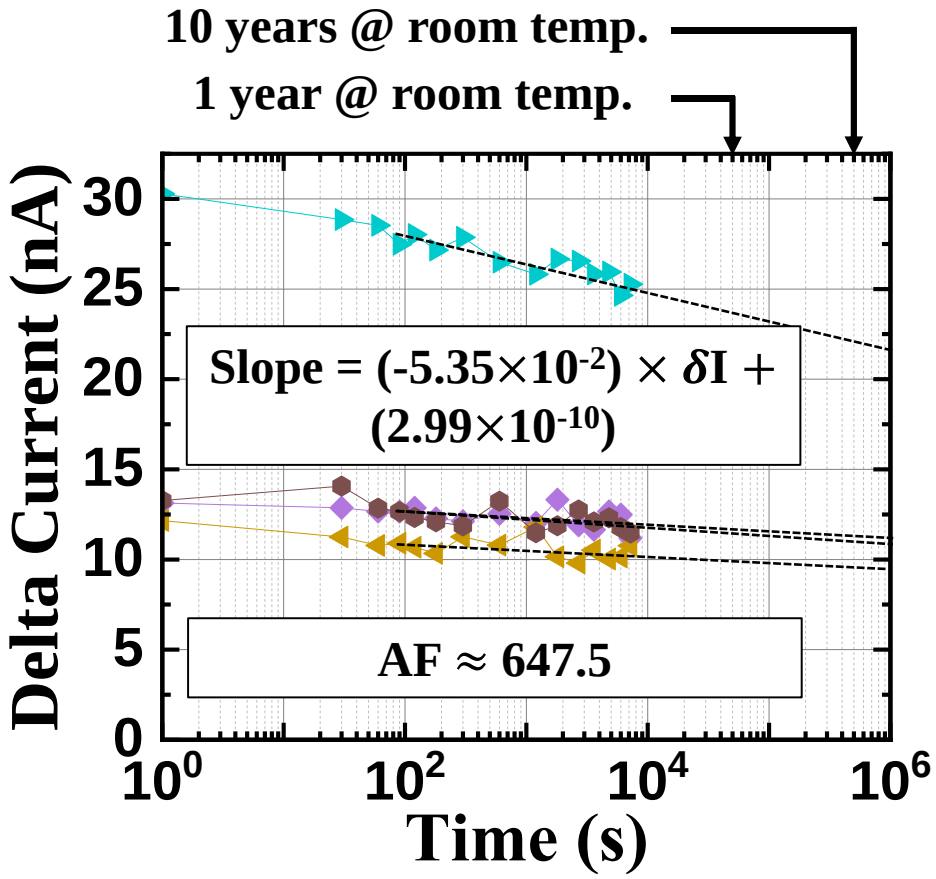


Figure 4.3. Retention characteristics of the OCS transistor in the 85 ℃. The acceleration factor is computed to estimate the current after several years.

Figure 4.4 shows the structure of CNNs with the fashion MNIST dataset. A stochastic gradient descent algorithm is used to train the networks, and hyperparameters are summarized in **Table 4.1** [57]. Since the spiking neuron and rectified linear unit (ReLU) are mathematically the same systems, the simulation is conducted by converting ANNs to spiking neural networks (SNNs). The SNNs with four-bit weights achieved 91.29% accuracy in a year, as shown in **Figure 4.5**.

In Chapter 3.4.4, the current variation according to the voltage variation is extracted and reflected in the same neural networks as depicted in **Figure 4.4**. The input voltage is set to follow the standard normal distribution for the two cases of $\mu=3$, $\sigma=0.065$, and 0.032, as shown in **Figure 4.6**. When σ is 0.065 and 0.032, $P(\mu-\sigma \leq X \leq \mu-\sigma)$ is 68.3% and 95.4%, reflecting the accuracy of HV-SPGU. The median classification accuracy is achieved greater than 92% with 50 samples in both cases, as shown in **Figure 4.7**.

In conclusion, the classification accuracy performance degradation according to the retention characteristics and input voltage variations of the OCS transistor is not significant, making it a suitable synaptic device for neuromorphic systems.

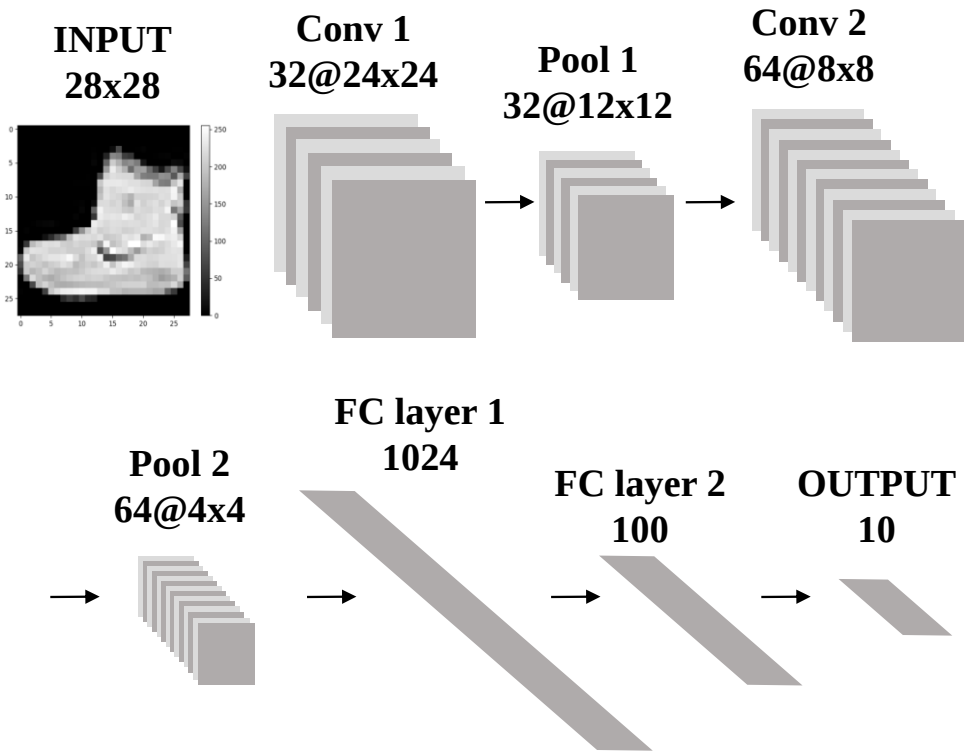


Figure 4.4. Diagrammatic representation of the simulated convolutional neural networks (CNNs) with fashion MNIST dataset.

Table 4.1. Hyperparameters of CNNs.

	Value
Batch size	1000
Learning rate	0.2 (StepLR)
Momentum	0.9
Dropout	0.3(C. layer) 0.5(F.C. layer)
ANN accuracy with test set	92.64%

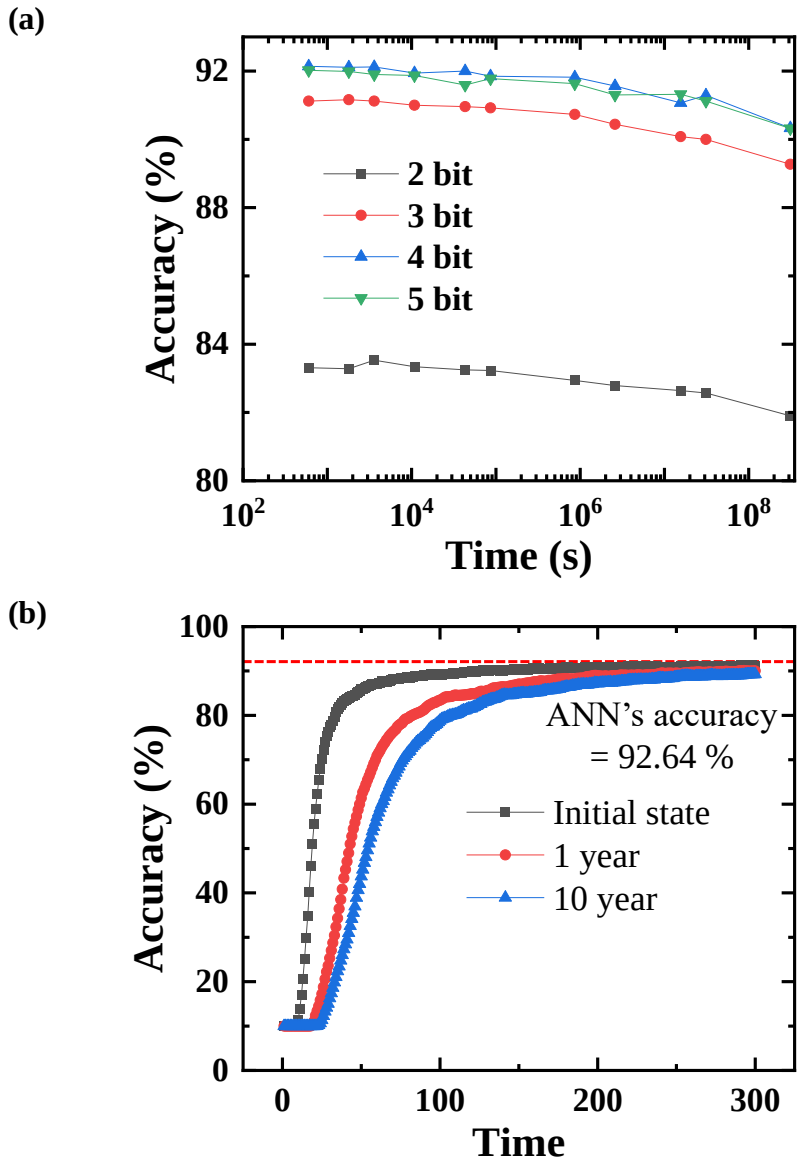


Figure 4.5. (a) Classification accuracy with quantization level as a function of time. (b) 4-bit quantization classification accuracy according to the simulation timestep.

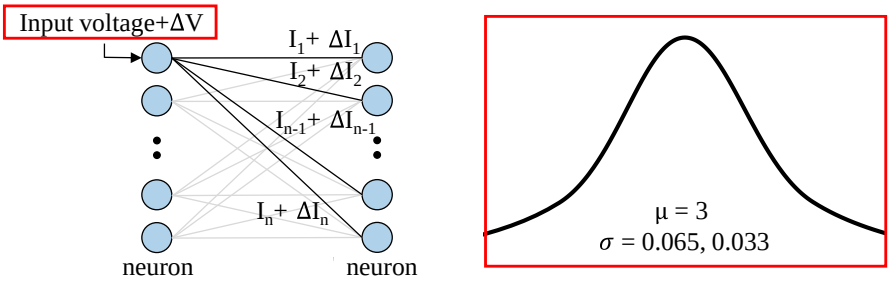


Figure 4.6. Input voltage variation with μ=3, and σ=0.065, 0.033.

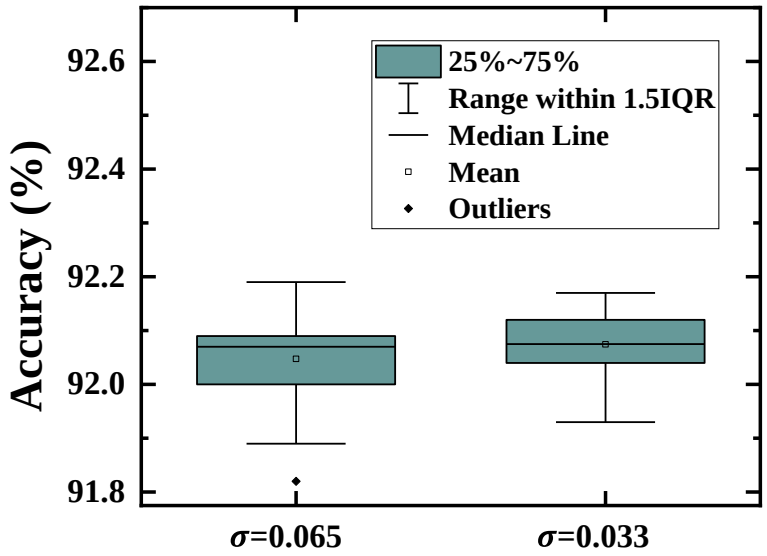


Figure 4.7. Classification accuracy with σ=0.065 and 0.032.

Chapter 5

Conclusion

In this dissertation, we propose and fabricate an overpass channel synaptic (OCS) transistor suitable for ultra-low power neuromorphic systems. The proposed OCS array is a flash-based NOR-type array structure, and we verify the key synaptic functions of vector matrix multiplication (VMM) and weight modulation. Moreover, by integrating the top gate line (TGL) and drain line (DL), we achieve cell size reduction and stable operation. The OCS array exhibits excellent performance in neuromorphic systems through high-level simulations that reflect its retention characteristics.

First, the short channel effect and decrease in the charge storage layer caused by scaling down the device hinder the stable operation. To fundamentally address this problem, the OCS transistor, in which the channel crosses the bottom gate fin, is

proposed to increase the effective gate length and charge storage layer as the fin height increases. Additionally, we optimize the on-current level for minimizing power consumption and fast operation speed through SPICE simulation to verify neuron compatibility.

Second, we use a mix-and-match process involving electron-beam lithography and photolithography for aggressively scaled devices. The OCS transistor is optimized for ultra-low power operation with a current level of several tens of nA, and we propose and verify the scheme for controlling conductance using two asymmetric gates. Power consumption is minimized by using Fowler-Nordheim tunneling. Moreover, the current sum error is 0.79% during the inference, enabling accurate VMM calculations. Furthermore, process-voltage-temperature variations are verified for the stable operation.

The diode-connected (D-C) synapse array, which integrates TGL and DL, reduces cell size and ensures stable current flow by continually operating in the saturation region. Consequently, the leakage current is minimized when the OCS transistor is off, and the program (PGM)/erase (ERS) scheme using the tail part of the top gate is proposed. The change in conductance of the target cell in the OCS array is measured with neighboring cells, and a cell inhibition value within 5% from the second PGM/ERS cycle is validated.

Finally, the retention characteristics are analyzed to verify the inference accuracy

of the fashion MNIST dataset in convolutional neural networks (CNNs). After a year, the spiking neural networks with four-bit weights exhibit a classification accuracy of 91.29%. It shows outstanding performance and significant advantages as an ultra-low power synaptic device.

Bibliography

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," (in eng), *Nature,* vol. 521, no. 7553, pp. 436-44, May 28 2015, doi: 10.1038/nature14539.

[2] R. A. Nawrocki, R. M. Voyles, and S. E. Shaheen, "A mini review of neuromorphic architectures and implementations," *IEEE Transactions on Electron Devices,* vol. 63, no. 10, pp. 3819-3829, 2016.

[3] A. R. Young, M. E. Dean, J. S. Plank, and G. S. Rose, "A review of spiking neuromorphic hardware communication systems," *IEEE Access,* vol. 7, pp. 135606-135620, 2019.

[4] C. Carvalho, "The Gap between Processor and Memory Speeds," 2002.

[5] A. Nowak, "Opportunities and choice in a new vector era," *Journal of Physics: Conference Series,* vol. 523, no. 1, p. 012002, 2014/06/06 2014, doi: 10.1088/1742-6596/523/1/012002.

[6] A. Jaiswal, I. Chakraborty, A. Agrawal, and K. Roy, "8T SRAM Cell as a Multibit Dot-Product Engine for Beyond Von Neumann Computing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 27, no. 11, pp. 2556-2567, 2019, doi: 10.1109/TVLSI.2019.2929245.

[7] Z. R. Wang *et al.*, "Efficient Implementation of Boolean and Full-Adder Functions With 1T1R RRAMs for Beyond Von Neumann In-Memory Computing," *IEEE Transactions on Electron Devices,* vol. 65, no. 10, pp. 4659-4666, 2018, doi: 10.1109/TED.2018.2866048.

[8] C. D. Wright, P. Hosseini, and J. A. V. Diosdado, "Beyond von-Neumann Computing with Nanoscale Phase-Change Memory Devices," *Advanced Functional Materials,* vol. 23, no. 18, pp. 2248-2254, 2013, doi: https://doi.org/10.1002/adfm.201202383.

[9] J. M. Cruz-Albrecht, M. W. Yung, and N. Srinivasa, "Energy-Efficient Neuron, Synapse and STDP Integrated Circuits," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 6, no. 3, pp. 246-256, 2012, doi:

10.1109/TBCAS.2011.2174152.

[10] G. Indiveri, F. Stefanini, and E. Chicca, "Spike-based learning with a generalized integrate and fire silicon neuron," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 30 May-2 June 2010 2010, pp. 1951-1954, doi: 10.1109/ISCAS.2010.5536980.

[11] M.-W. Kwon *et al.*, "Integrate-and-fire neuron circuit using positive feedback field effect transistor for low power operation," *Journal of Applied Physics*, vol. 124, no. 15, 2018, doi: 10.1063/1.5031929.

[12] C. Mayr *et al.*, "A Biological-Realtime Neuromorphic System in 28 nm CMOS Using Low-Leakage Switched Capacitor Circuits," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 1, pp. 243-254, 2016, doi: 10.1109/TBCAS.2014.2379294.

[13] J. K. Han *et al.*, "Mimicry of Excitatory and Inhibitory Artificial Neuron With Leaky Integrate-and-Fire Function by a Single MOSFET," *IEEE Electron Device Letters*, vol. 41, no. 2, pp. 208-211, 2020, doi: 10.1109/LED.2019.2958623.

[14] Y. Jiang *et al.*, "Design and Hardware Implementation of Neuromorphic Systems With RRAM Synapses and Threshold-Controlled Neurons for Pattern Recognition," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 9, pp. 2726-2738, 2018, doi: 10.1109/TCSI.2018.2812419.

[15] J. Lin and J. S. Yuan, "Analysis and Simulation of Capacitor-Less ReRAM-Based Stochastic Neurons for the in-Memory Spiking Neural Network," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 5, pp. 1004-1017, 2018, doi: 10.1109/TBCAS.2018.2843286.

[16] J. Luo *et al.*, "Capacitor-less Stochastic Leaky-FeFET Neuron of Both Excitatory and Inhibitory Connections for SNN with Reduced Hardware Cost," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 7-11 Dec. 2019 2019, pp. 6.4.1-6.4.4, doi: 10.1109/IEDM19573.2019.8993535.

[17] X. Zhang *et al.*, "An Artificial Neuron Based on a Threshold Switching Memristor," *IEEE Electron Device Letters*, vol. 39, no. 2, pp. 308-311,

2018, doi: 10.1109/LED.2017.2782752.

[18] F. Akopyan *et al.*, "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 34, no. 10, pp. 1537-1557, 2015, doi: 10.1109/TCAD.2015.2474396.

[19] M. Davies *et al.*, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro,* vol. 38, no. 1, pp. 82-99, 2018, doi: 10.1109/MM.2018.112130359.

[20] M. Davies *et al.*, "Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook," *Proceedings of the IEEE,* vol. 109, no. 5, pp. 911-934, 2021, doi: 10.1109/JPROC.2021.3067593.

[21] M. V. DeBole *et al.*, "TrueNorth: Accelerating From Zero to 64 Million Neurons in 10 Years," *Computer,* vol. 52, no. 5, pp. 20-29, 2019, doi: 10.1109/MC.2019.2903009.

[22] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker Project," *Proceedings of the IEEE,* vol. 102, no. 5, pp. 652-665, 2014, doi: 10.1109/JPROC.2014.2304638.

[23] E. Painkras *et al.*, "SpiNNaker: A 1-W 18-Core System-on-Chip for Massively-Parallel Neural Network Simulation," *IEEE Journal of Solid-State Circuits,* vol. 48, no. 8, pp. 1943-1953, 2013, doi: 10.1109/JSSC.2013.2259038.

[24] M. Ishii *et al.*, "On-chip trainable 1.4 M 6T2R PCM synaptic array with 1.6 K stochastic LIF neurons for spiking RBM," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019: IEEE, pp. 14.2. 1-14.2. 4.

[25] H. Jeong and L. Shi, "Memristor devices for neural networks," *Journal of Physics D: Applied Physics,* vol. 52, no. 2, p. 023003, 2018/10/30 2019, doi: 10.1088/1361-6463/aae223.

[26] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale Memristor Device as Synapse in Neuromorphic Systems," *Nano Letters,* vol. 10, no. 4, pp. 1297-1301, 2010/04/14 2010, doi: 10.1021/nl904092h.

[27] S. Kim, H. Kim, S. Hwang, M.-H. Kim, Y.-F. Chang, and B.-G. Park, "Analog Synaptic Behavior of a Silicon Nitride Memristor," *ACS Applied Materials & Interfaces*, vol. 9, no. 46, pp. 40420-40427, 2017/11/22 2017, doi: 10.1021/acsami.7b11191.

[28] T.-H. Kim, H. Nili, M.-H. Kim, K. K. Min, B.-G. Park, and H. Kim, "Reset-voltage-dependent precise tuning operation of TiOx/Al2O3 memristive crossbar array," *Applied Physics Letters*, vol. 117, no. 15, p. 152103, 2020.

[29] W. Wang *et al.*, "Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses," *Science Advances*, vol. 4, no. 9, p. eaat4752, doi: 10.1126/sciadv.aat4752.

[30] Q. Zhang *et al.*, "Sign backpropagation: an on-chip learning algorithm for analog RRAM neuromorphic computing systems," *Neural Networks*, vol. 108, pp. 217-223, 2018.

[31] Z. Zhou *et al.*, "A new hardware implementation approach of BNNs based on nonlinear 2T2R synaptic cell," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018: IEEE, pp. 20.7. 1-20.7. 4.

[32] R. Aluguri and T.-Y. Tseng, "Notice of violation of IEEE publication principles: overview of selector devices for 3-D stackable cross point RRAM arrays," *IEEE Journal of the Electron Devices Society*, vol. 4, no. 5, pp. 294-306, 2016.

[33] T.-H. Kim, B. Song, I.-J. Jung, and S.-O. Jung, "A Sneak Current Compensation Scheme With Offset Cancellation Sensing Circuit for ReRAM-Based Cross-Point Memory Array," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 4, pp. 1583-1594, 2021.

[34] W. Y. Park *et al.*, "A Pt/TiO2/Ti Schottky-type selection diode for alleviating the sneak current in resistance switching memory arrays," *Nanotechnology*, vol. 21, no. 19, p. 195201, 2010.

[35] D. Schor. "Intel Labs Builds A Neuromorphic System With 64 To 768 Loihi Chips: 8 Million To 100 Million Neurons." https://fuse.wikichip.org/news/2519/intel-labs-builds-a-neuromorphic-

system-with-64-to-768-loihi-chips-8-million-to-100-million-neurons/ (accessed April. 29, 2023).

[36] S.-T. Lee *et al.*, "Operation scheme of multi-layer neural networks using NAND flash memory as high-density synaptic devices," *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 1085-1093, 2019.

[37] W. Zhou *et al.*, "Unsupervised learning in winner-takes-all neural network based on 3D NAND flash," *IEEE Electron Device Letters*, vol. 43, no. 3, pp. 374-377, 2022.

[38] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017: IEEE, pp. 6.5. 1-6.5. 4.

[39] J. N. Kim, J. Lee, J. E. Kim, S. W. Hong, M. Koo, and Y. Kim, "NOR-Type 3-D Synapse Array Architecture Based on Charge-Trap Flash Memory," *IEEE Journal of the Electron Devices Society*, vol. 10, pp. 813-820, 2022.

[40] Y.-T. Seo *et al.*, "3-D AND-type flash memory architecture with high-κ gate dielectric for high-density synaptic devices," *IEEE Transactions on Electron Devices*, vol. 68, no. 8, pp. 3801-3806, 2021.

[41] M.-H. Baek, T. Jang, M.-W. Kwon, S. Hwang, S. Kim, and B.-G. Park, "Polysilicon-based synaptic transistor and array structure for short/long-term memory," *Journal of Nanoscience and Nanotechnology*, vol. 19, no. 10, pp. 6066-6069, 2019.

[42] H. Kim, S. Hwang, J. Park, S. Yun, J.-H. Lee, and B.-G. Park, "Spiking neural network using synaptic transistors and neuron circuits for pattern recognition with noisy images," *IEEE Electron Device Letters*, vol. 39, no. 4, pp. 630-633, 2018.

[43] B.-g. Park, M.-H. Baek, and T. Jang, "Semi-conductor device having double-gate and method for setting synapse weight of target semi-conductor device within neural network," ed: Google Patents, 2022.

[44] Y. Yang and M. H. White, "Charge retention of scaled SONOS nonvolatile memory devices at elevated temperatures," *Solid-State Electronics*, vol. 44,

no. 6, pp. 949-958, 2000/06/01/ 2000, doi: https://doi.org/10.1016/S0038-1101(00)00012-5.

[45] S. Ambrogio *et al.*, "Novel RRAM-enabled 1T1R synapse capable of low-power STDP via burst-mode communication and real-time unsupervised machine learning," in *2016 IEEE Symposium on VLSI Technology*, 2016: IEEE, pp. 1-2.

[46] S. Ambrogio *et al.*, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM," *IEEE Transactions on Electron Devices*, vol. 63, no. 4, pp. 1508-1515, 2016.

[47] J. Hur *et al.*, "A Recoverable Synapse Device Using a Three-Dimensional Silicon Transistor," *Advanced Functional Materials*, vol. 28, no. 47, p. 1804844, 2018.

[48] C.-H. Kim *et al.*, "Demonstration of unsupervised learning with spike-timing-dependent plasticity using a TFT-type NOR flash memory array," *IEEE Transactions on Electron Devices*, vol. 65, no. 5, pp. 1774-1780, 2018.

[49] H. Kim, J. Park, M.-W. Kwon, J.-H. Lee, and B.-G. Park, "Silicon-based floating-body synaptic transistor with frequency-dependent short-and long-term memories," *IEEE Electron Device Letters*, vol. 37, no. 3, pp. 249-252, 2016.

[50] S.-T. Lee *et al.*, "High-density and highly-reliable binary neural networks using NAND flash memory cells as synaptic devices," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019: IEEE, pp. 38.4. 1-38.4. 4.

[51] V. Milo *et al.*, "Multilevel HfO2-based RRAM devices for low-power neuromorphic networks," *APL materials*, vol. 7, no. 8, p. 081120, 2019.

[52] H. Zhao *et al.*, "Implementation of discrete Fourier transform using RRAM arrays with quasi-analog mapping for high-fidelity medical image reconstruction," in *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021: IEEE, pp. 12.4. 1-12.4. 4.

[53] A. W. De Groot, G. C. McGonigal, D. J. Thomson, and H. C. Card,

"Thermionic-field emission from interface states at grain boundaries in silicon," *Journal of Applied Physics*, vol. 55, no. 2, pp. 312-317, 1984, doi: 10.1063/1.333099.

[54] K. C. P. Byung-Gook Park, "NEURON CIRCUIT AND NEUROMORPHIC DEVICE FOR COMPENSATING FOR CHAGE IN SYNAPTIC PROPERTIES," 2023.

[55] S. Dai *et al.*, "Recent advances in transistor-based artificial synapses," *Advanced Functional Materials,* vol. 29, no. 42, p. 1903700, 2019.

[56] E. Tiomkin. "Industrial Temperature and NAND Flash in SSD Products." https://www.eeweb.com/industrial-temperature-and-nand-flash-in-ssd-products/ (accessed April.26, 2023).

[57] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

초 록

Von Nueumann 기반의 심층 신경망은 컴퓨팅 파워의 급속한 발전으로 놀라운 성능을 달성해왔다. 그러나 데이터 이동은 중앙처리장치(CPU)와 메모리의 직렬 연결로 인해 상당한 시간과 에너지를 소모한다. 그 결과, 뉴로모픽 시스템은 매우 낮은 전력 소비를 유지하면서 인공 신경망의 계산 문제를 해결하는 강력한 후보로 주목받았다. 특히, 뉴로모픽 시스템의 중요한 구성 요소인 시냅스는 뉴런 사이의 가중치를 저장하고 신호를 전달한다. 일반적으로 인공 시냅스는 벡터곱 연산이 Kirchhoff의 전류 법칙에 따라 전류 합으로 표현되기 때문에 빠르고 낮은 전력 소모로 작동한다.

따라서 멤리스터 기반의 2단자 소자, 플래시 메모리 기반의 시냅스 소자 등 다양한 시냅스 소자가 제안 및 연구되었다. 멤리스터 시냅스 소자는 구조가 단순하고 고집적 · 대용량 집적에 유리하지만 소자 편차, 신뢰성, 누설 전류 등에 한계가 있다. 반면 플래시 메모리 기반 시냅스 소자는 오랜 연구 역사를 가진 성숙한 분야로 안정적인 멀티비트 동작을 제공한다. 일반적으로 NAND, NOR 및 AND 타입의 어레이가 시냅스 어레이로 구현되며, NAND형 어레이는 가중치 값을 순차적으로 읽어야 하므로 추가 회로가 필요하다. 반면 NOR 및

AND 타입 어레이는 소스에서 전류를 감지하여 벡터곱 연산을 수행할 수 있으나 고집적에 한계가 있다.

본 논문에서는 초저전력 동작 및 RC 지연 감소를 위한 poly-Si 기반의 OCS(overpass channel synaptic) 트랜지스터를 제안하였다. OCS 트랜지스터는 두 가지 주요 구조적 이점을 보여준다. 첫째, 온 전류가 100nA 미만으로 핀 모양의 바닥 게이트를 둘러싼 채널로 인해 높은 온/오프 전류 비율을 가지고 있다. 둘째, 전하 저장층의 부피를 늘려 OCS 트랜지스터의 가중치를 미세하게 나눌 수 있다. NOR 타입의 OCS 어레이의 추론 및 가중치 업데이트를 실험적으로 증명하였다.

제작된 다이오드 연결형(D-C) OCS 어레이는 추론 시 1% 미만의 벡터곱 연산 오차를 나타냄을 검증하였다. 또한, D-C OCS 어레이의 시냅스 가중치는 비대칭 게이트가 있는 FN(Fowler-Nordheim) 터널링을 사용하여 nA 미만의 간격으로 조정된다. 마지막으로 Fashion MNIST 데이터 세트의 분류 정확도는 스파이킹 신경망(SNN)의 시냅스 가중치를 4비트 양자화 한 상태에서 1년 후에도 91.29%를 달성하였다.

주요어 : 폴리 실리콘 시냅스 소자, NOR 타입 어레이, 플래시 메모리, 저전력 동작, 스파이킹 뉴럴 네트워크(SNN), 뉴로모픽 시스템

학번 : 2016-20970