Ph.D. THESIS

Silicon-Based Synaptic Transistor for Neuromorphic Computing Systems

신경계 모방 시스템을 위한 실리콘 기반 시냅스 모방 트랜지스터

BY

HYUNGJIN KIM

February 2017

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

Silicon-Based Synaptic Transistor for Neuromorphic Computing Systems

신경계 모방 시스템을 위한 실리콘 기반 시냅스 모방 트랜지스터

지도교수 박 병 국

이 논문을 공학박사 학위논문으로 제출함

2017 년 2 월

서울대학교 대학원

전기컴퓨터공학부

김 형 진

김형진의 공학박사 학위논문을 인준함

2017 년 2 월

위 원 장 : 박 영 준 (인)

부위원장 : 박 병 국 (인)

위 원 : 이 종 호 (인)

위 원 : 송 윤 흡 (인)

위 원 : 박 완 준 (인)

Abstract

Current computing systems based on the von Neumann architecture suffer from the fact that serious leakage current issues have risen up in nanoscale devices. Neuromorphic computing system has been believed to solve fundamental challenges in current computing system by mimicking a biological nervous system, especially in terms of parallel signal processing. Synaptic devices are considered as the one of the most important parts of neuromorphic systems because a biological synapse is thought to control signal transmissions and memory effects in our nervous system. However, the memristor, one of the strongest candidates for an artificial synapse, requires additional switching parts in order to transfer and receive signals using the same electrode, leading to extra overheads that may compromise the advantages of massive parallelism inherent in neuromorphic systems.

In this dissertation, a silicon-based synaptic transistor with asymmetric dual-gate structure is investigated. The structural feature enables the synaptic transistors to interact with both pre- and post-synaptic neuron circuits directly. A TCAD device simulator and a circuit simulator are used to verify its synaptic learning properties and study its mechanism fundamentally. After verifying all the fabrication flow using a process simulator, the synaptic transistors are

fabricated through process techniques including two-step CMP processes. The electrical and synaptic characteristics of the fabricated devices are measured with a semiconductor parameter analyzer, and a device model is created based on the measured data. Furthermore, spiking neural network composed of them is verified systematically using the device model.

From the simulation study and electrical measurement, synaptic learning rules are observed in the synaptic transistors including the transition from short-term to long-term memory and spike-timing dependent plasticity. In addition, the spiking neural network composed of the synaptic transistors boasted its ability of pattern recognition using MNIST data set. The total recognition accuracy of the hardware-based neural network system having 784 input nodes and 10 output nodes is improved to nearly 70% by adding inhibitory synapses.

These results indicate that the synaptic transistor studied in this dissertation can be used as a synaptic device in neuromorphic systems thanks to its direct connectability with neuron circuits and synaptic learning properties.

Keywords : asymmetric dual-gate structure, neuromorphic system, pattern recognition, silicon-based synaptic transistor, spiking neural network, synaptic learning.

Student number : 2012-30934

Table of Contents

List of Figures

Chaper 1. Introduction

Chaper 2. Silicon-based Synaptic Transistor

Chaper 3. Device Fabrication

Chaper 4. Device Characteristics

Chaper 5. System Level Simulation

Appendix A. Multi-Threshold Voltages in Ultra Thin Body Devices by Asymmetric Dual-Gate Structure

Appendix B. Asymmetric Dual-Gate-Structured 1-T DRAM Cell for Retention Characteristics Improvement

Appendix C. A Single Memory Cell with Voltaile and Non-Volatile Memory Functions

Chapter 1. Introduction

1.1. Fundamental Challenges in Current Computing System

Since nearly the beginning of semiconductor history, semiconductor electronics have been enormously developed by reducing device size based on Moore's law, the observation that the number of transistors on a microprocessor chip doubles every two years, in order to obtain a high integration density, a high operation speed and even lowered cost per transistor [1]−[3]. Along with Moore's law, electronic systems have been allowed to expand their functionalities widely in various kinds of computing systems, such as desktop computers and portable electronics, to smart sensor systems. This trend was roughly kept for nearly past three decades, but it has become extremely difficult to maintain Moore's law because of severe leakage current issues especially in nanoscale devices [4]−[9], which make power dissipation problem as one of the most serious issues in current computing systems. According to Gordon Moore, the author of Moore's law, the leakage power consumption generated when transistors are off has been quickly approaching to the active power consumption generated when transistors are on as shown in Fig. 1.1 [9].

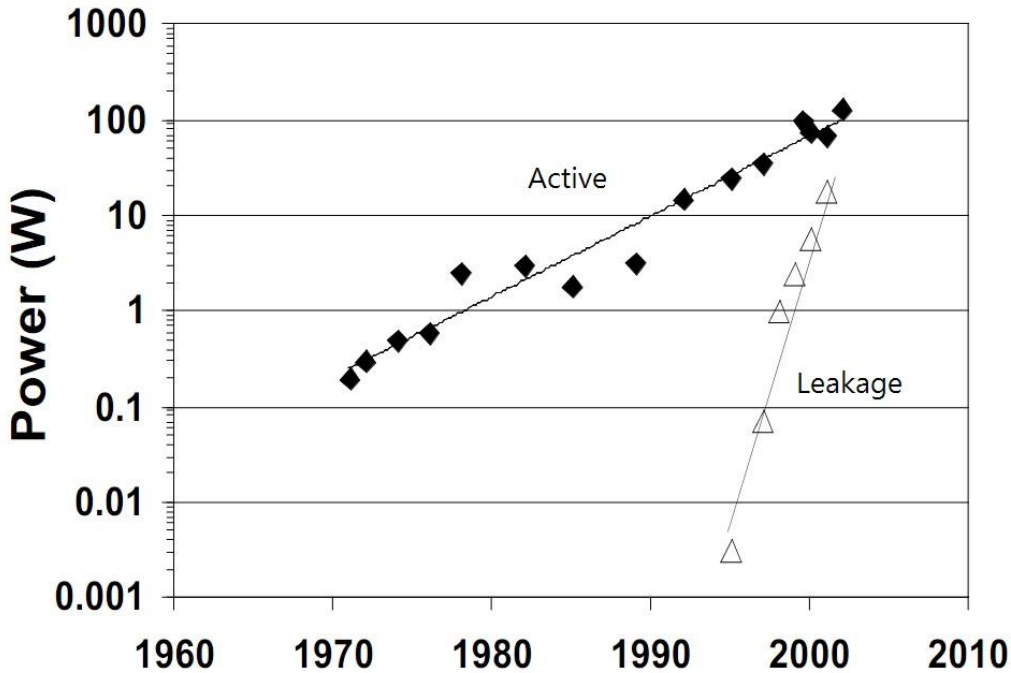


Fig. 1.1. Active and leakage power consumption with years [9].

Another problem originates from the structure of the von Neumann architecture itself. The von Neumann architecture is a computer architecture, devised by John von Neumann, which consists of a central processing unit (CPU), memory and input/output (I/O) devices and has been adopted in almost all computing systems in use to date [10]. However, the problem is that the throughput rate is lower than that at which the CPU can work because computer instructions should be executed sequentially owing to a shared common bus between an instruction fetch and a data operation. This is referred to as the von Neumann bottleneck, a term coined back in

1977 by John Backus, and often limits the performance of computing systems based on the von Neumann architecture [11], [12]. The von Neumann bottleneck is well illustrated in Fig. 1.2. All components of a digital computer such as CPU, memory and I/O devices are connected to a shared common bus, which seriously limits the effective processing speed when processing is required to be performed with large amounts of data.

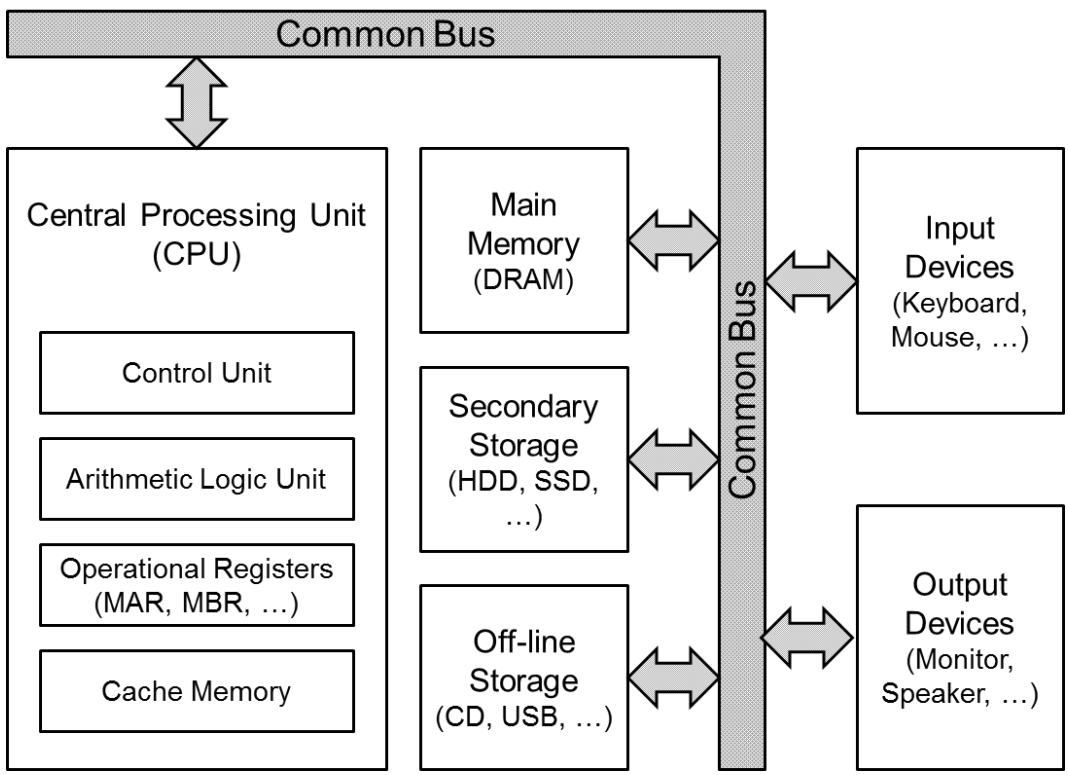


Fig. 1.2. Von Neumann bottleneck due to the shared common bus between program memory and data memory.

1.2. Neurobiological Background

In order to solve the problems of current computing systems mentioned above, there have been a number of attempts to realize neuromorphic systems, which have totally different structure and operation method compared to the computing systems based on the von Neumann architecture, by imitating a biological nervous system. Before discussing specific details about these kind of systems, the fundamental characteristics of a biological nervous system, especially biological synapses, are introduced first in this section because it is necessary to understand the basics of the way it operates.

1.2.1. Synaptic Transmission

Synaptic transmission, also called neurotransmission, is the information transfer through a biological synapse [13]−[17]. Synaptic transmission is essential for the communication between two neurons because the axon of a pre-synaptic neuron and the dendrite of a post-synaptic neuron are connected to a synapse. Action potentials (spikes) are generated when the membrane potential of a neuron exceeds a threshold, and the membrane potential is determined by the spatiotemporal integration of the neuron's input signals. Thus, the frequency and patterns of action potentials are determined by synaptic transmissions. This is the reason why it is believed that memory and synaptic weights are closely related.

The transmission of information is accomplished in two ways: electrical synapses and chemical synapses as shown in Fig. 1.3 [13]. Electrical synapses allow the electrical current of ions and small molecules between two neurons through gap junctions where very fast synaptic transmission can occur in a bidirectional manner. They are typically found where fast responses are required such as reflex reactions. On the other hand, the synaptic transmission of chemical synapses arises with neurotransmitters and receptors. Neurotransmitters stored in synaptic vesicles are released from pre-synaptic terminal and then receptors respond to them at post-synaptic terminal. Most synaptic transmission in a biological nervous system is in

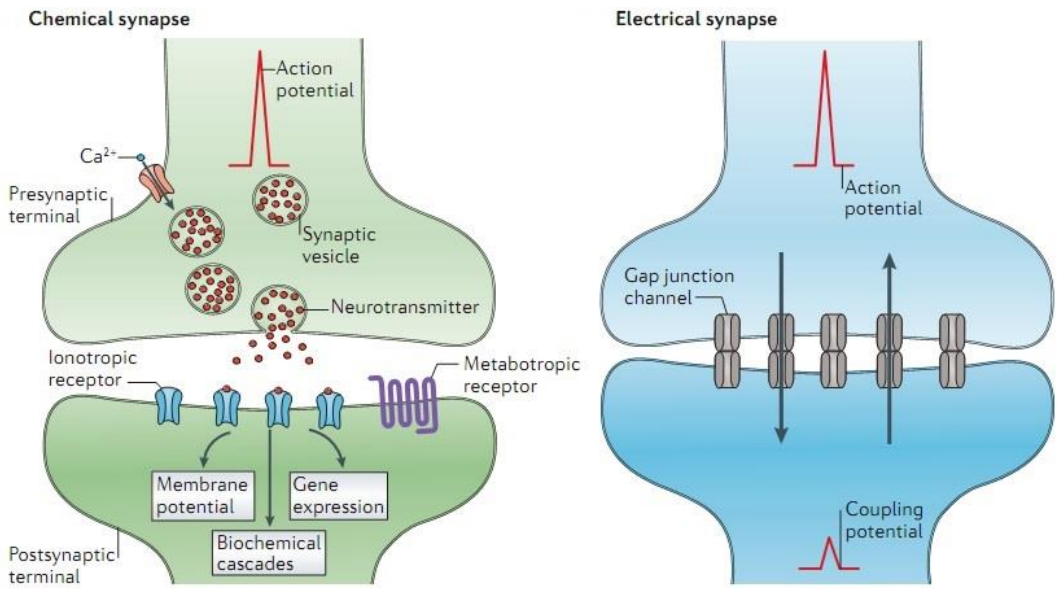


Fig. 1.3. Two kinds of biological synapses: chemical synapse and electrical synapse [13].

this way so drug therapies are widely used to deal with mental disorders because this synaptic transmission can be easily affected by specific chemicals.

The transferred signals through synaptic transmission are all integrated to membrane potential at the axon hillock of the post-synaptic neuron. and this process is known as synaptic integration or summation [18]–[20]; here, excitatory signals increase membrane potential but inhibitory signals take membrane potential away from threshold, and an action potential is triggered at the post-synaptic neuron when that membrane potential reaches its threshold. Figure 1.4 shows two kinds of synaptic integration: temporal summation and spatial summation. Temporal summation occurs when action potentials from the same pre-synaptic neuron are transferred in a high

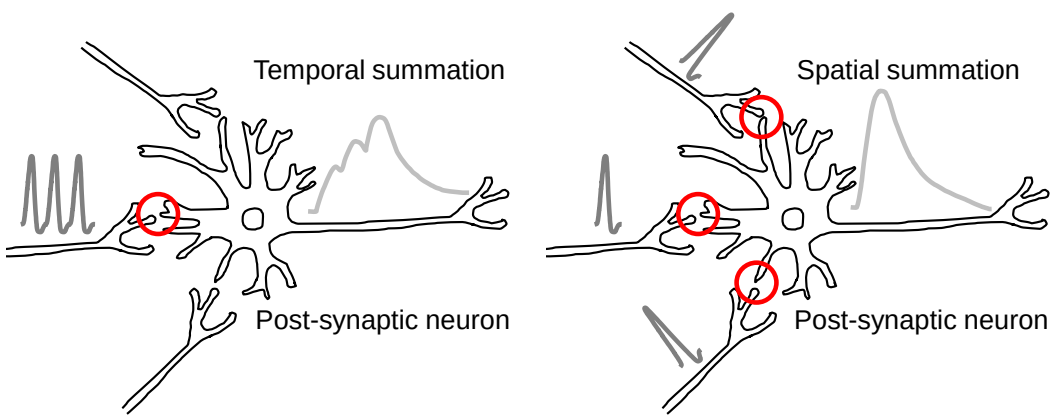


Fig. 1.4. Synaptic integration: temporal summation and spatial summation.

frequency repeatedly and those consecutive signals are summated together. On the other hand, spatial summation occurs when multiple pre-synaptic cells are stimulated simultaneously and those parallel signals from different pre-synaptic elements are incorporated all together into membrane potential.

1.2.2. Short-Term and Long-Term Memory

The creation of a memory is believed to occur in following three ways: firstly in the sensory stage; then in short-term memory; and finally in long-term memory through repeated and persistent stimulations. Kandel et al. revealed that those short-term memory and long-term memory in a sea slug, Aplysia, are located at the synapse [21], [22]. The schematic view of how short- and long-term memory form in a synapse is shown in Fig. 1.5 [21]. Weaker stimuli cause short-term memory. A stimulus on its tail releases serotonin (5HT) by the presynaptic axon which activates receptors releasing cyclic adenosine monophosphate (cAMP). The temporally increased cAMP activates protein kinases (PKA) which boost releasing neurotransmitters, leading to an amplification of the reflex from minutes to hours. This is the process of short-term memory formation. When repeatedly stimulated several times, however, long-term memory is formed. Stronger stimuli to its tail lead to not temporal but prolonged high level of cAMP, which results in the production of cAMP response element-binding protein 1 (CREB-1). CREB-1 activates nuclear

transcription factors CCAATT/enhancer binding proteins (C/EBP) that are thought to lead to the growth of new synaptic connections. Neurons make new physical connections and synapses in this way when long-term memory is formed by repeated stimuli, and these newly grown connections endure regardless of whether they are being used or not, once they are made.

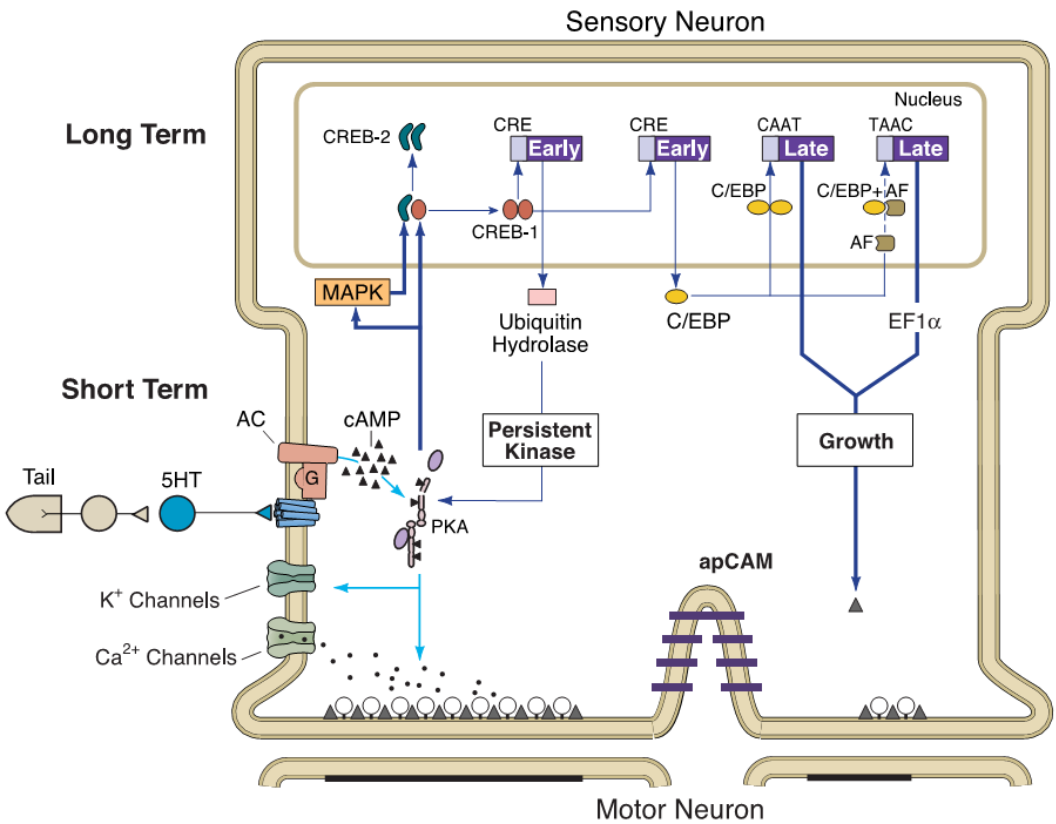


Fig. 1.5. Schematic view of mechanisms of short- and long-term memory formation in Aplysia [21].

1.2.3. Spike-Timing Dependent Plasticity

Spike-timing-dependent plasticity (STDP) is a process by which synaptic transmission is enhanced or depressed. In STDP, one of the fundamental mechanisms of learning in biological nerve systems, the precise relative timing of presynaptic and postsynaptic action potentials determines the strength of synaptic potentiation or depression [23]–[28]. It is widely believed as one of the basic rules of learning and storing information in the brain. Spike timing difference $\Delta t = t_{post}-t_{pre}$ is very important in STDP learning rule because it may refer to the causality of neural connections. If the presynaptic neuron becomes active slightly before the postsynaptic neuron on average ($\Delta t > 0$), which means the firing of the former may have a causal link to the firing of the latter, the input spike makes the connection between them stronger, resulting in long-term potentiation (LTP). On the other hand if the presynaptic neuron becomes active slightly after the postsynaptic neuron on average ($\Delta t < 0$), which means the firing of the former is not the cause of the firing of the latter, the input spike makes the connection weaker, leading to long-term depression (LTD). In this sense, STDP can be understood as a form of a Hebbian learning rule which says that a synapse is strengthened if a presynaptic neuron repeatedly or persistently contributes to the firing of the postsynaptic neuron [29].

The change of synaptic weight, excitatory postsynaptic potential (EPSP), plotted as a function of Δt is usually called the STDP function or learning window.

Bi and Poo reported STDP function observed in hippocampal neurons as shown in Fig. 1.6 [23]. In their experiments, repetitive postsynaptic spiking within 50 ms after presynaptic spiking resulted in LTP, while repetitive postsynaptic spiking within 50

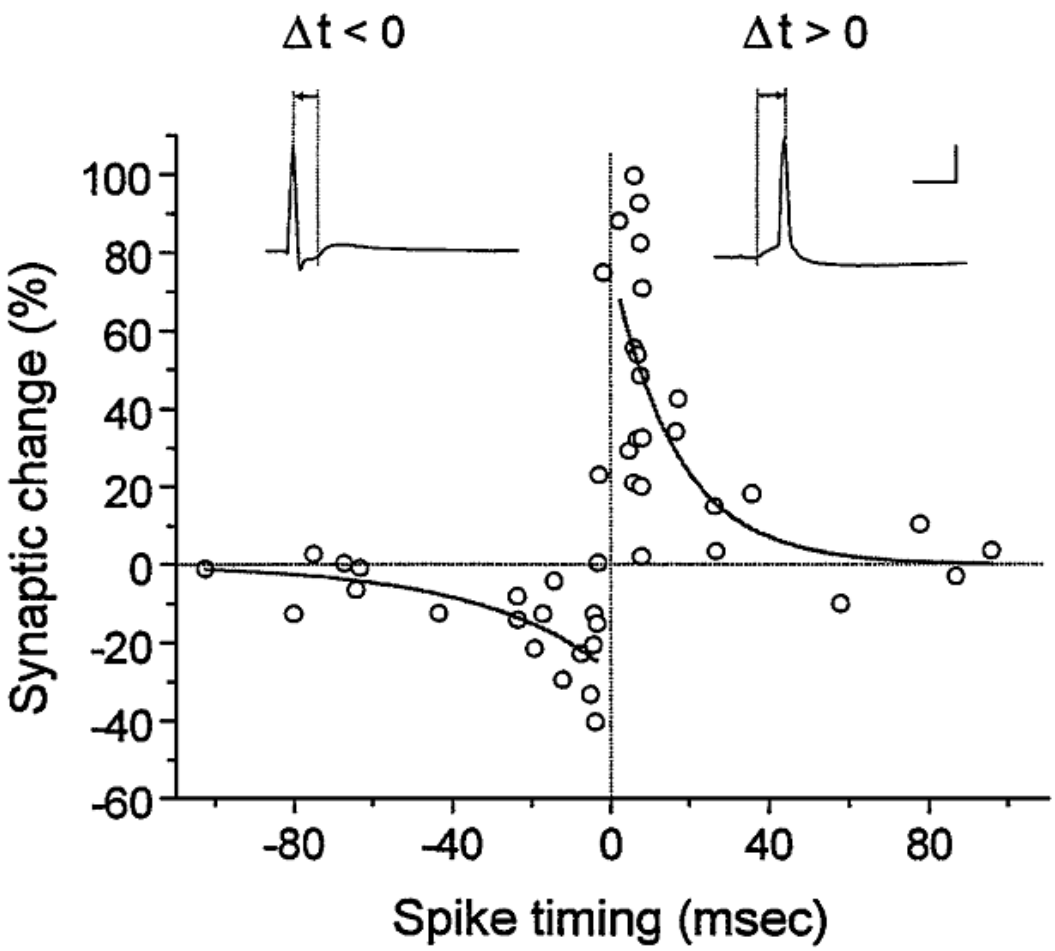


Fig. 1.6. The change of the excitatory postsynaptic potential (EPSP) amplitudes as a function of the time difference Δt [23].

ms before the presynaptic spiking resulted in LTD. Additionally, more significant synaptic modifications were obtained with Δt close to 0 s, which might indicate a high level causality between the two neurons.

Furthermore, STDP has a dependence on a synaptic weight [28]. Weaker synapses are more likely strengthened than stronger ones as shown in Fig. 1.7, which is called soft-bounds. In contrast, depression dominates over potentiation for stronger synapses, which is called hard-bounds. This weight dependent plasticity limits the synaptic weights in a finite range in both potentiation and depression directions [30]−[32].

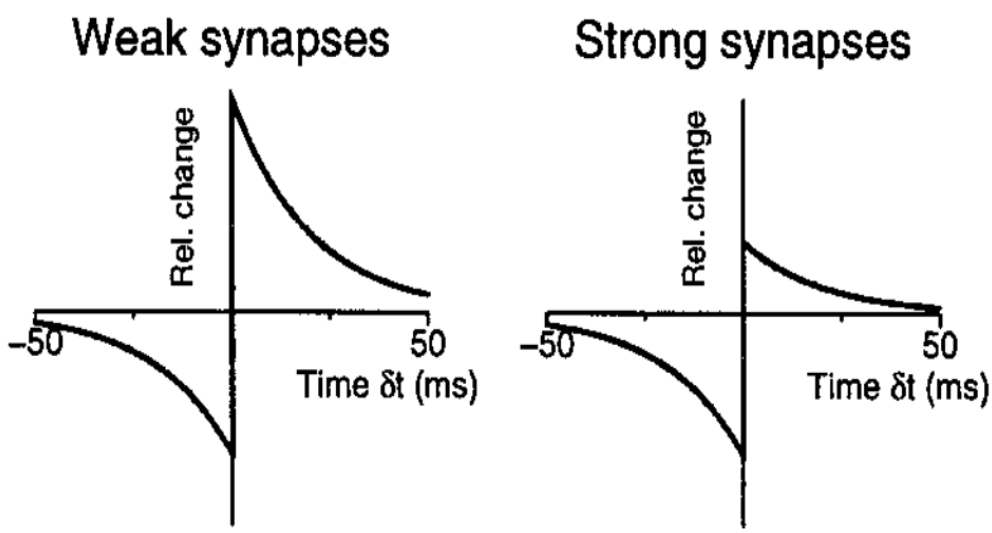


Fig. 1.7. Weight dependent STDP characteristics. The weaker synapses are more easily strengthened [28].

1.3. Neuromorphic Computing

Current computing systems based on von Neumann architecture have been suffering from the way they operate through serial information processing, called von Neumann bottleneck as discussed in Section 1.1. To solve this fundamental problem, there have been a lot of interesting researches about neuromorphic systems by imitating a biological nervous system. The concept of neuromorphic computing was firstly developed by Mead in the late 1980s using the very-large-scale integration (VLSI) implementation of analog circuits that mimic biological functions in the nervous system [33]. Since then, brain-inspired neuromorphic systems have been considered as beyond-von-Neumann-architecture computing systems for their energy-efficiency, parallel signal processing and fault tolerance [33]−[48]. A significant reduction in power consumption comes from the absence of CPU clock. Continuous power consumption is unavoidable for current CPUs controlled by a clock. Also, large fault tolerance is obtained from their massive parallel architecture. Even if one unit such as neurons or their connections is in false operation, it can be compensated by other units connected in parallel.

Synaptic devices are considered as one of the most important elements of neuromorphic systems because it is believed that biological synapses are in charge of signal transmissions and memory effects in our nervous systems as discussed in Section 1.2. Recently, various synaptic devices have been demonstrated to realize

hardware-based neural network systems [49]−[65]. Memristor is one of the strongest candidates for an artificial synapse because of their gradual switching characteristics and simple structure with two electrodes such as phase change materials and resistive switching materials, leading to high density nanocrossbar arrays [53]−[65]. Figure 1.8 shows a cation-based memristor as the application of a synaptic device [53]. This Ag-based memristor represents gradual switching properties using potentiating and depressing pulses, which have different pulse amplitudes, as shown in Fig. 1.8(c). In addition, STDP characteristics in this memristor are obtained by reducing the amplitude of each pulse in half as shown in Fig. 1.8(d). The reduced each single spike is unable to cause a conductance change, but the overlap between the presynaptic and postsynaptic spikes allows the memristor to change its conductance as a result of exceeding threshold voltage for resistance change.

However, these 2-terminal memristors requires additional switching components because artificial synapses have to transfer signals to the post-synaptic parts and receive back-propagation signals from the post-synaptic parts for online learning under STDP rules through the same electrodes [43]−[48]. Figure 1.9 shows one example of a neuromorphic system composed of memristors [43]. It requires a lot of switches and control logic to operate in two modes: training and testing. In the training mode, the connection (SW_4) between the memristors and the postsynaptic neuron circuit should be cut in order to apply the training voltage to their electrodes

for conductance change. Here, the other switches (SW_{1-3}) are also controlled depending on the direction of conductance changes: potentiation or depression. However, the memristors should be connected to the postsynaptic neuron circuit for the integration of the transferred signals from the postsynaptic ones and the firing of the postsynaptic one in the testing mode. In this sense, even though 2-terminal memristor has a simple structure itself, neuromorphic computing systems composed of them become quite complicated because of a lot of switching parts to operate in the two modes. Such a scheme inevitably incurs extra overheads that may compromise the advantage of massive parallelism inherent in neuromorphic systems.

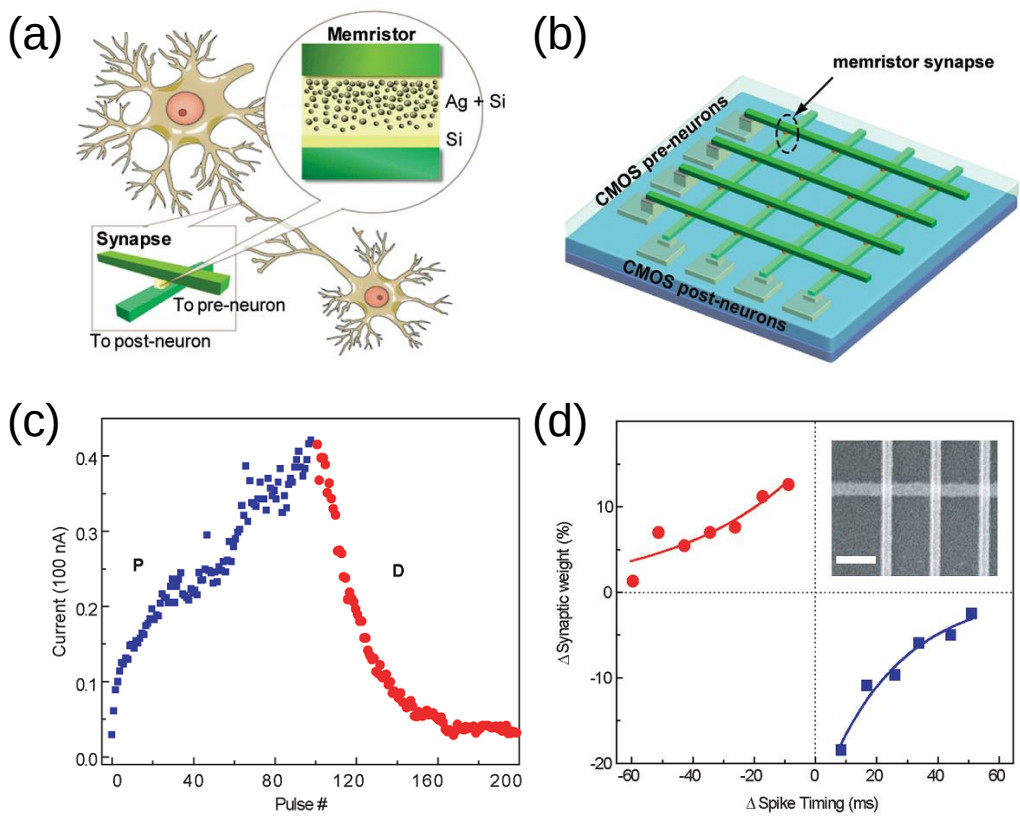


Fig. 1.8. Nanoscale memristor characteristics as a synaptic device [53]. (a) Schematic view of the concept of using memristor as a synaptic device. (b) Schematic view of connections between CMOS neurons and memristors. (c) Gradual switching charateristics by consecutive potentiating or depressing pulses. (d) Demonstration of STDP in the memristor synapse.

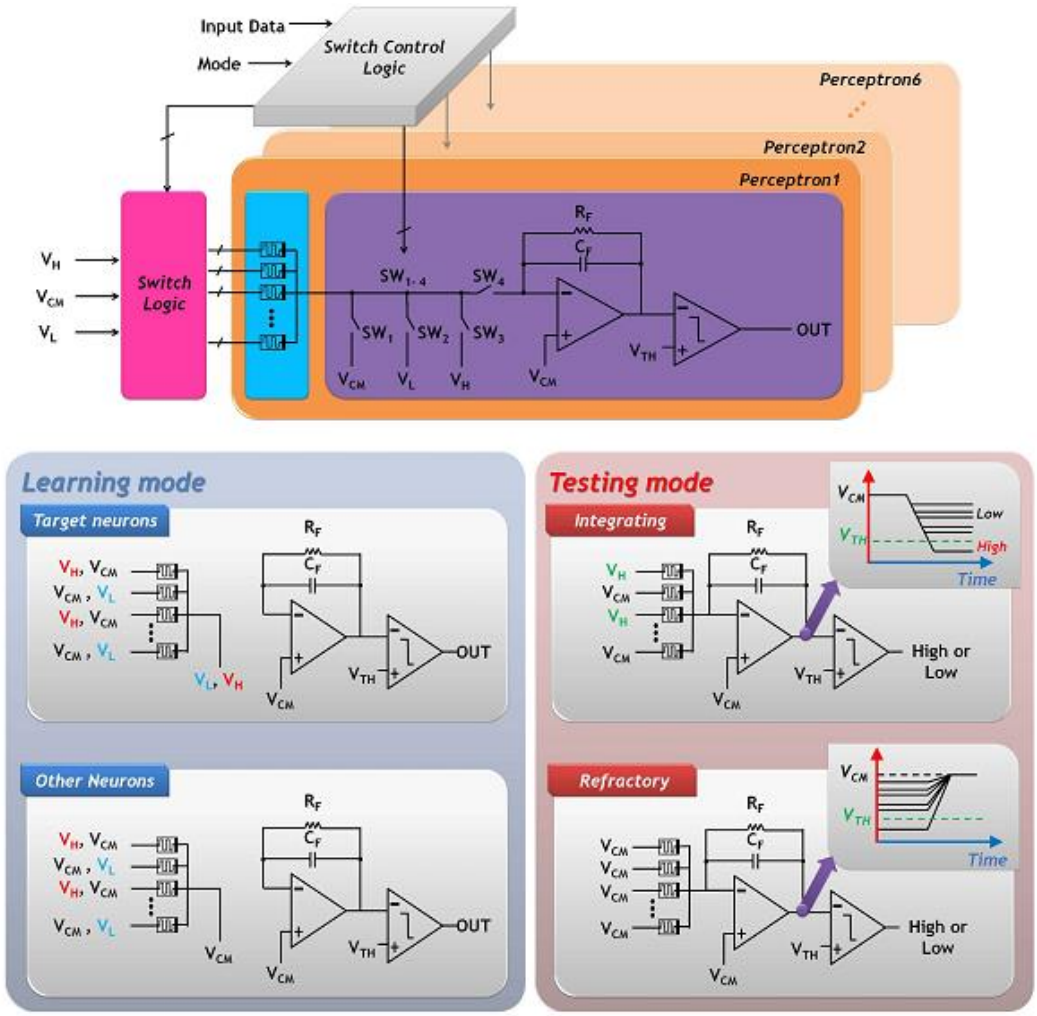


Fig. 1.9. One example of a neuromorphic system composed of 2-terminal memristors with switches and control logic circuit [43].

1.4. Outline of the Dissertation

Based on the discussions above, this work mainly focuses on a silicon-based synaptic transistor and its applications to neuromorphic computing. The remainder of this dissertation is organized as follows. Chapter 2 introduces the basic idea of the device and covers simulation studies verifying its functions both in terms of the device itself and systems composed of them. In Chapter 3, the process flow for the fabrication is briefly described. The key experimental conditions for the fabrication are explained in detail. Chapter 4 covers the measured data of the fabricated device in two respects: field effect transistor (FET) characteristics and synaptic properties. The measured FET characteristics confirms its structural feature, electrically independent two gates, and the measured synaptic properties including the transition from short-term to long-term memory and STDP characteristics show its potential as a synaptic component in neuromorphic systems. In Chapter 5, spiking neural networks based on a device model developed by the current equation of hot carrier injections are discussed. The pattern classification function of 28×28 hand-written digit images is studied in three ways how to construct spiking neural networks: with only excitatory synapse part, using transferred weights from artificial neural network, and the addition of inhibitory synapse part. Chapter 6 concludes this dissertation with a summary and recommendations for future work. Other applications including logic and memory devices are discussed in Appendices.

Chapter 2. Silicon-based Synaptic Transistor

2.1. Device Configuration

The schematic block diagram of the connection between the synaptic transistors and neuron circuit is shown in Fig. 2.1. The main feature of the fabricated synaptic transistor is electrically independent two gates with different gate stacks; the detailed process flow appears in Chapter 3. The first gate (G1) with a silicon dioxide (SiO_2) layer as a gate dielectric is used to receive signals from pre-synaptic neuron circuits and the second gate (G2) with ONO stacks is used to receive signals from post-synaptic ones and store charges in its nitride layer. In other words, G1 is used as a switching node whereas G2 is used as a memory node. This structural feature makes it possible that multiple pre-synaptic neuron inputs are transferred to each synaptic transistor and currents flowing through them are collected at a single node called synaptic summation in a biological nervous system as discussed in Section 1.2.1. In addition, these two electrically independent gates make the device have the capability of direct interaction with the back-propagation signal of post-synaptic neuron circuit by G2 without any extra selection device and control circuit which is

the main problem when the neuromorphic system is built with the 2-terminal memristors as discussed in Section 1.3.

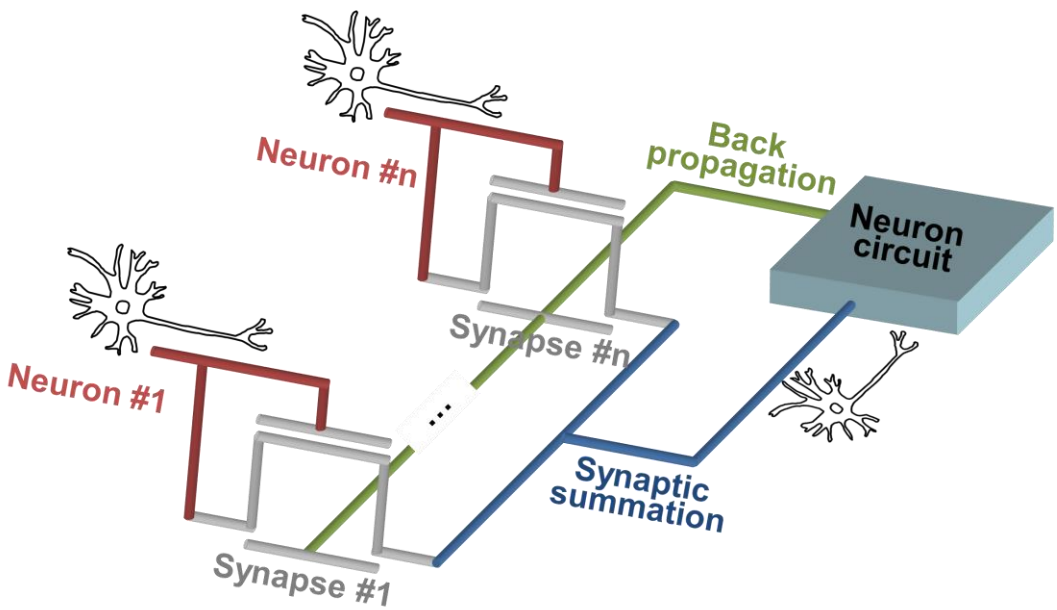


Fig. 2.1. Schematic diagram of the direct connection between the synaptic transistors and a neuron circuit.

2.2. Device Simulation Study

A 2-dimensional (2-D) technology computer-aided design (TCAD) simulation tool of Silvaco Inc. (Atlas ver. 5.20.2.R 2015) has been used to demonstrate that the learning mechanism of the biological system is realized in the silicon-based synaptic transistor [66]. The long-term memory capability is realized by the charge storage layer and G2 placed on the other side of G1 across the channel of a floating body. Consequently, the simulated device structure appears to be a double-gate transistor with 50 nm body thickness and 3 nm gate oxide thickness. The type and concentration of the body doping is *p*-type and 1×10^{18} cm^{-3} boron atoms, respectively, and source/drain doping profiles with Gaussian distribution are adopted with *n*-type and 1×10^{21} cm^{-3} arsenic atoms. Multiple models are included in each simulation, such as Shockley-Read-Hall generation and recombination model [67], [68], Selberherr's impact ionization model [69], and the hot carrier injection model based on Tam's equation [70]. Especially, impact ionization and mobility parameters are calibrated for more accurate electrical estimation in comparison to the previous experimental data. Figure 2.2 shows an overlay of the measured output characteristics and simulated ones under the same design conditions. It is well-fitted and gives a guarantee of the accuracy of following device simulation results.

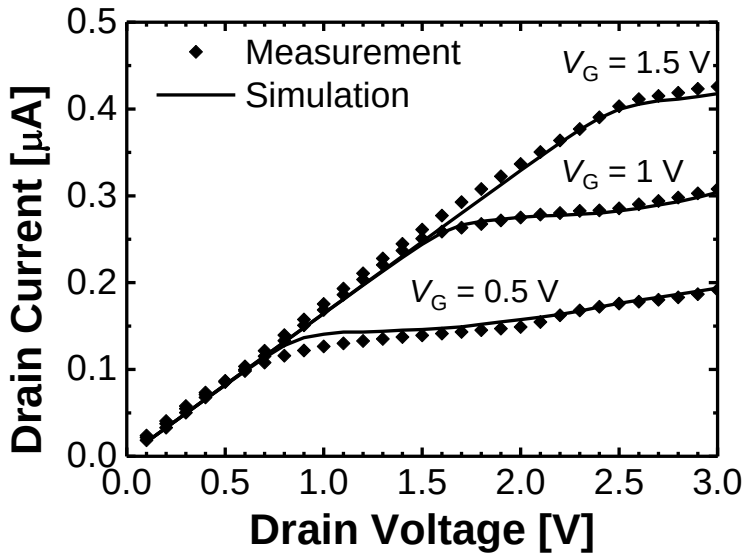


Fig. 2.2. Comparison of measured and simulated output characteristics.

2.2.1. Transition from Short-Term to Long-Term Memory

In the designed device structure and under the given process conditions, it has been verified that the transition from short-term to long-term memory occurs when input pulses with a time width of 50 ns and a repetition interval (T_i) of 1 μs are applied. This is not the unique solution but one of the possible time-dependent input schemes leading to the transition. The bias conditions for the memory transition are $V_{G1} = V_D = 2$ V, and $V_{G2} = -2$ V (here, V_{G1}: G1 voltage, V_D: drain voltage, and V_{G2}: G2 voltage). When input pulses are fed to the device only a few times, the operation of the device can be approximated to that of the floating body devices [71]−[76]. Figure 2.3 shows that excess holes are generated by impact ionization near the top

gate and accumulated near G2. The accumulation of holes results in the increase of the conductivity for a short time since these impact-ionization generated holes increase the potential of the floating body, which corresponds to the short-term potentiation of a biological synapse.

As the pulsing is repeated many times, however, the operation mode of the device changes significantly. The accumulated excess holes do not easily disappear since the repetition interval is not long enough for the generated holes to be fully recombined with electrons. Figure 2.4 shows that the region where impact ionization occurs is expanded down to the bottom as the number of input pulses increases. The impact ionization rate after the sixth input pulse is much higher than after the first input pulse. The reason is that the body potential is increased by the generated holes with the input pulses, hence increasing the output current, I_D, of the device. At the moment when the source-to-body junction becomes forward-biased due to the accumulated holes, the impact ionization occurs near the back-side gate with higher probability, and the generated hot holes begin to enter the charge storage layer over the energy barrier of gate dielectric near the drain end.

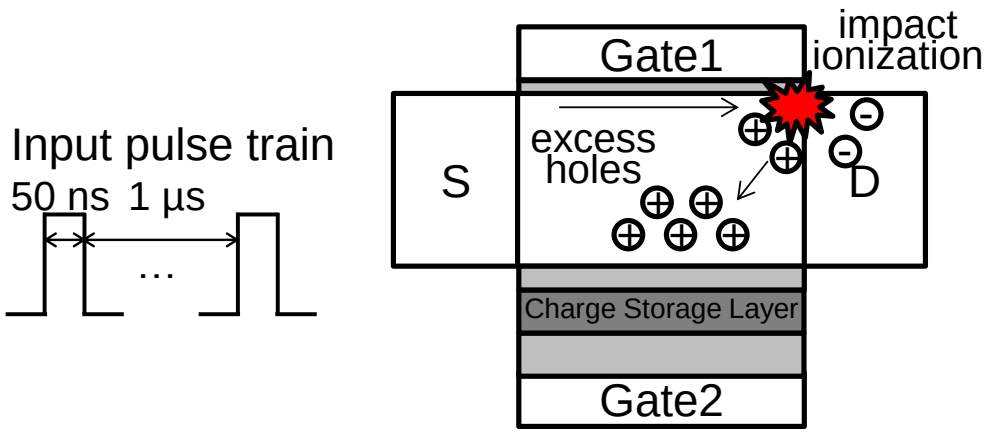


Fig. 2.3. Short-term learning operation of the device.

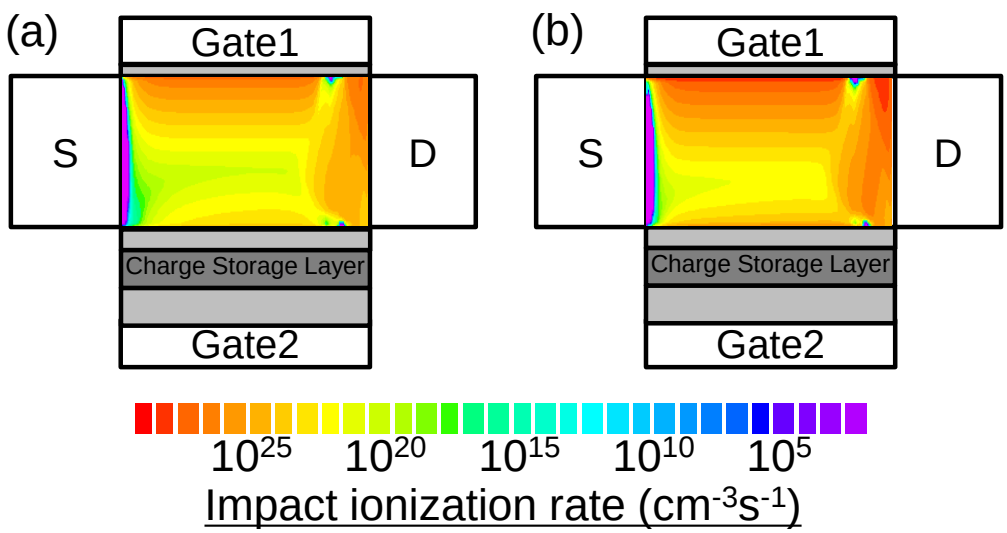


Fig. 2.4. Simulated contours of impact ionization rate after (a) the first input pulse and (b) the sixth input pulse.

After this event, the device is operated like a programmed flash memory cell having charges trapped in the charge storage layer ($Q_{trapped}$), which emulates the biological function of transition from short-term memory to long-term memory in a synapse as shown in Fig. 2.5 [21], [22]. The excess holes can be compared to the temporally increased cAMP in a biological system because both of them decay without the next input or serotonin and the long-term memory arises when the accumulation of both of them exceeds threshold points. Also, the hot hole injection to the charge storage layer corresponds to the growth of a biological synapse because they bring out the long-term increase in channel conductance and biological synaptic weight.

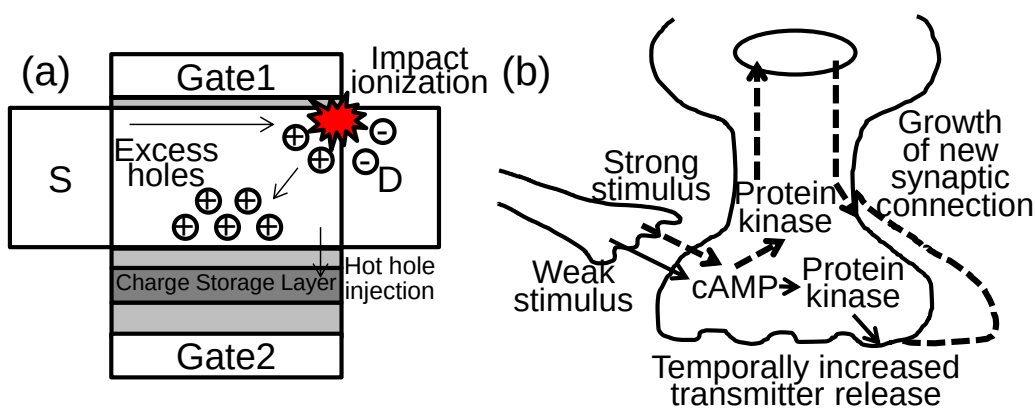


Fig. 2.5. Schematic views of how short- and long-term memories are formed in (a) the device and (b) a biological synapse.

Figure 2.6(a) shows that the hot hole injection into the charge storage layer starts from the seventh input signal pulsing. It is noteworthy that the transition in the device occurs without any change in the bias condition. It depends on the timing scheme only. Figure 2.6(b) shows the retention of stored information with N under the read bias condition. The data reading was conducted at V_D = 0.1 V which is assumed to be small enough to minimize the interruption of the device field and carrier distributions. Here, $I_{read}/I_{initial}$ is defined as the ratio between the learned state where hot carriers are injected to the charge storage layer and the initial state where there is no carrier injection. When the number of input pulses (N) is no more than 7, the information is lost in several milliseconds. On the other hand, if N is increased above 7, long-term memory is formed and the stored information remains for more than 100 seconds. The flat tails of the upper four curves (N = 7~10) indicate that the information is stored into a long-term memory.

In order to examine the effect of input pulses on the synaptic learning, input pulses with longer width of 100 ns and the same T_i of 1 μs are applied to the device at first. The transition to long-term memory occurs at the third input signal pulsing as shown in Figs. 2.7(a) and 2.7(b). It is earlier than when the pulse width is 50 ns. Understandably, this comes from the fact that more excess holes are generated in a single pulse with longer width compared to shorter one.

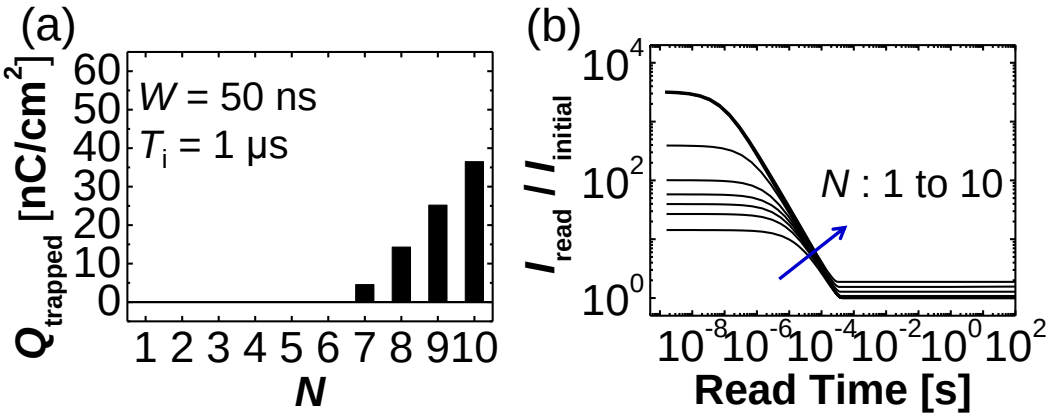


Fig. 2.6. Synaptic device operation. (a) Simulated trapped charges in the charge storage layer as a function of N and (b) read retention characteristics according to the number of the applied pulses with T_i of 1 μs.

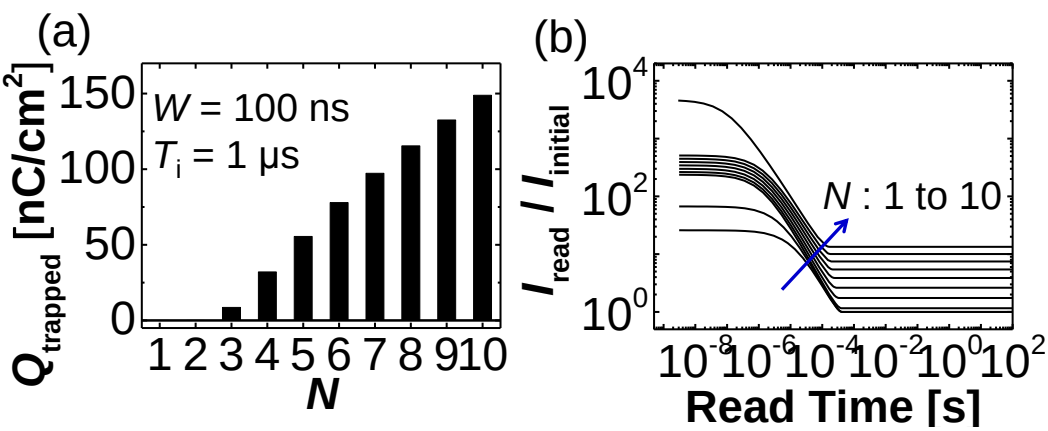


Fig. 2.7. Simulated learning characteristics with pulse width of 100 ns. (a) Trapped charges and (b) read retention characteristics.

In addition, the time response of the device to different T_i is studied in depth in order to mimic the biological system more closely. For an input with a shorter interval (T_i = 0.1 μs), the transition needs fewer number of input pulses than when T_i = 1 μs as shown in Fig. 2.8(a). This is because the next input for the shorter interval is applied to the device while more excess holes are remaining in the floating body. Moreover, for an input with a longer interval (T_i = 10 μs), hole injection does not occur in spite of the increased N as shown in Fig. 2.8(b). Furthermore, there is no

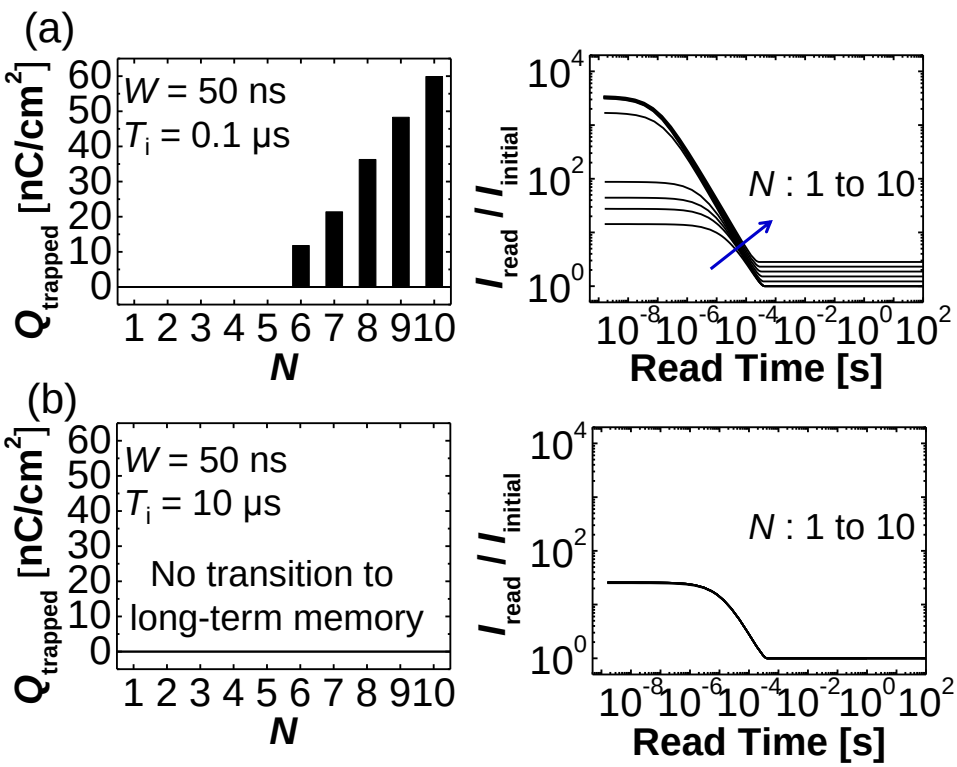


Fig. 2.8. Simulated learning operations of the device according to the number of the applied pulses with T_i of (a) 0.1 μs and (b) 10 μs.

transition from short-term to long-term memory because the hole concentration in the floating body decreases back to the initial value during the increased time interval between pulses, which obviously realizes time-dependent synaptic learning of the biological system through an electronic system.

2.2.2. Spike-Timing Dependent Plasticity Characteristics

STDP characteristics are emulated with time difference (Δt) between pre- and post-synaptic spikes. Here, the pre-synaptic spikes are applied to G1 and drain of the devices and the post-synaptic spikes are applied to G1 of the devices. Spikes with T_i of 1 μs and duration of 0.5 μs are applied to the device ten times. In case of positive Δt, hot hole injection occurs by the same mechanism explained previously since negative V_{G2} is applied while impact ionization rate is high due to large V_D as shown in Fig. 2.9(a). However, hot electrons generated by the impact ionization begin to enter the charge storage layer with a negative Δt because V_{G2} is positive while large V_D is applied as shown in Fig. 2.9(b). This is analogous to the depression process in the biological systems. It is noticeable that these contrasting results are obtained by nothing but the spike timing scheme.

Figure 2.10 depicts the conductivity of the channel after 10 spikes as a function of Δt. It is confirmed that the obtained STDP characteristics bear great deal of resemblance with those of the biological synapses [23]−[28]. Regardless of the

polarity of Δt, smaller $|\Delta t|$ substantially increases potential difference between drain and G2, which results in the increase of hot carrier injection rate. For this reason, the $I_{read}/I_{initial}$ shows a large deviation from the unity near $\Delta t = 0$ in Fig. 2.10.

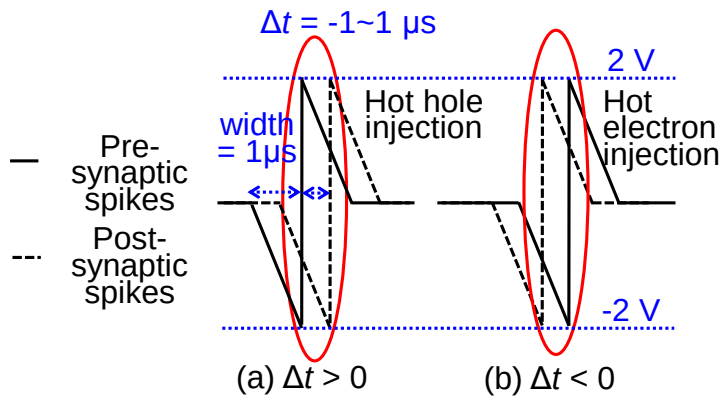


Fig. 2.9. Timing diagrams of biasing scheme of pre- and post-synaptic spikes with (a) positive Δt and (b) negative Δt.

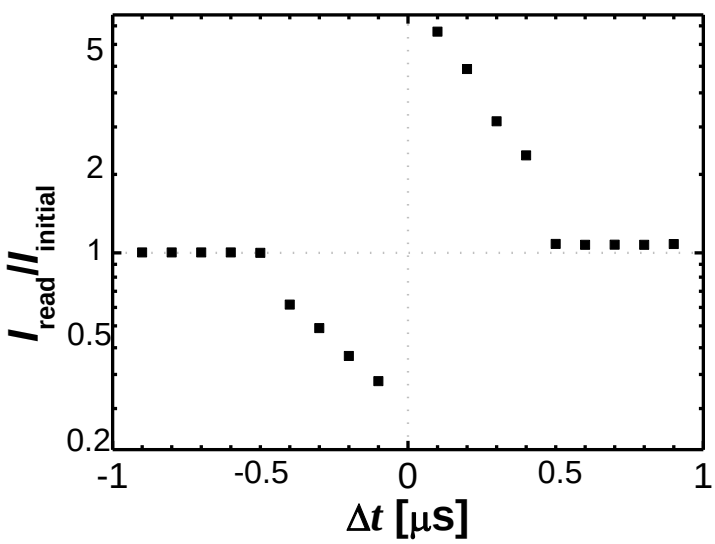


Fig. 2.10. Simulated STDP characteristics after 10 triangular spikes.

2.3. Circuit Simulation Study

In order to generate the triangular-shaped spikes ranging in amplitude from supply voltage (V_{dd}) to $-V_{dd}$ as shown in Fig. 2.9, an integrate-and-fire neuron circuit and the connection between the neuron circuit and the synaptic devices are designed as shown in Fig. 2.11. The integration part received signals from two parts: excitatory synapse part and inhibitory synapse part. Transferred signals from the pre-synaptic neuron circuits are added through excitatory synaptic devices but taken out through inhibitory synaptic devices at a node of capacitor. Therefore, signals through the excitatory synaptic devices contribute to the firing of the postsynaptic neuron circuit by integrating charges at the capacitor, while ones through the inhibitory synaptic devices suppress the firing by discharging the capacitor. When the node voltage of the capacitor (V_C) as a result of the integration of charges at the capacitor exceeds the threshold voltage of NMOS (N_1) at the next stage included in the firing part, the output voltage (V_{out}) starts to fall down to $-V_{dd}$ because the source of N_1 is connected to $-V_{dd}$ and it is transferred to V_{out}. At the same time, a low output voltage of the first inverter coming from V_C exceeding over the threshold voltage of the inverter boosts V_C up to V_{dd} immediately. And then, V_{out} rises up to V_{dd} later as a result of the delay of the two-stage inverters. Finally, the firing of the neuron circuit turns on N_2, leading V_{out} to return to ground, the initial state, by discharging the capacitor.

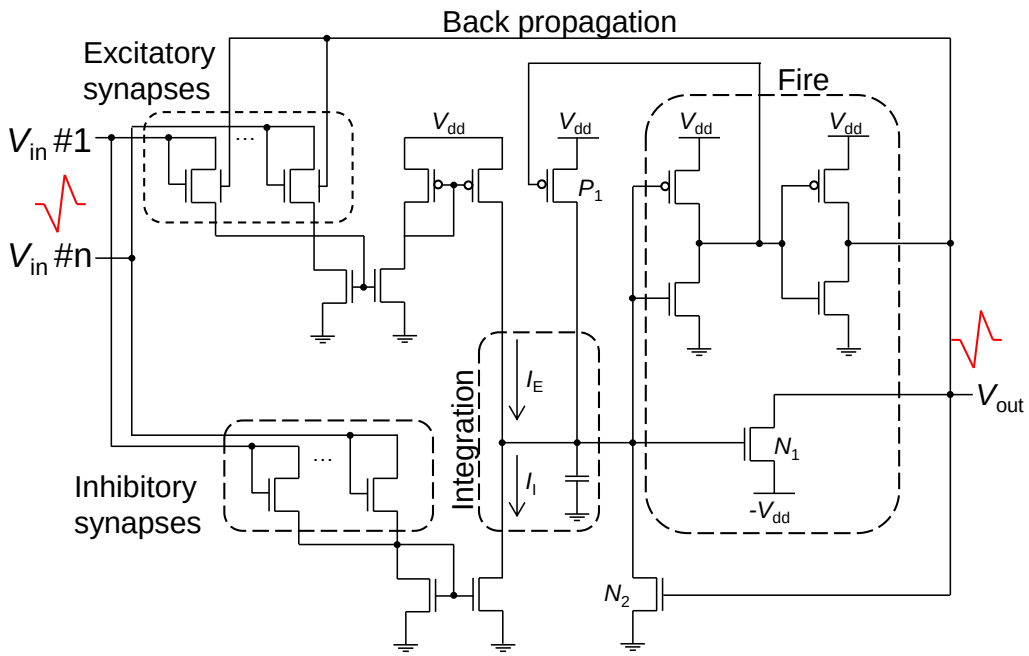
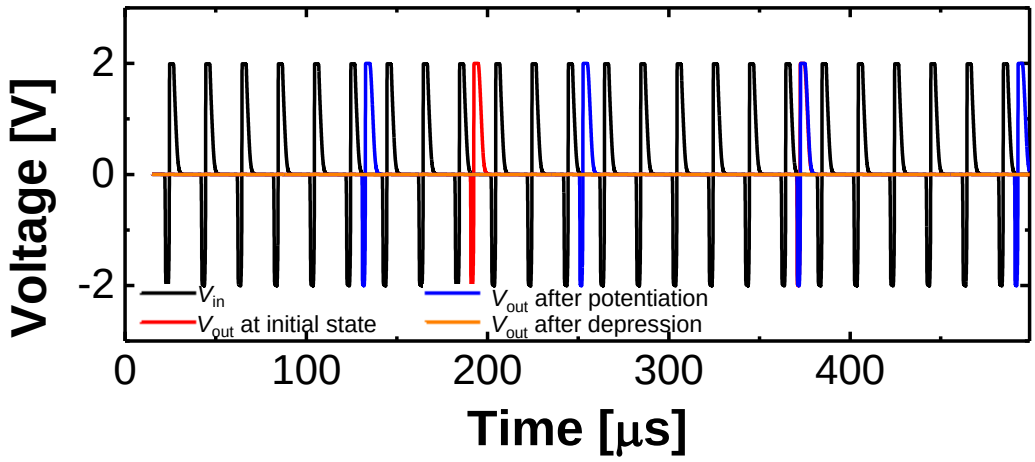


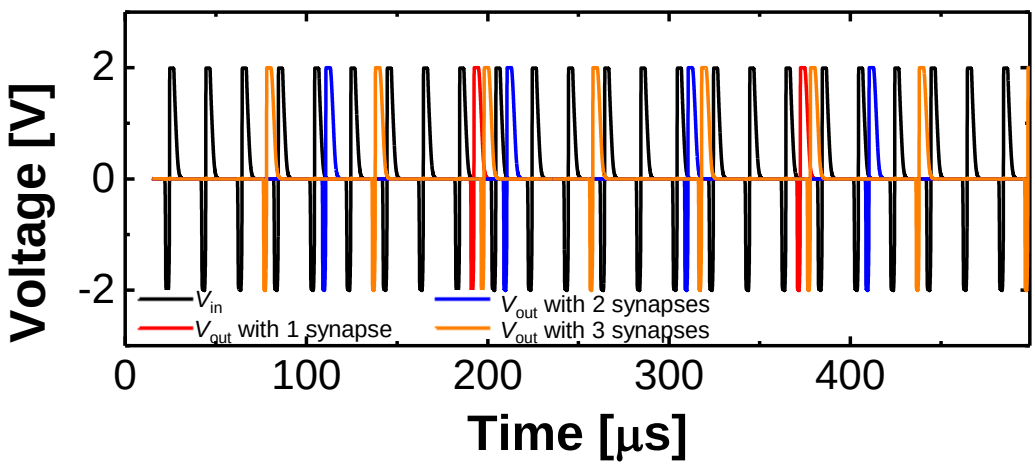
Fig. 2.11. Neuromorphic system composed of the synaptic devices and the neuron circuit.

Spice simulation studies of this neural network are conducted using a circuit simulation tool of Silvaco Inc. (SmartSpice ver. 4.18.16.R 2016) [77]. At first, the firing characteristics of the postsynaptic neuron circuit, which is depending on the state of the connected synaptic device, are plotted in Fig. 2.12(a). Here, the neuron circuit is connected with an excitatory synaptic device and an inhibitory one which has a slightly higher threshold voltage compared to the excitatory one. The postsynaptic neuron circuit fires after receiving input signals 9 times when the

excitatory synaptic device is at the initial state, but only after 6 times when it is potentiated. This is because the increased conductance of the excitatory synaptic device makes the current from it (I_E) larger compared to the current from the inhibitory synaptic device (I_I), leading to a higher firing rate of the postsynaptic neuron circuit. Besides, the postsynaptic neuron circuit does not fire when the excitatory synaptic device is depressed because the inhibitory synaptic device has a higher conductance than the excitatory synaptic device. In this condition, charges cannot be accumulated at the capacitor because I_I is larger than I_E, meaning that V_C cannot exceed the threshold voltage of the transistors at the next stage. Figure 2.12(b) shows the firing characteristics depending on the number of connected excitatory synaptic devices. As the number increases from 1 to 3, the required the number of input signals decreases from 9 to 5 and 3, respectively. This is because the transferred currents of the excitatory synaptic devices in parallel are integrated at the common source node, increasing I_E relative to the number of the connected excitatory synaptic devices. This could correspond to the spatial summation of a biological nervous system as discussed in Section 1.2.1. These simulation results indicate that the firing rates of the postsynaptic neuron circuits are strongly related to the conductance and the number of connected synaptic devices like a biological nervous system.



(a)



(b)

Fig. 2.12. Simulated transient characteristics of V_{out}. (a) Depending on the number of the connected synaptic devices. (b) Depending on the state of the synaptic device.

Chapter 3. Device Fabrication

3.1. Process Design and Fabrication Flow

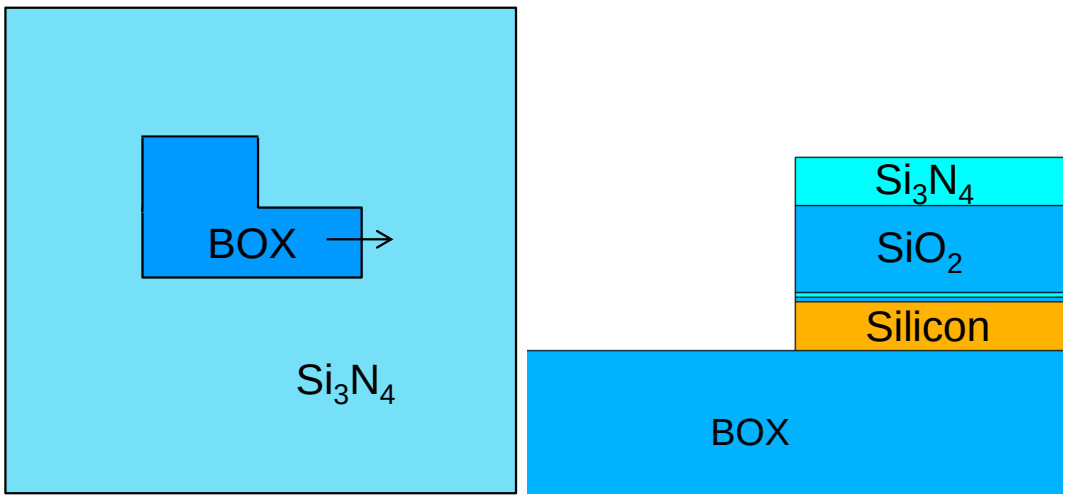
In order to fabricate the synaptic transistor with asymmetric dual-gate structure, the fabrication sequence follows the process flow shown in Fig. 3.1. A process simulation tool of Silvaco Inc. (Athena ver. 5.20.0.R 2012) is used to estimate the experimental results more accurately [78]. First, a hard mask stack is deposited and the 1st active patterns are etched as shown in Fig. 3.1(a). Because two gates are formed in sequence, the active region which contacts G1 is formed first. The hard mask stack is comprised of silicon dioxide (SiO_2) and silicon nitride (Si_3N_4) which is used for selective etching with poly-silicon and stopping layer of chemical mechanical planarization (CMP) process, respectively. Steep etch slope is required because SiO_2 sidewall is formed at the region where the hard mask stack is removed.

After forming G1 stack, CMP process is done to leave the poly-silicon layer only at the G1 region as shown in Fig. 3.1(b). The CMP slurry which has good selectivity between doped poly-silicon and Si_3N_4 is required to protect the height of the hard mask stack, which allows a margin for gate splitting process later. And then, the Si_3N_4 layer is removed with dry etch process due to bad selectivity of phosphoric

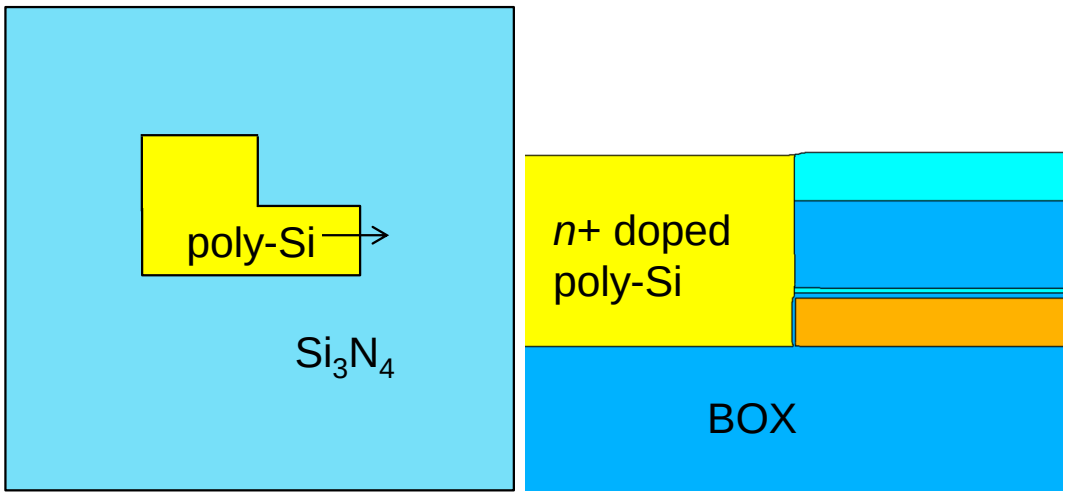
acid (H_3PO_4) wet etch process between doped poly-silicon and Si_3N_4 and the SiO_2 layer is removed with hydrogen fluoride (HF) wet etch process which shows good selectivity between silicon and oxide (Fig. 3.1(c)).

Thin sidewall is formed through deposition and dry etch process of medium temperature chemical vapor deposition oxide (MTO) which has slow deposition rate, hence forming the thin sidewall (Fig. 3.1(d)). The unnecessary sidewall is removed using HF wet etch process to suppress the parasitic capacitance around G1 region as illustrated in Fig. 3.1(e). Then, a silicon fin is formed using dry etch process (Fig. 3.1(f)). During this process, the surface of the 2nd active region is formed. The fin is achieved with little loss of oxide sidewall because hydrogen bromide (HBr) plasma etch process has good selectivity between silicon and SiO_2 and the fin width is the same as the thickness of oxide sidewall.

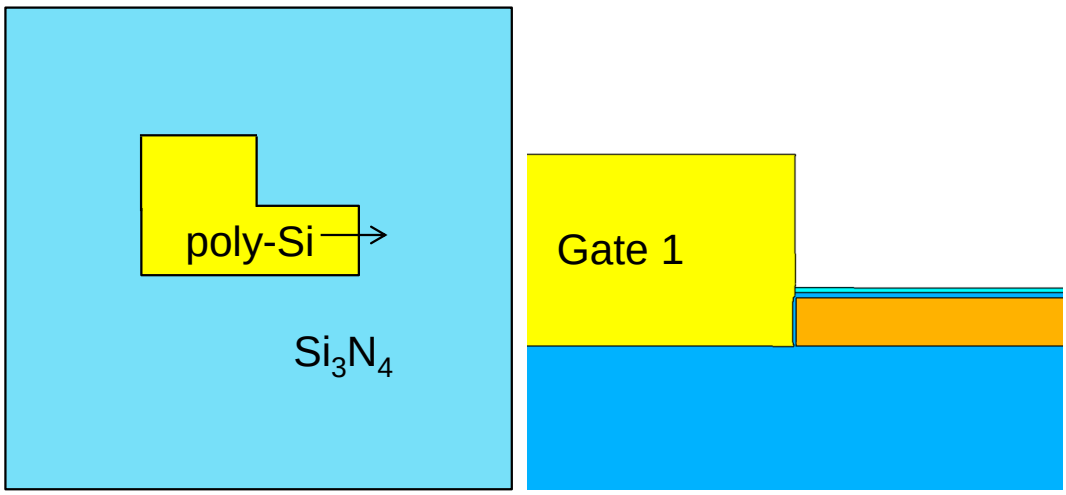
After the fin formation, G2 stack comprised of oxide/nitride/oxide/doped poly-silicon is deposited over the fin and G1 using low-pressure chemical vapor deposition (LPCVD) as shown in Fig. 3.1(g). The CMP process makes a difference in the thickness of doped poly-silicon and this difference separates two gates by dry etchback process (Fig. 3.1(h)). The gate length is defined at the same time by photolithography and dry etch process. The overall process flow is summarized in Fig. 3.1(i).



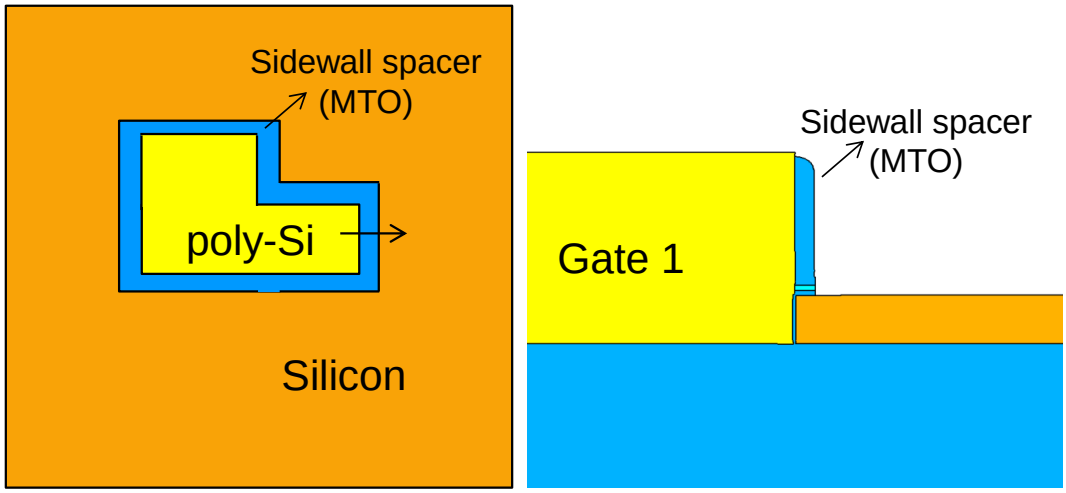
(a) Deposition of hard mask and patterning



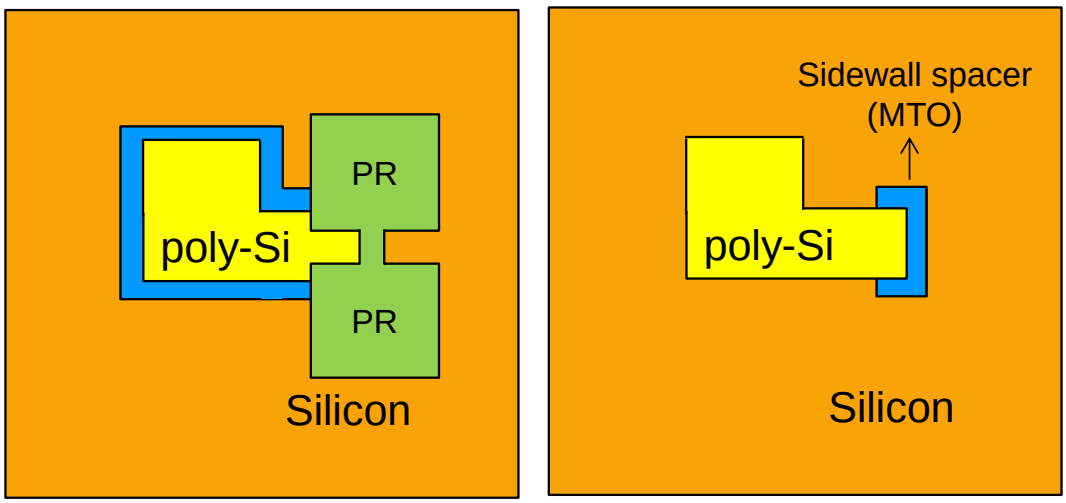
(b) Formation of G1 through CMP



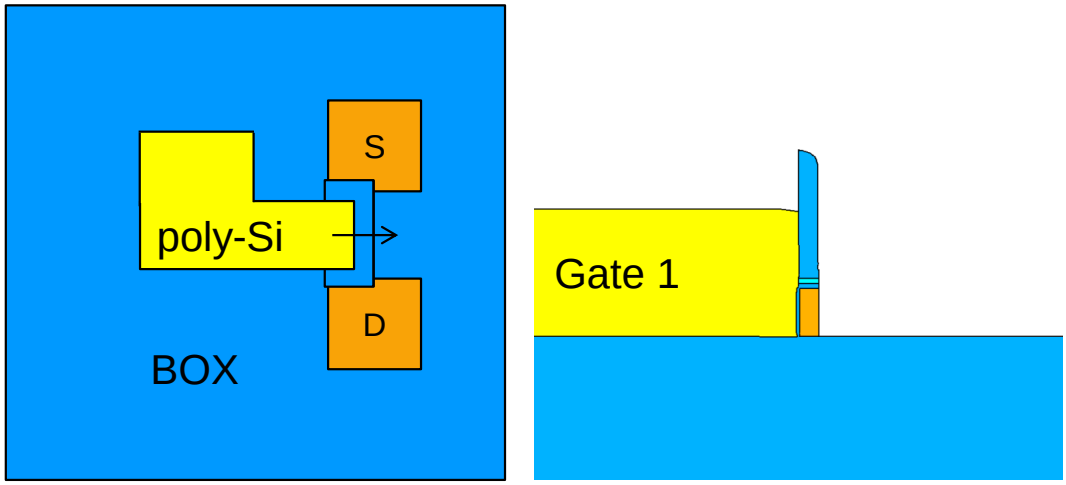
(c) Removal of hard mask



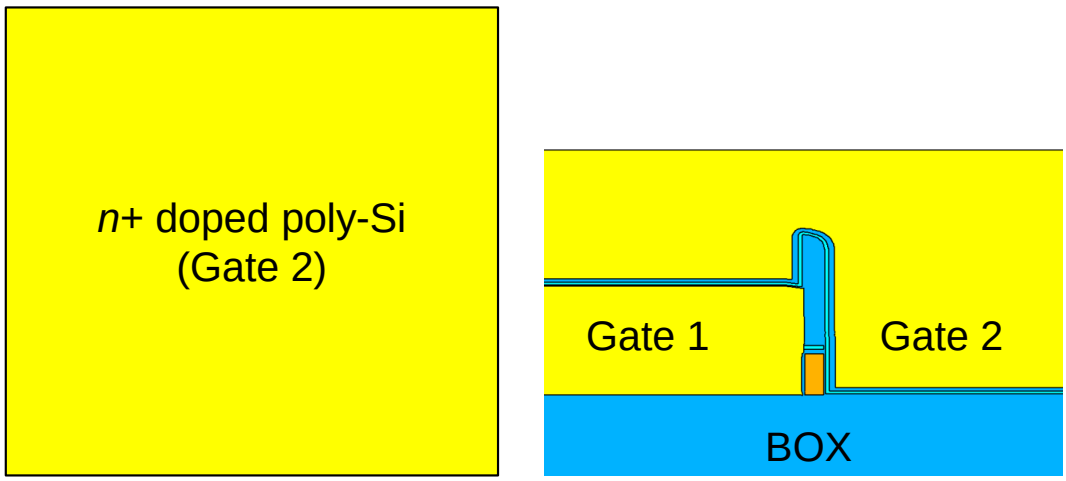
(d) MTO sidewall formation



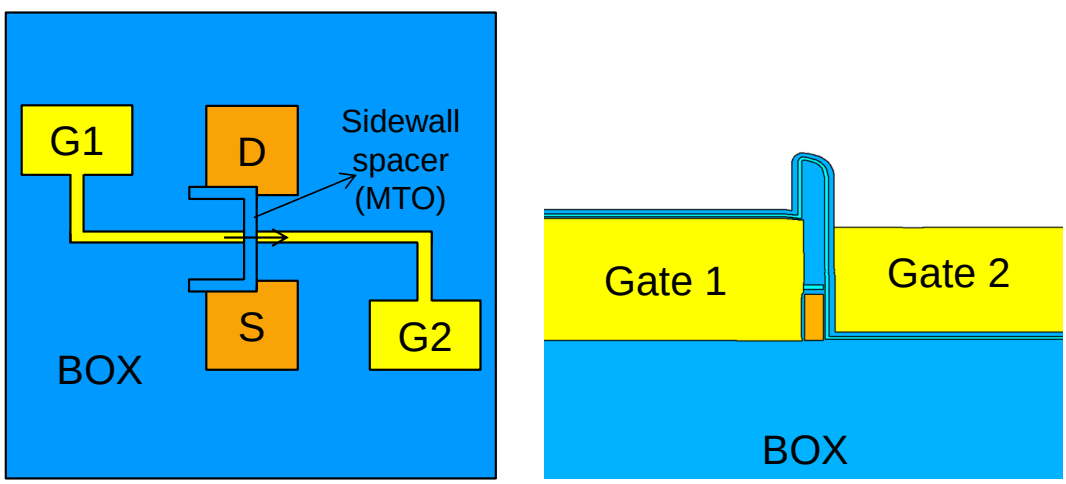
(e) Sidewall cutting



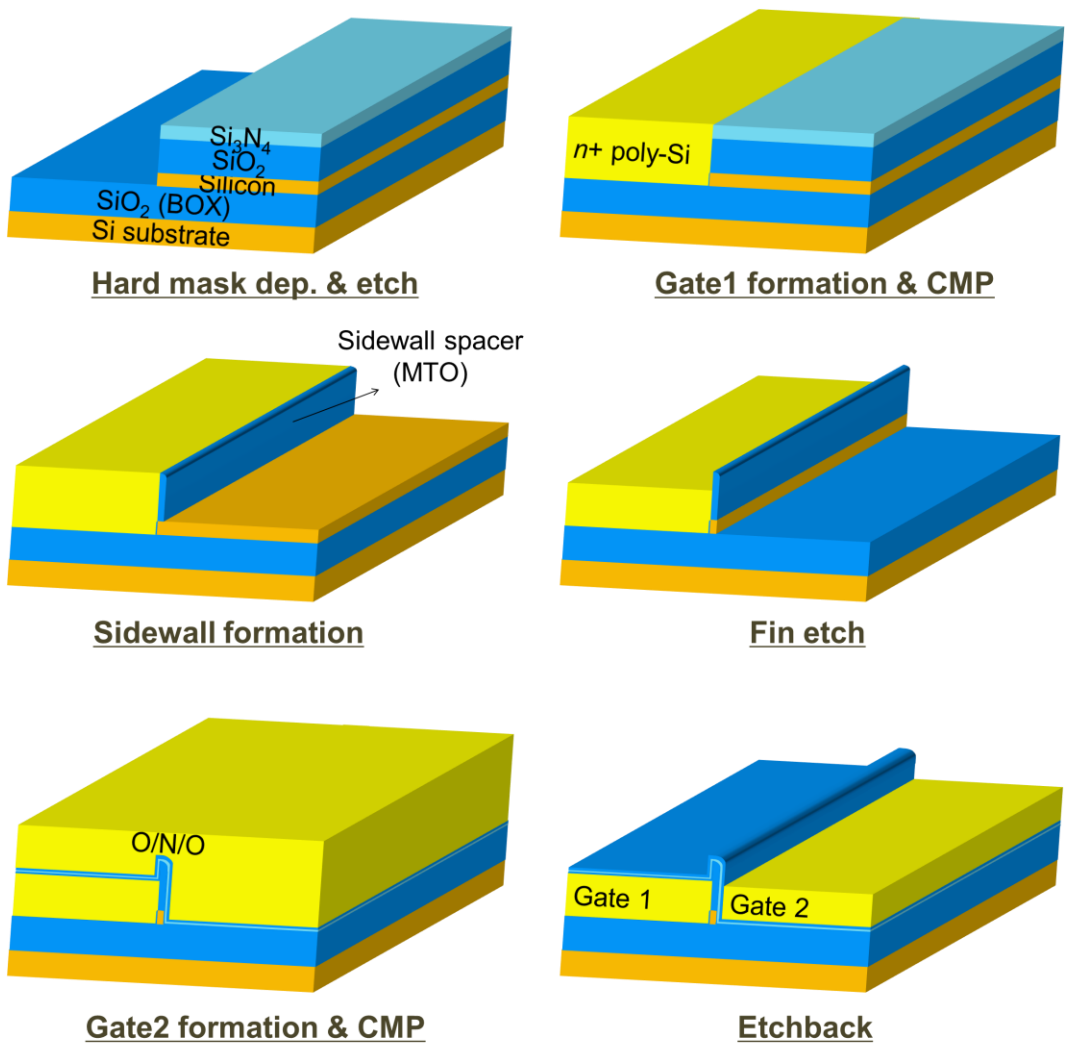
(f) Fin channel formation



(g) G2 stack deposition and CMP



(h) Etchback for gate splitting and gate patterning



(i) Overall view of the fabrication flow.

Fig. 3.1. Device fabrication flow.

3.2. Experimental Results

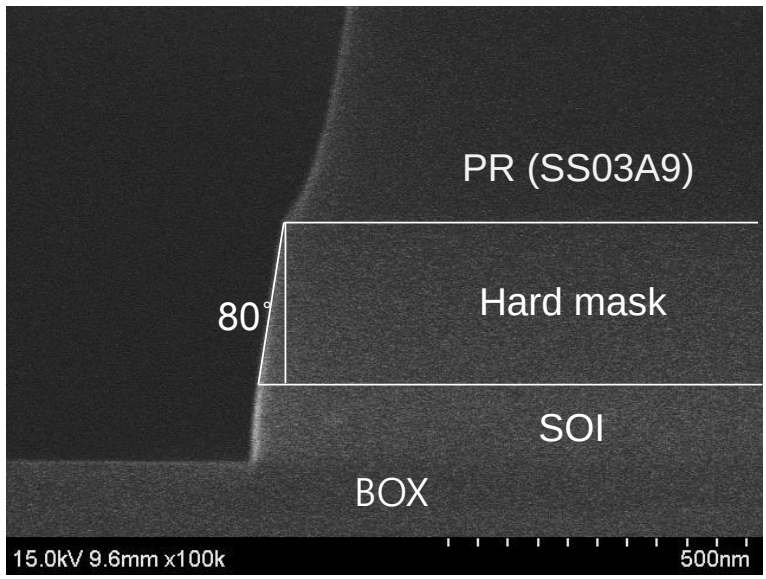
In Section 3.2, the experimental conditions and issues of the key process steps described in Section 3.1 are explained in more detail. Most of fabrication procedures were carried out with the fabrication facilities of the Inter-university Semiconductor Research Center (ISRC) at Seoul National University except the deposition of doped poly-silicon layer which was done with the help of National Nano-Fab Center (NNFC) located at Daejeon.

3.2.1. Deposition of Hard Mask and Patterning

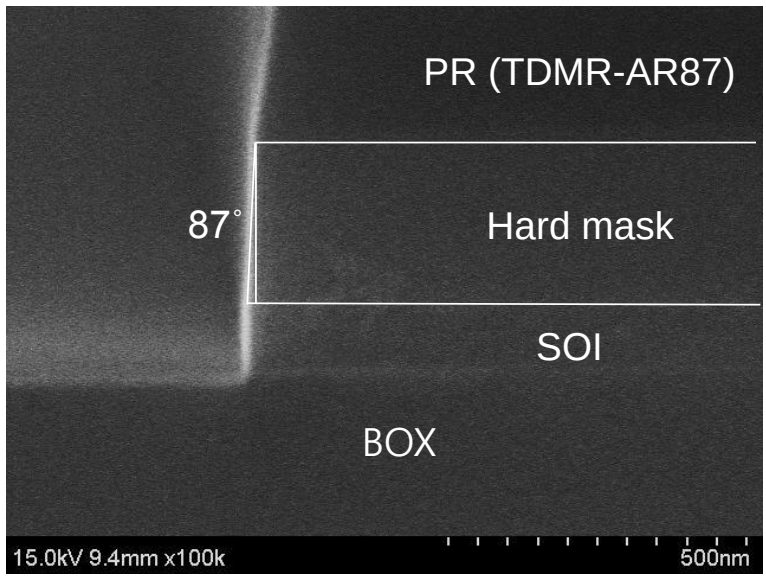
First, BF_2^+ ions for body doping are implanted with a dose of 2×10^{11} and 2×10^{12} cm^{-2} over a silicon-on-insulator (SOI) wafer fabricated by SIMOX technology [79]. The thicknesses of SOI layer and buried oxide (BOX) layer are 100 and 375 nm, respectively. A hard mask stack is deposited and G1 region etched. The hard mask stack is comprised of 300-nm oxide and 100-nm nitride used for selective etching with poly-silicon and stopping layer of CMP process, respectively. The oxide layer is deposited by LPCVD using tetraethoxysilane ($Si(OC_2H_5)_4$, TEOS) gas at 710 ℃ and 250 mTorr, and the nitride layer is deposited by LPCVD in an ambient of dichlorosilane (SiH_2Cl_2) of 30 sccm and ammonia (NH_3) of 100 sccm at 785 ℃ and 200 mTorr.

At the etch process after then, it is very important to obtain a steep etch slope because it is transmitted at the angle of the sidewall spacer for fin channel formation. If the hard mask is etched gradually, the width of a fin channel would be thicker than the spacer thickness, which means the channel width cannot be controlled finely. When SS03A9, which is based on phenol formaldehyde resin, is used as a photoresist (PR), the hard mask is etched with a slope of 80 degree as shown in Fig. 3.2(a). This is because of the interaction between nitrogen atoms in the Si_3N_4 layer and acid which is produced and diffused during exposure process as the by-product of SS03A9, called PR poisoning effects [80]–[82]. It leads to a tail located close to the edge of the exposed area, resulting in the gradual etch slope of 80 degree.

In order to avoid these PR poisoning effects, TDMR-AR87 replaces SS03A9 as a PR. In addition, oxygen plasma treatment is performed to oxidize the Si_3N_4 layer slightly. As a result, a pretty steep etch slope of 87 degree, nearly close to perpendicular, is achieved as shown in Fig. 3.2(b). After etching the hard mask and the active region, chemical dry etcher (CDE) process is carried out to remove plasma-damaged silicon active region during the previous etch step using tetraflouromethane (CF_4) gas of 50 sccm and oxygen (O_2) gas of 30 sccm, suppressing the reduction of carrier lifetime due to the defects formed during the etch step.



(a)



(b)

Fig. 3.2. Cross-sectional SEM images after patterning depending on PR materials. (a) SS03A9. (b) TDMR-AR87.

3.2.2. Formation of G1 through CMP

After 3.5 nm oxide and 587 nm doped poly-silicon fills that region, CMP process is done. The oxide layer is thermally grown through dry oxidation process at 800 ℃ for 30 sec, and the in-situ phosphorus doped poly-silicon is deposited by LPCVD using silane (SiH_4) gas and phosphine (PH_3) gas. The CMP process is done for 190 sec after that in order to remove G1 stack at the other regions, except G1 region as shown in Fig 3.3. Here, the nitride layer is used as a stopping layer of CMP process since the slurry used in this study has an excellent selectivity between nitride and silicon (about 1:100). Figure 3.4 shows the wafer maps representing the remaining thicknesses of the nitride layer and poly-silicon depending on the location in a test wafer. They are measured by observing cross-sectional scanning electron microscope (SEM) images for each location. The remaining thickness of the nitride layer which was initially about 100 nm is 67 to 89 nm after the CMP process, and for the poly-silicon layer 384 to 420 nm. The uniformity of the remaining thickness of the nitride and poly-silicon layer is 8.1% and 3.1%, respectively, confirming that high uniformity of each layer even after the CMP process comes from the slurry having a good selectivity between those two layers.

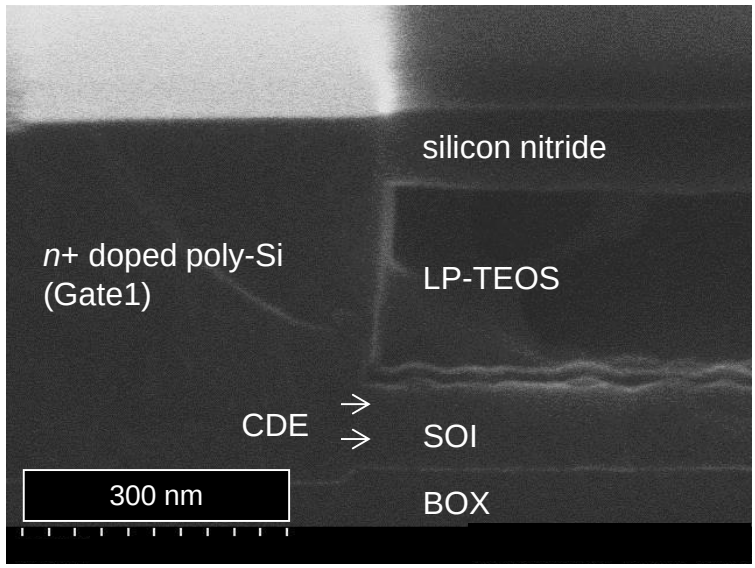


Fig. 3.3. Cross-sectional SEM image after G1 formation.

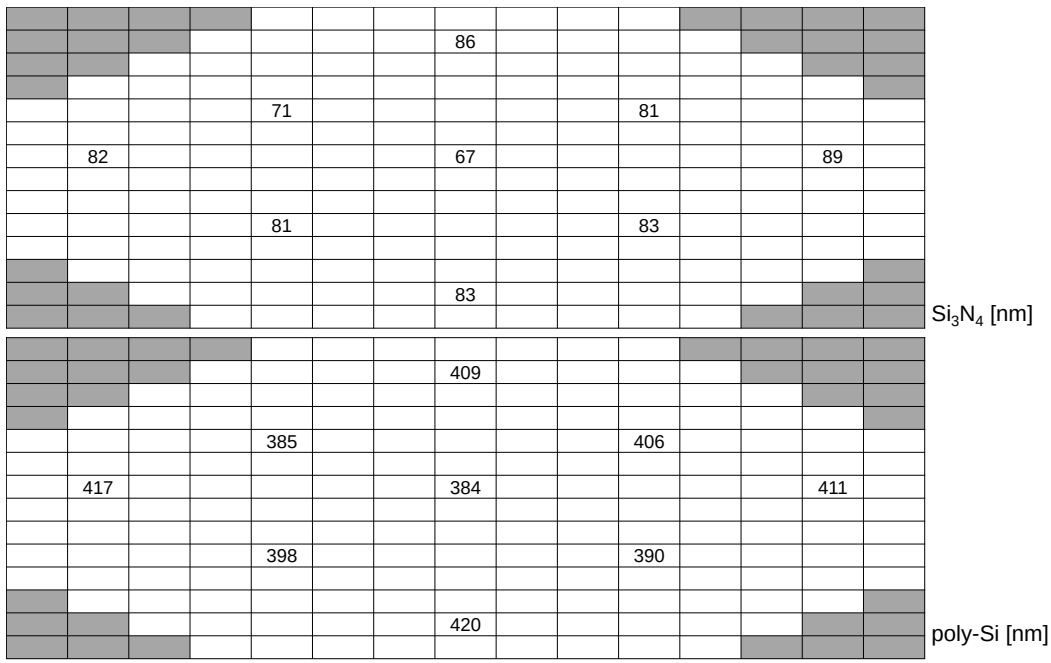
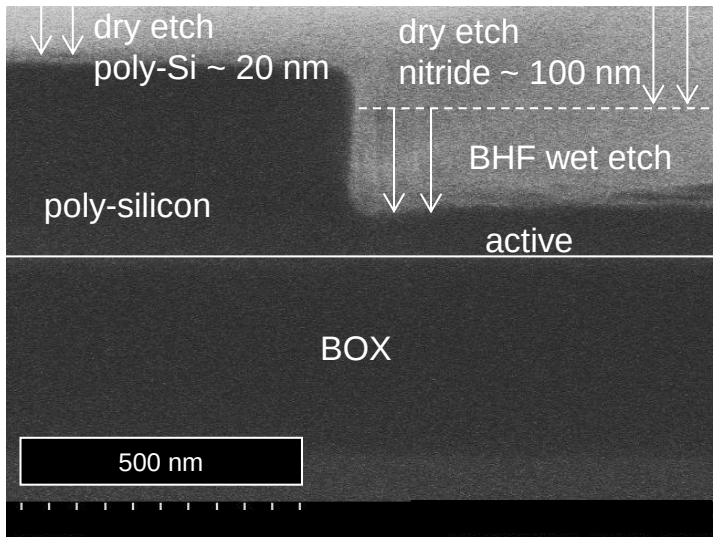


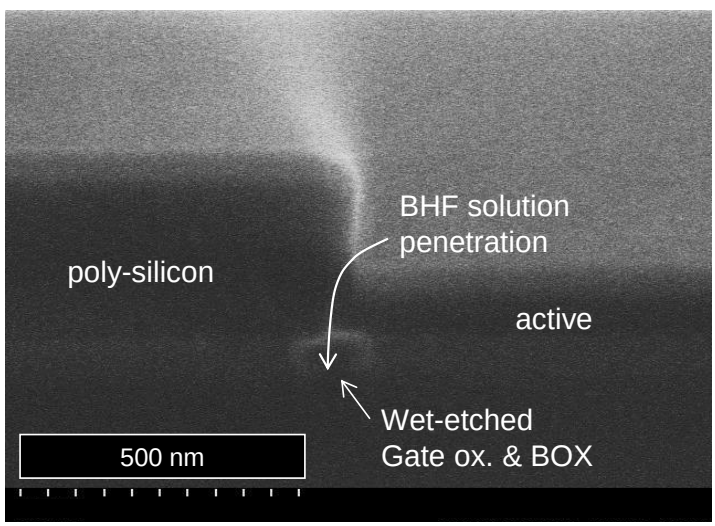
Fig. 3.4. Wafer maps of remaining thickness of Si_3N_4 layer and poly-silicon layer.

3.2.3. Removal of Hard Mask

During the removal process of the hard mask, the most import thing is to protect the poly-silicon layer located at G1 region as much as possible so that G1 can fully cover the side of the active region even after the etchback process step for splitting G1 and G2 electrically later. Figure 3.5(a) shows an overall view of the removal process in a SEM image. Even though H_3PO_4 solution in a heated bath is commonly used to strip silicon nitride films, the nitride layer of the hard mask in this study is removed with dry etch process because H_3PO_4 solution has a relatively poor etch selectivity between silicon nitride and n^+ doped poly-silicon [83]. The target etching thickness is 130 nm which is over the nitride layer thickness, 100nm, to remove the layer perfectly all over the wafers, leading to the fact that the top 20 nm of poly-silicon layer is etched at the same time during this dry etch process. Then, the lower TEOS layer is removed with buffered HF wet etch process (NH_4F : HF = 7:1) which shows good selectivity between silicon and oxide. In fact, very thin oxide and nitride layers, 5 nm and 10 nm, respectively, are placed between the silicon active and the hard mask to prevent the penetration of BHF solution. The gate oxide and BOX layers can be wet-etched if those layers do not block the penetration as shown in Fig. 3.5(b).



(a)



(b)

Fig. 3.5. Cross-sectional SEM images after the hard mask removal step. (a) When successfully done. (b) When BHF solution penetrated into BOX layer.

3.2.4. Fin Channel Formation Using Sidewall Spacer

Thin MTO sidewall spacer of 53 nm is formed through the deposition and dry etch process as shown in Fig. 3.6. MTO film is deposited by LPCVD at 782 ℃ and 350 mTorr using SiH_2Cl_2 of 40 sccm and N_2O of 160 sccm as the precursors. A silicon fin is then achieved through HBr plasma etch process using the sidewall spacer as a hard mask because HBr plasma etch process has good selectivity between silicon and oxide. The fin's width is defined almost the same as the thickness of oxide sidewall spacer because the hard mask stack is etched steeply, 88 degree, as a result of the fine control in the etch slope at the previous step.

Figure 3.7 shows the wafer maps representing the remaining thicknesses of the poly-silicon layer, which would become G1, depending on the location in a test wafer. Because doped poly-silicon is etched at the same time when the fin is formed, a sufficient thickness is required to cover up the silicon active region after dry etch process for the fin formation. Although dry etch process is done, the active region is well surrounded by poly-silicon layer with the minimum thickness (174 nm), which is pretty thicker than the thickness of the active SOI region (100nm). After that, CDE process is also carried out to remove damaged active region as described in Section 3.2.1.

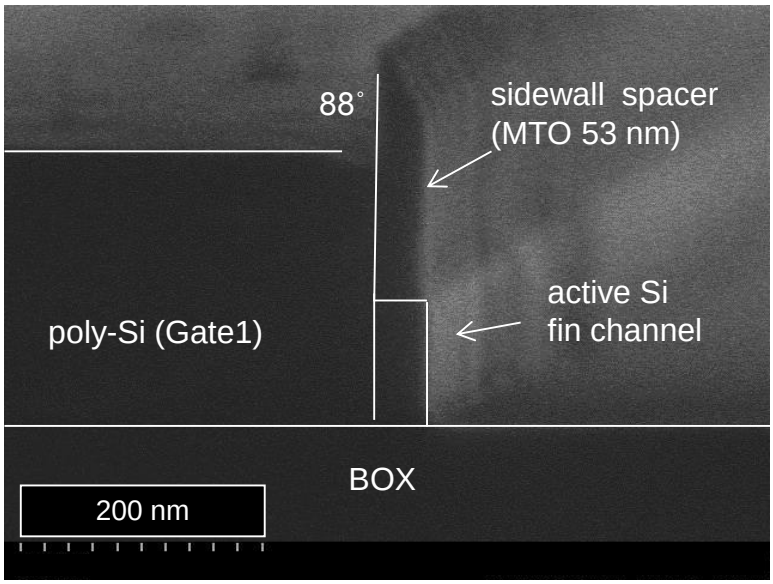


Fig. 3.6. Cross-sectional SEM image after the removal of the hard mask.

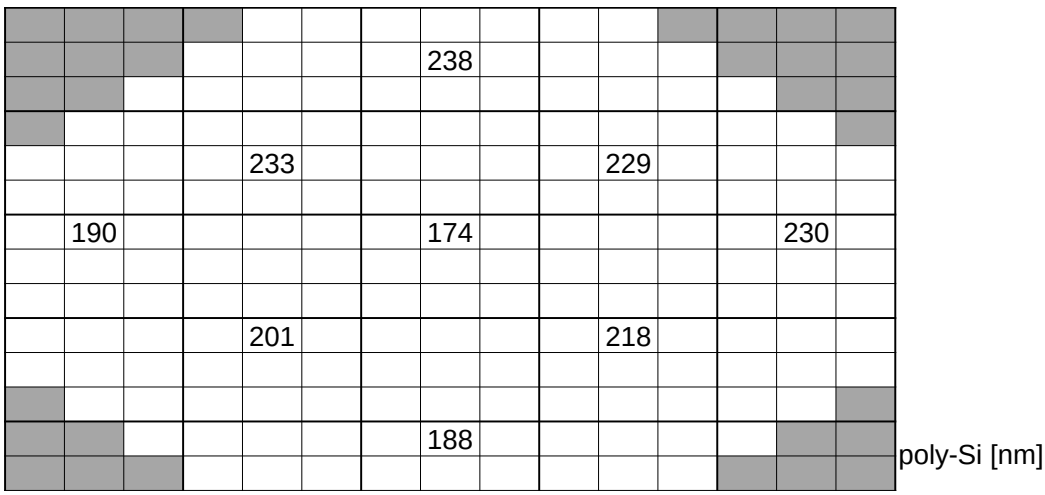


Fig. 3.7. Wafer maps of remaining thickness of poly-silicon layer after the fin formation.

3.2.5. Gate Splitting through CMP and Etchback Processes

After the fin formation, G2 stack comprised of 3.5 nm oxide/5.5 nm nitride/8.6 nm oxide/doped poly-silicon is formed over the fin and G1 region using dry oxidation and LPCVD processes at the same conditions described in previous sections. The CMP process makes a difference in the thickness of doped poly-silicon layer depending on the location, which means that the thickness of the poly-silicon layer deposited over G1 region becomes thinner than over G2 region because the layer is flattened by the CMP process. The later-deposited doped poly-silicon acts as G2. Figure 3.8, the cross-sectional high resolution transmission electron microscopy (HR-TEM) image, confirms that the two gates, G1 and G2, are finally separated after the CMP and dry etchback process. During the etchback process, MTO sidewall is rarely etched out because of good selectivity of HBr plasma etch process between silicon and silicon oxide.

The length of G1 and G2 is defined at the same time by photolithography and dry etch process with values ranging from 0.5 μm to 5 μm. After 5×10^{15} cm^{-2} of P^{+} ions are implanted to make source/drain junction, rapid thermal annealing is carried out to activate dopants at 950 ℃ for 5 seconds. The back-end-of-line (BEOL) process is composed of inter-layer dielectric (ILD) deposition, contact hole formation and metallization.

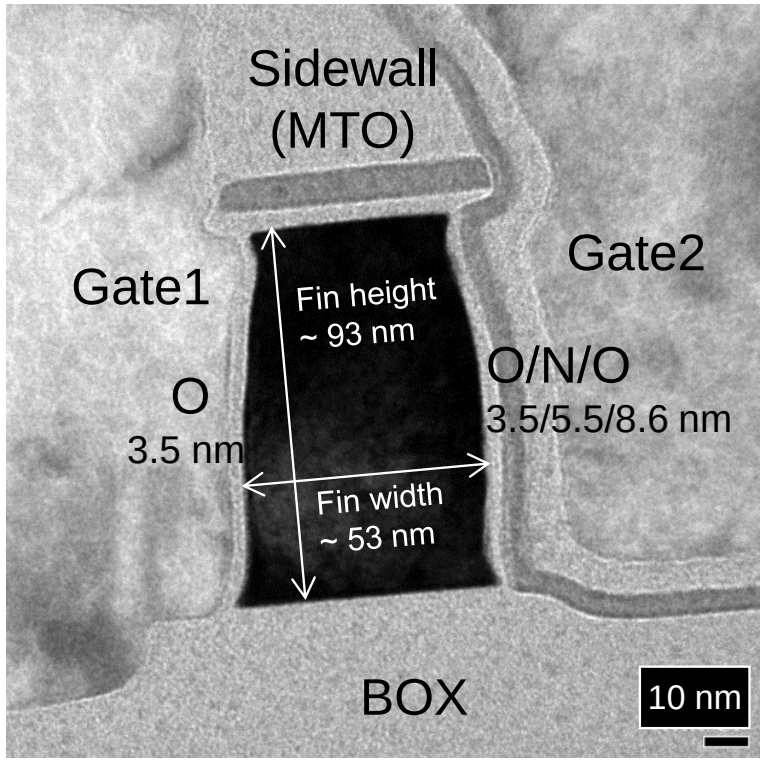


Fig. 3.8. Cross-sectional TEM image after the gate splitting process.

Chapter 4. Device Characteristics

4.1. Field-Effect Transistor Characteristics

Firstly, the two gates are confirmed to be well separated. The transfer characteristics of G_1 and G_2 are shown in Fig. 4.1. The former shows the characteristics of conventional MOSFET with 89 mV/decade whereas the latter shows higher threshold voltage and subthreshold swing with 264 mV/decade when the other gate is grounded. This is because G2 has thick gate insulator composed of oxide / nitride / oxide and this property confirms that two gates are electrically independent of each other.

Additionally, floating body effect is observed in the device. Figure 4.2 plots the measured output characteristics. There is kink effect for large V_D because additional holes are generated by impact ionization and accumulated. The kink observed in the measured data verifies that the generated excess holes are accumulated in the floating silicon body. These holes increase body potential temporarily and give the capability of short-term memory.

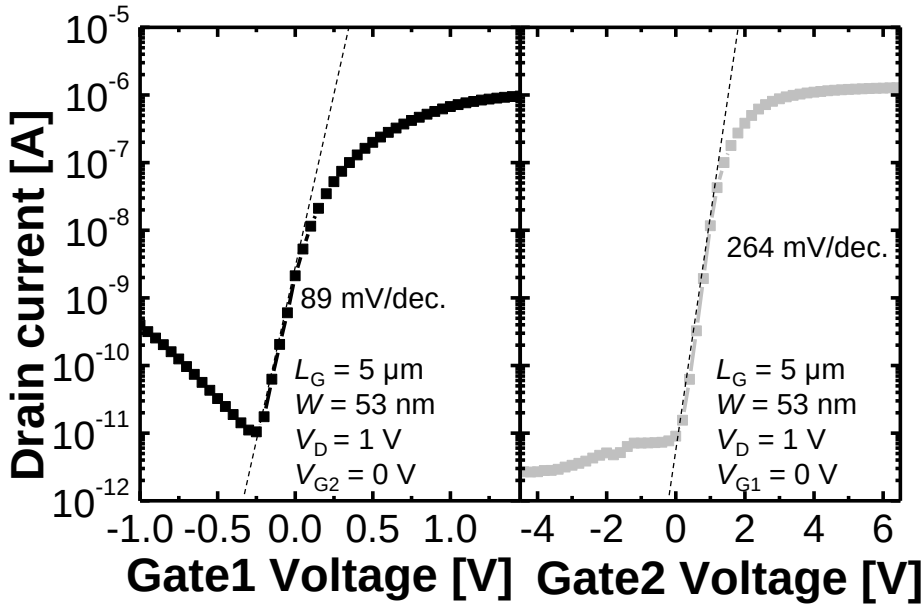


Fig. 4.1. Measured transfer curves of G1 and G2.

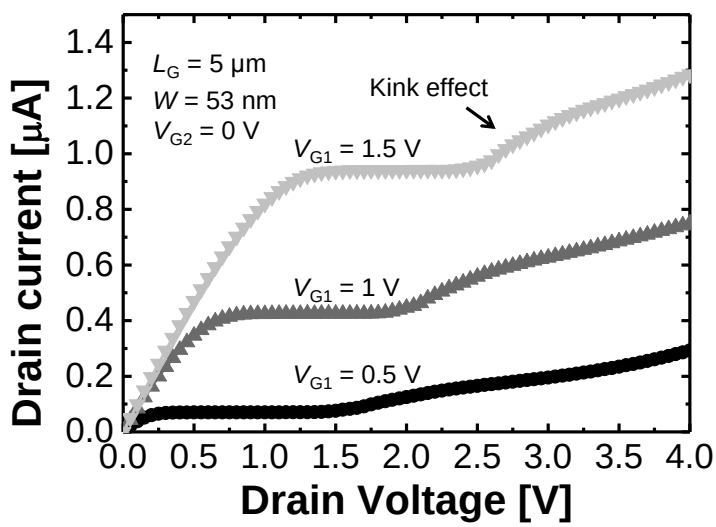


Fig. 4.2. Measured output characteristics depending on V_{G1}. A kink occurs as an evidence of hole accumulation due to floating body effect.

4.2. Synaptic Learning Properties

4.2.1. Transition from Short-Term to Long-Term Memory

To verify synaptic learning properties of the device, the response of the device with input pulse train is investigated. Input pulses with a width of 1 μs and various interval times of 10 μs and 100 μs are applied to the device. Synaptic learning is done with $V_{G1} = V_D = 2$ V and $V_{G2} = -2$ V to generate hot holes and inject them to the charge storage layer.

Figure 4.3 shows the responses of the device to the input pulses with the interval times of 10 μs and 100 μs. When the interval time is 10 μs, the more input pulses are applied to the device, the higher source current flows. On the other hand, the response of the device for each input pulse is almost the same when the interval time is 100 μs. It means that the synaptic weight of the device is increased and it can deliver higher current to the post-synaptic neuron when the input pulses are applied with a higher repetition rate, whereas the synaptic weight is unchanged and the same amount of source current is delivered to the post-synaptic neuron when the input pulses are applied with a lower repetition rate.

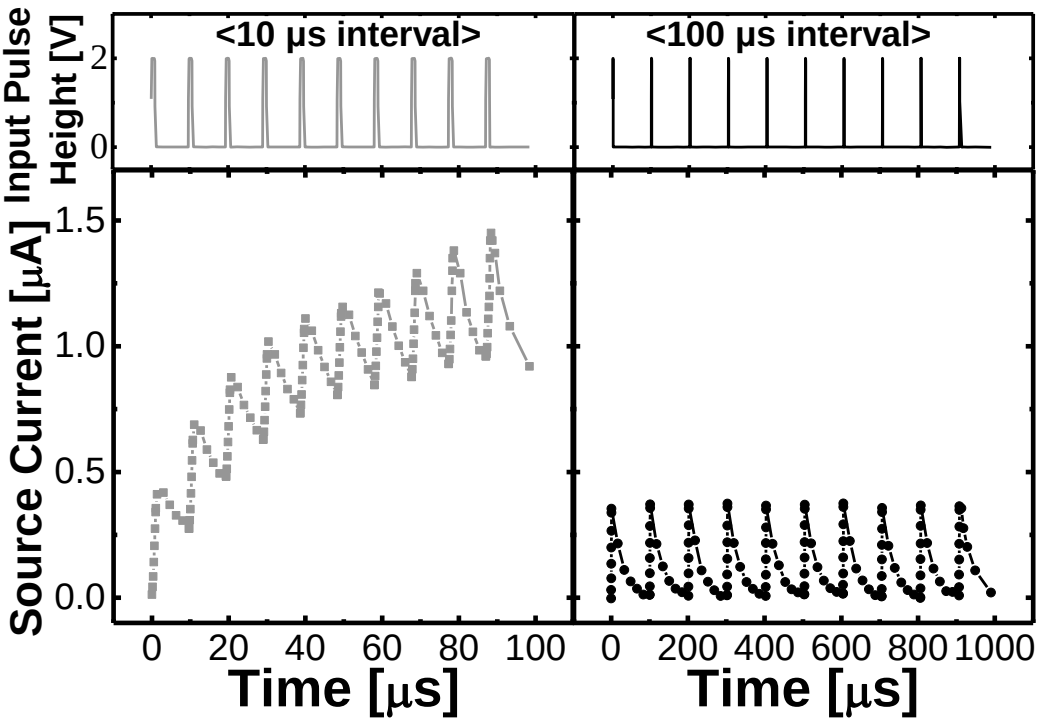


Fig. 4.3. Measured transient responses of source current when the device learning through several times of input pulses with different interval times 10 μs and 100 μs.

Figure 4.4 presents the retention characteristics of the synaptic device right after several input pulses with 10 μs interval were applied to the device. The data reading is done at V_{G1} = V_{G2} = 0 V, V_D = 1 V not to influence the device state. When the number of applied input pulses is less than 7, the drain current temporarily increases and then decreases to the initial level in tens of microseconds, due to the recombination of the excess holes, which is the

forgetting process. Only short-term memory is established unless there are enough number of applied input pulses. Read current is also increased as the number of input pulses increase because many excess holes are accumulated in the floating silicon body.

On the contrary, temporarily-increased read current is decreased to a higher level than the initial level and the increased conductivity remains high for more than 10^2 seconds after 7 or more input pulses are applied to the device. It is because holes are trapped in the nitride layer through hot hole injection. When the body potential is increased enough to generate hot holes near G_2 through positive feedback loop between body potential and accumulated excess holes, newly generated hot holes start to be injected to the nitride layer. Once this happens, the device operates as a non-volatile memory due to the positive charge in the nitride layer. This corresponds to the transition from short-term memory to long-term memory in the biological system.

Furthermore, the synaptic learning discussed above is strongly dependent upon the interval of pulses. Figure 4.4 also shows the transient characteristics of drain current at the same condition except that interval time is 100 μs. With the input with a long interval time, there is no result of the synaptic learning compared with 10 μs interval time. This results from the fact

that the next input pulse is applied to the device while all of generated holes are removed during the long interval time because of recombination. Such characteristic can be considered as time-dependent synaptic learning [21], [22].

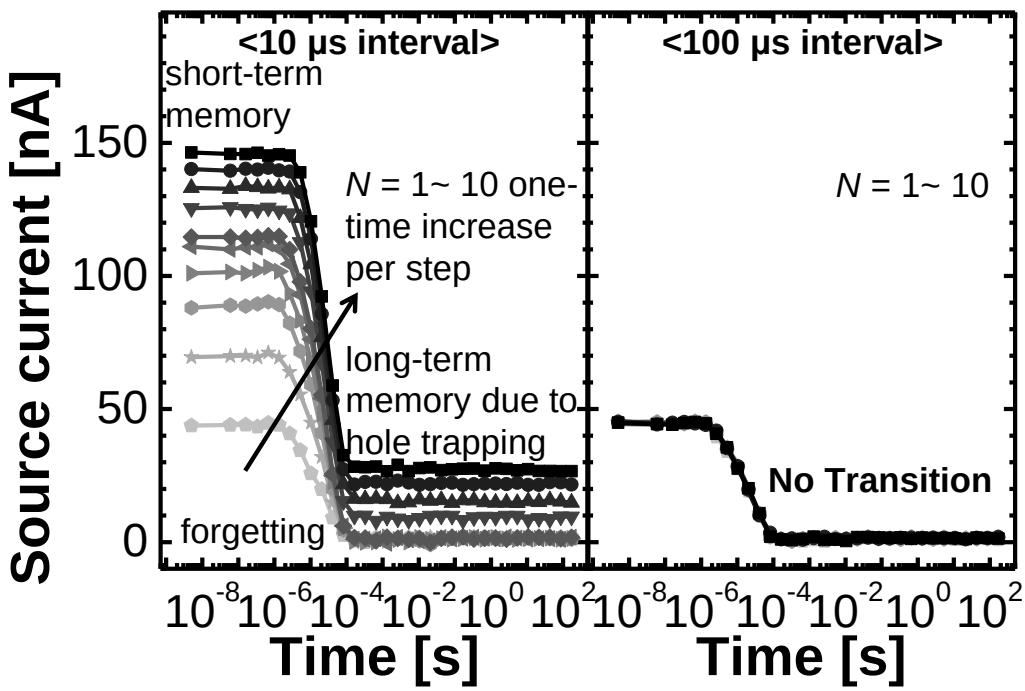
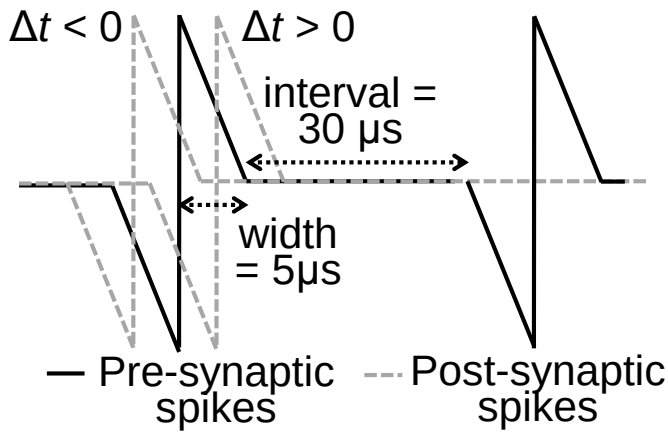


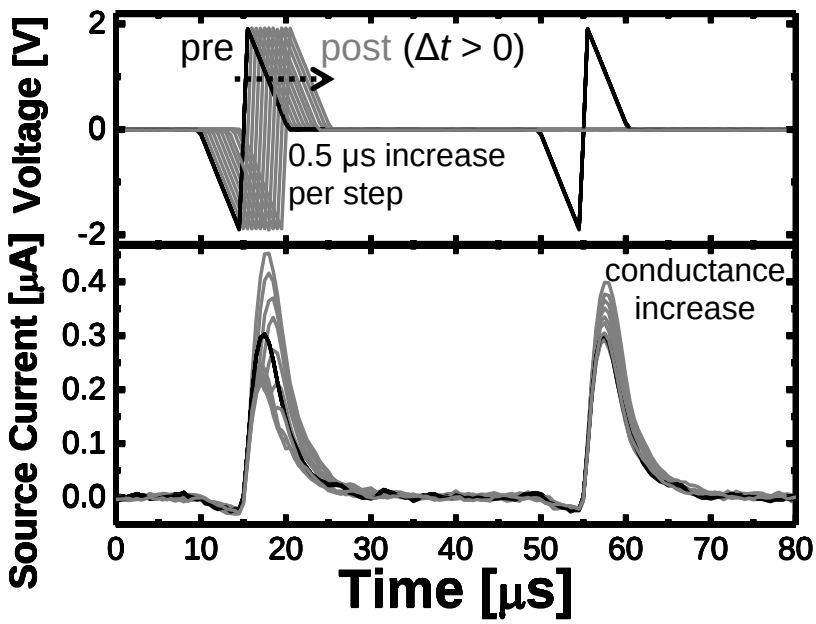
Fig. 4.4 Measured retention characteristics of source current under the read condition (V_D = 1 V) after several times of input pulses with different interval times as 10 μs and 100 μs.

4.2.2. Spike-Timing Dependent Plasticity Characteristics

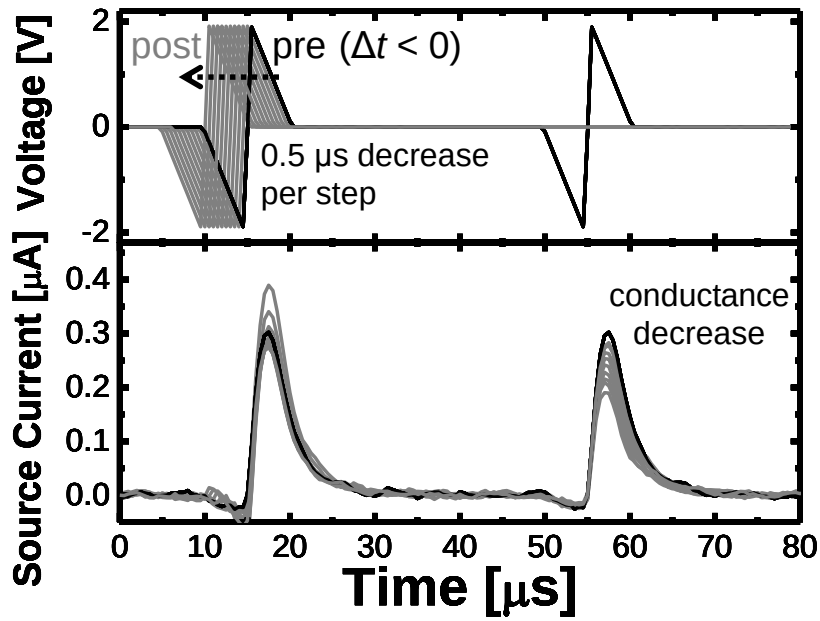
Triangular pre- and post-synaptic spikes are applied to the device to verify its STDP characteristics as shown in Fig. 4.5(a). The pre-synaptic spikes are applied to G1 and drain (D); the post-synaptic ones are applied to G2. Both spikes are applied together with different timing (Δt) for STDP learning process and only a pre-synaptic spike follows to check conductance changes after that; Δt varies from −5 μs to 5 μs. When the device receives the pre-synaptic spikes before the post-synaptic spikes ($\Delta t > 0$), its conductance increases and, the smaller Δt, the larger change in the conductance because of hot hole injection as shown in Fig. 4.5(b). This is because hot holes are generated, injected and trapped in to the nitride layer in a region where both spikes overlap and potential difference between them is larger at smaller Δt. However, the conductance decreases for $\Delta t < 0$ because of hot electron injection as shown in Fig. 4.5(c). The device is depressed in the exactly opposite way of potentiation because of positive V_{G2} at the region where both spikes overlap, leading to the greatest potential difference between them.


Δt < 0
Δt > 0
interval =
30 μs
width
= 5μs
— Pre-synaptic
spikes
Post-synaptic
spikes

(a)


pre
post (Δt > 0)
0.5 μs increase
per step
conductance
increase
Source Current [μA] Voltage [V]
Time [μs]

(b)

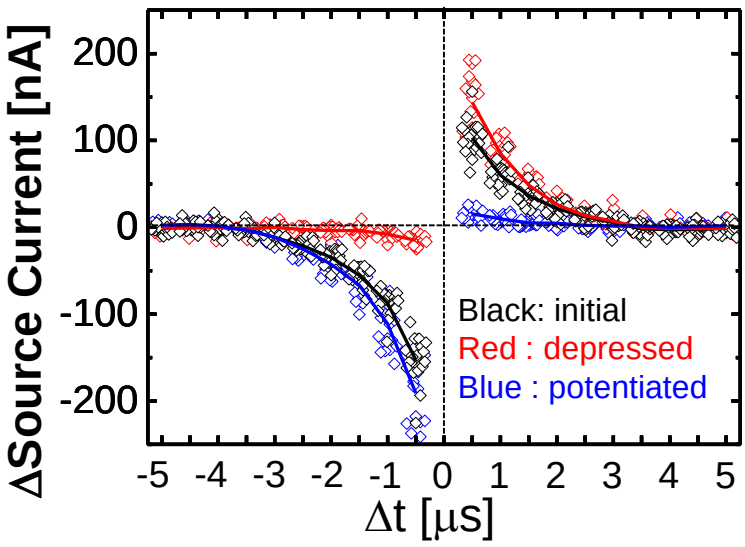


(c)

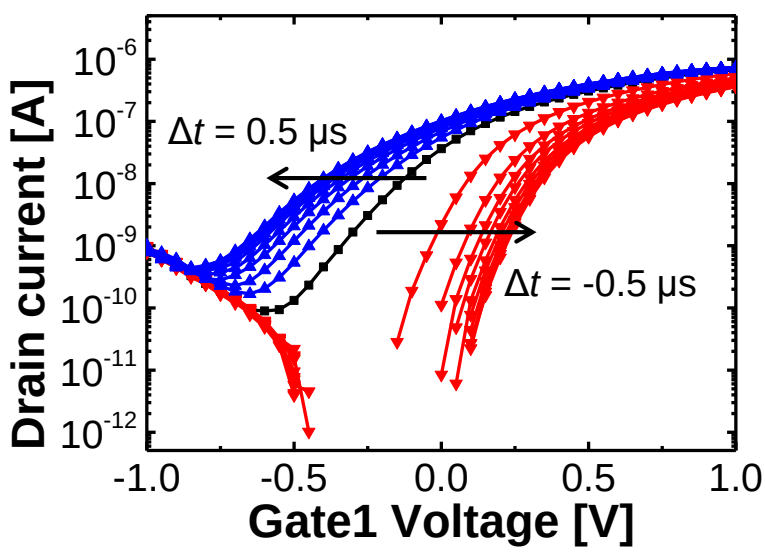
Fig. 4.5. Transient measurement of spike timing-dependent plasticity. (a) Timing diagrams of pre- and post-synaptic spikes with a width of 5 μs and an interval of 30 μs. (b, c) Measured transient responses of source current when the devices learn under STDP rules for positive Δ*t* and negative Δ*t*.

The source current changes after applying spikes with different Δ*t* are plotted depending on the device state in Fig. 4.6(a). The most important feature of the STDP characteristics is that the conductance changes can be obtained by nothing but the spike timing. In addition, the conductance is

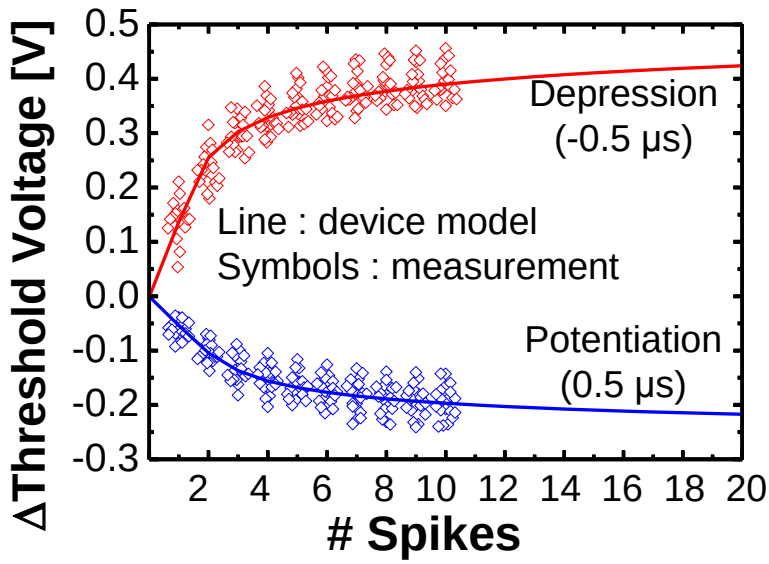
increased a lot more at the depressed state but decreased easily at the potentiated state. This strong dependence of the device's plasticity on its initial state is in line with previous numerous studies on charge trap flash cells having the nitride layer [84]−[86]. The shifted transfer curves as applying pre- and post-synaptic spikes 10 times with keeping Δt as 0.5 and −0.5 μs are plotted in Fig. 4.6(b). The amount of threshold voltage change (ΔV_T) per spike becomes smaller when those spikes are applied to the device repetitively. Note that it corresponds to the fact that the learning rate of the artificial neural network (ANN) system becomes smaller with the consecutive spikes having the same sign of Δt. ΔV_T of 20 samples at Δt = 0.5 and −0.5 μs are measured with the number of applied spikes and summarized in Fig. 4.6(c). V_T is changed in a positive direction (depression) about 2 times more than in a negative direction (potentiation) because of effective mass difference between holes and electrons, and about 0.6 V of ΔV_T is obtained after 10 spikes on average. In addition, the STDP characteristics have been modeled based on the measured data (symbols) and the hot carrier gate current model reported in [87]. The injection rates of hot carriers drop dramatically because the trapped carriers reduce the electric field in the G2 dielectric exponentially, and the device model reflects that tendency well (line).


ΔSource Current [nA]
200
100
0
-100
-200
-5 -4 -3 -2 -1 0 1 2 3 4 5
Δt [μs]
Black: initial
Red : depressed
Blue : potentiated

(a)


Drain current [A]
10-6
10-7
10-8
10-9
10-10
10-11
10-12
Δt = 0.5 μs
Δt = -0.5 μs
-1.0 -0.5 0.0 0.5 1.0
Gate1 Voltage [V]

(b)



(c)

Fig. 4.6. Experimental results of spike timing-dependent plasticity. (a) Statistical STDP characteristics with 20 samples. (b) Shifted transfer curves of G1 (I_D-V_{G1}). (c) Potentiation and depression characteristics (ΔV_T) with the number of applied spikes when Δt = 0.5 μs and −0.5 μs, respectively.

Chapter 5. System Level Simulation

5.1. Hardware-Based Spiking Neural Network

To demonstrate the functionality of the synaptic transistors discussed above for pattern classification through supervised learning, the single-layer SNN system composed of them is designed and simulated as shown in Fig. 5.1. A total of 784 input nodes represent 28×28 black-and-white pixels of the modified national institute of standards and technology (MNIST) dataset and 10 output nodes represent 10 digits; here, all the input and output nodes are fully connected through the synaptic transistors. Transferred signals from the input nodes are integrated through excitatory synaptic transistors at a node of capacitor so the post-synaptic neuron circuit can generate output spikes at a higher rate when those excitatory ones have been potentiated and the integrated voltage at the node exceeds the threshold voltage at a higher rate because of it.

Figure 5.2(a) explains the way this spiking neural network (SNN) system learns the MNIST digit samples under supervised learning. The pre-synaptic spikes are applied to corresponding synaptic transistors with different

timing depending on their colors: black with $\Delta t = -0.5$ μs and white with $\Delta t = 0.5$ μs compared to teaching signal which is given to the output node matching to the digit of training sample. Therefore, the V_T is increased for the synaptic transistors representing black pixels and decreased for white pixels. This is a kind of temporal coding in artificial neural network where the intensity of signals can be expressed as their firing timing [88]−[90]. The weight maps of all the synaptic transistors in the form of ΔV_T after training of 10,000 samples are illustrated in Fig. 5.2(b). The value of ΔV_T is the most negative (white) at the most duplicated area with digit samples, whereas the synaptic transistors representing the background area have the most positive ΔV_T (black) because input nodes of the background area received all the spikes in the condition of $\Delta t < 0$.

The way this trained SNN system classifies untrained test samples is illustrated in Fig. 5.3(a). Testing input signals are applied to the input nodes corresponding to white pixels, and the output node that fires first is regarded as the classification result. The reason why classification accuracy is calculated in this manner is that the weight sum of transferred currents (I_E) to the output node, which is the most congruous to the testing sample, is expected to be the largest owing to the potentiated synaptic transistors in the shape of the digit, leading to high current flows. The recognition rate with

1,000 untrained testing samples is shown in Fig. 5.3(b). After very quick increase to 37%, the accuracy increases gradually up to 60% as the number of training samples increases; however, it is saturated over 3,000 trained samples. This is because the output nodes having more white pixels in their weight maps, such as eight or zero, have a higher probability to fire, even though they do not match the digits of test samples, leading to a low recognition rate of the ones (such as 8) that have less white pixels (such as 1).

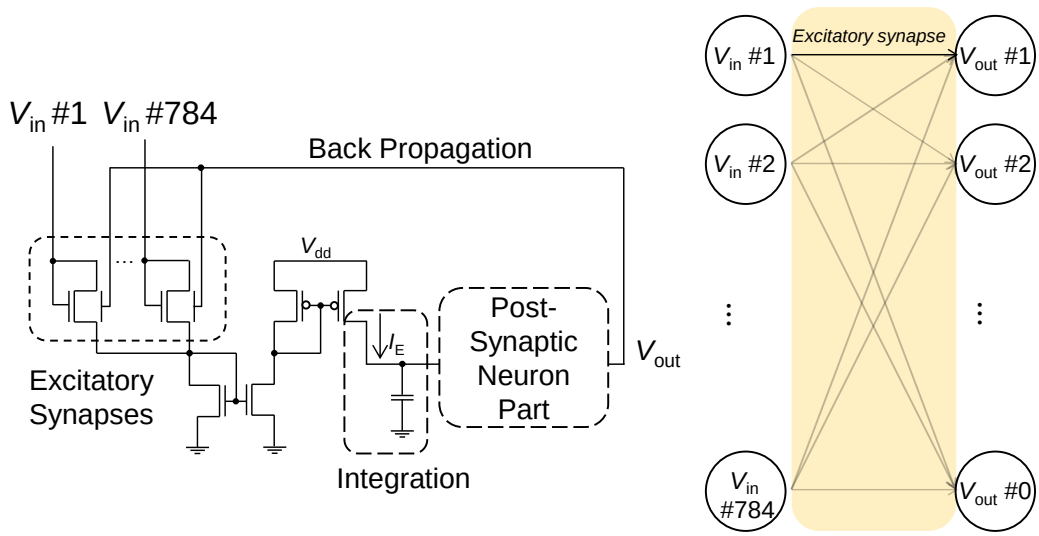
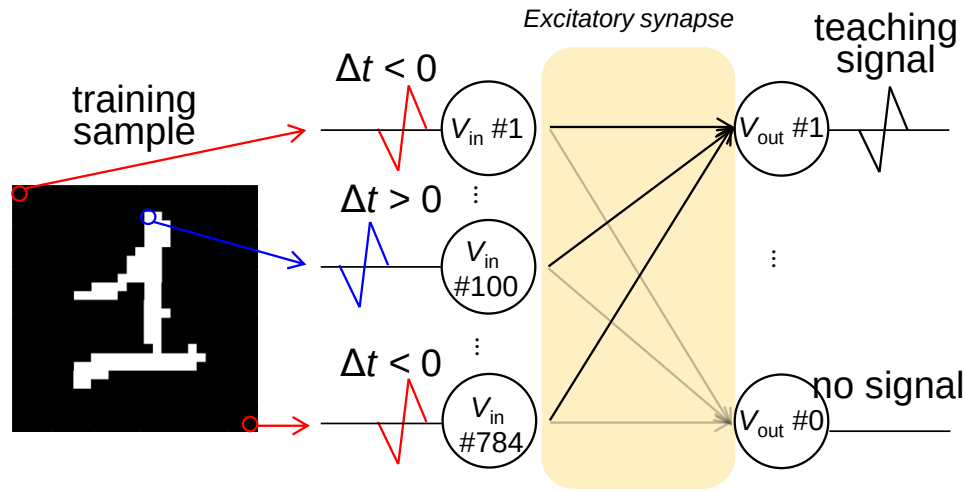
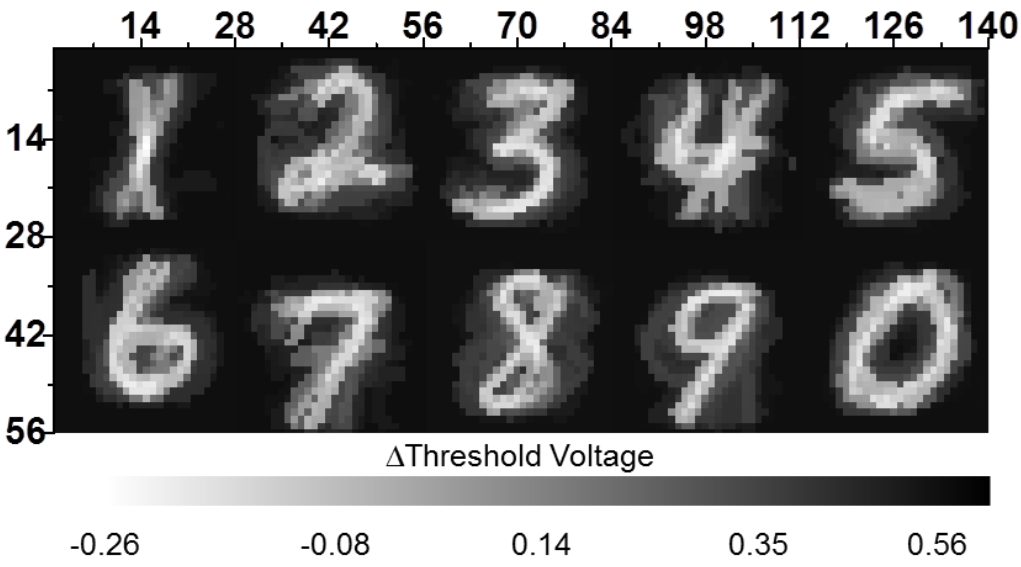


Fig. 5.1. Schematic illustration of the single-layer neural network composed of excitatory synaptic transistors.

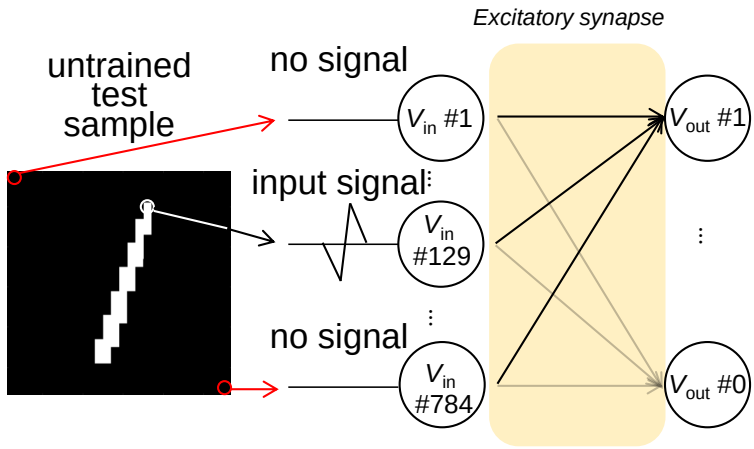


(a)

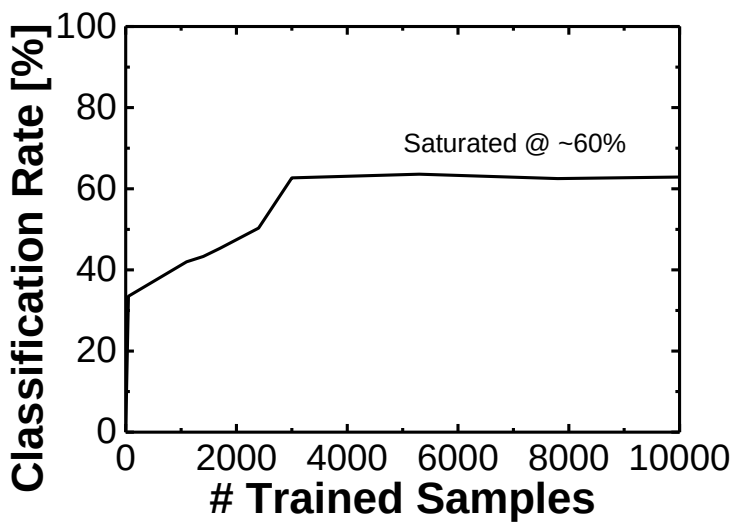


(b)

Fig. 5.2. Learning process. (a) The illustration of how to train samples under supervised learning using temporal coding method. (b) The weight map (ΔV_T) of 784×10 excitatory synapses after training of 10,000 samples.



(a)



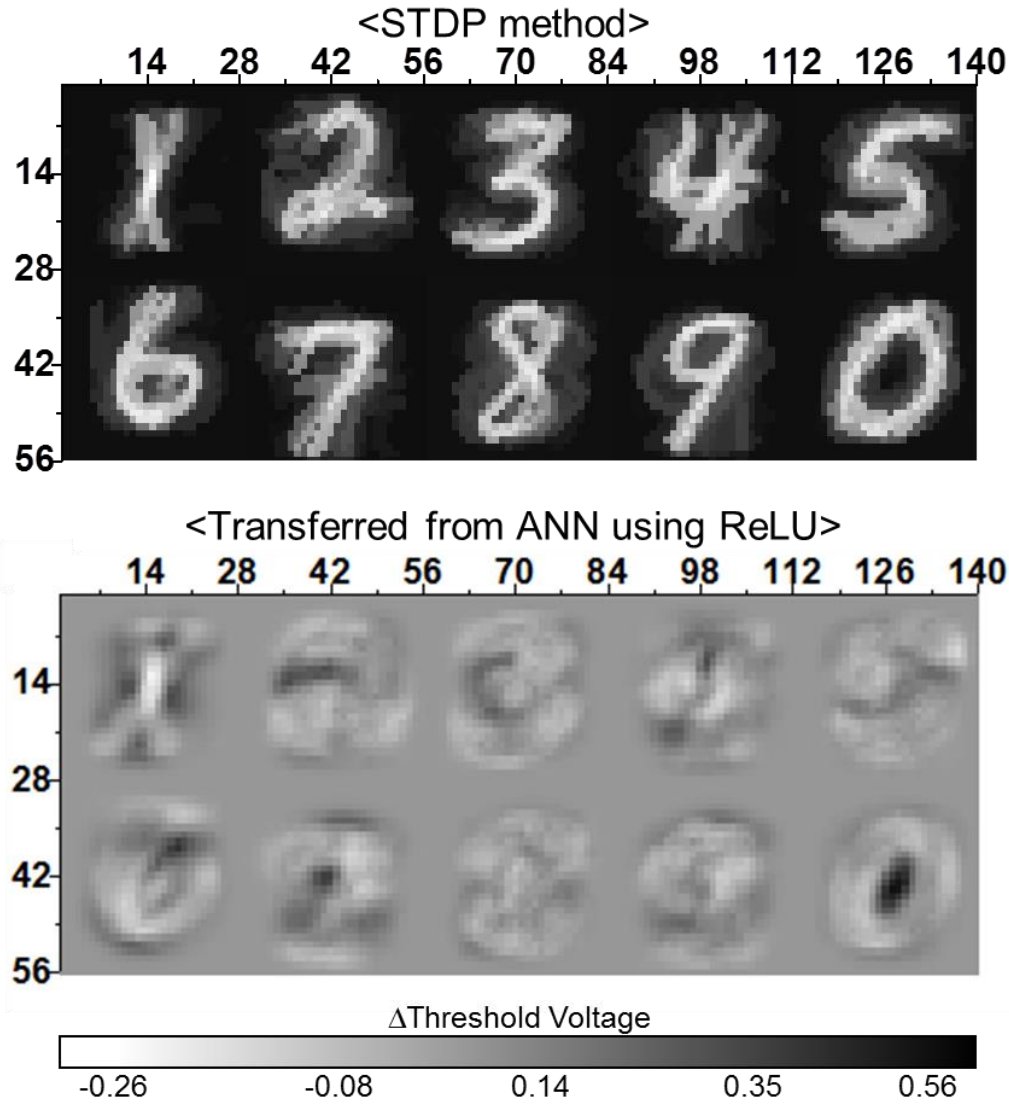
(b)

Fig. 5.3. Verification of classification functionality. (a) The illustration of how to verify whether SNN system classifies test samples correctly or not. (b) Classification rate using 1,000 test samples as a function of the number of training sample.

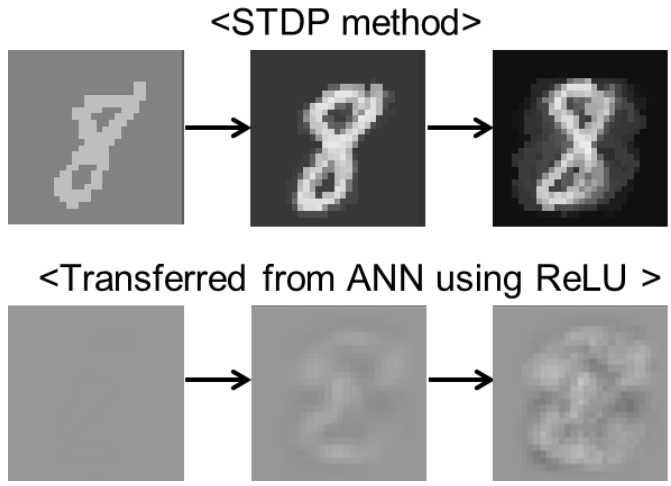
5.2. Transferred Synaptic Weights from ANN Using ReLU

In artificial neural networks (ANNs), a rectified linear unit (ReLU) is one of the most popular activation functions because of the lack of vanishing gradients problems compared to other ones such as sigmoid or hyperbolic tangent [91]–[94]. To analyze the reason of the classification accuracy difference between the hardware-based SNN and ANN, the weight values of ANN, which has the same network structure (784×10), using ReLU as an activation function are transferred to ΔV_T values of SNN and the pattern recognition rates for those two cases are evaluated. Figure 5.4(a) compares two weight maps which are learned through STDP rule and transferred from ANN in the same scale, respectively; here, the synaptic weights of ANN are converted to the ones of SNN to be proportional to their square roots so the transferred I_E can be in line with the weight sum of ANN with ReLU. The former one looks like imprinting digits into the synaptic devices, whereas the latter one is well characterized by their unique features of each digit. That is why the hardware-based SNN has a poor accuracy of 60%: the weight map of it does not reflect the unique characteristics of each digit. The transformation processes of the weight maps for digit 8 as the training progressed are illustrated in Fig. 5.4(b) for the two cases. The imprinted pattern on the weight

map by STDP method becomes clearer in the direction that it can fire frequently by digit input samples; however, the weight map transferred from ANN refines its unique feature as learning proceeds so that the weight map results in higher classification accuracy. Figure 5.5 plots the classification rates for each digit depending on the methods. The poor accuracies, especially digit one and digit nine, have been highly improved by adopting the transferred synaptic weights, leading to 87.6% of the total accuracy. In addition, the most noteworthy thing is that the classification rates of the transferring method and ANN itself are almost the same for every single digit. It is believed that SNN using the transferred weight map and ANN with ReLU are equivalent in their operations in respect that the intensity of output nodes can correspond to the firing rate [95].


<STDP method>
14 28 42 56 70 84 98 112 126 140
14
28
42
56
<Transferred from ANN using ReLU>
14 28 42 56 70 84 98 112 126 140
14
28
42
56
ΔThreshold Voltage
-0.26 -0.08 0.14 0.35 0.56

(a)



(b)

Fig. 5.4. Transferred synaptic weights from artificial neural network. (a) Comparison of the weight maps learned by STDP method and transferred from ANN, respectively. (b) Training progress of each weight map.

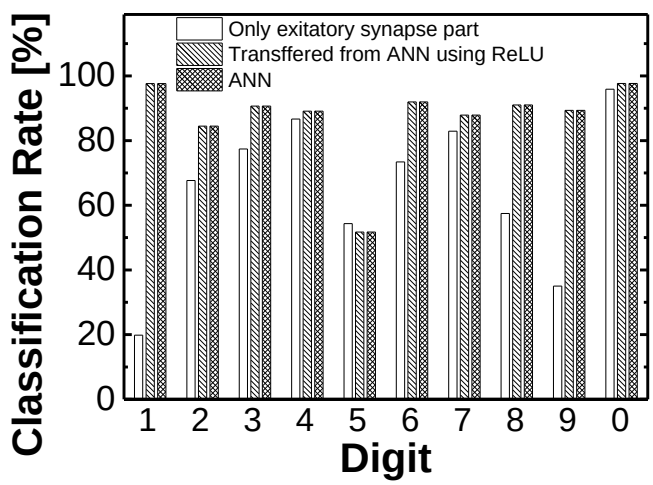
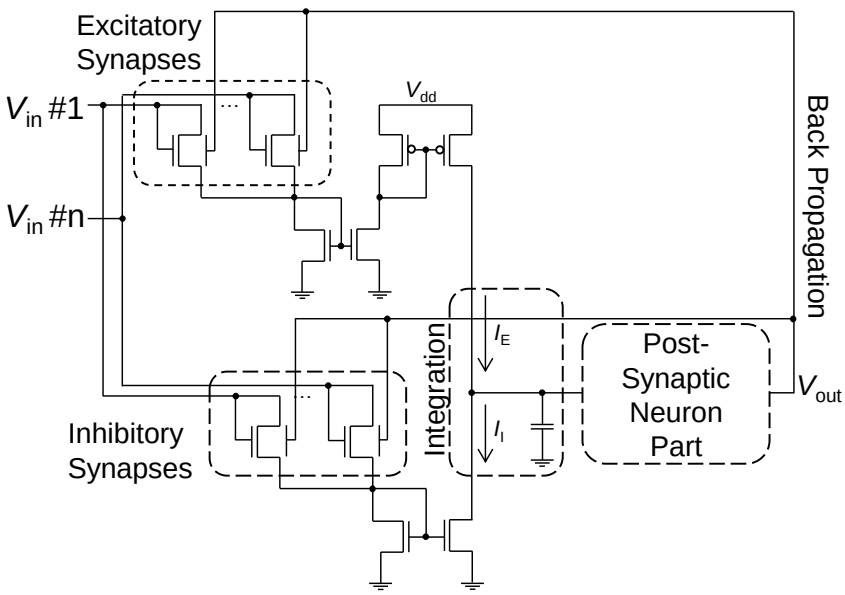


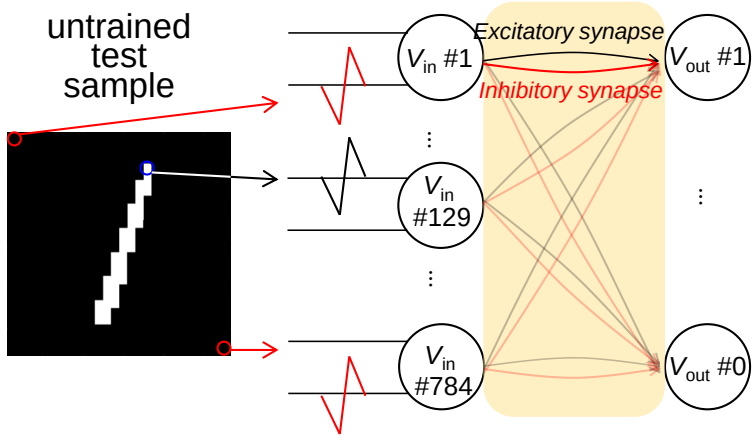
Fig. 5.5. Comparison of the classification accuracy of each digit for STDP method, transferred synaptic weights method, and ANN, respectively.

5.3. Addition of Inhibitory Synapse Part

In order to reduce the classification error caused by the overlapping pattern issue discussed above, the inhibitory synaptic devices with the same weight maps as the excitatory ones are added to the circuit as shown in Fig. 5.6(a), so that the integrated charges, the weight sum, at the node by transferred signals from the excitatory synapse part are taken out through inhibitory synapse part. How to recognize test samples through both the excitatory and inhibitory parts is explained in Fig. 5.6(b). Like the previous method, the input signals are applied to the excitatory synapses corresponding to their own pixels in the case of the white pixels; at the same time, the input signals are applied to the inhibitory synapses in the case of the black pixels. This change in the manner of classification leads to the result that if the testing samples which cover not only their own digits but also other digits, the remaining parts contribute to subtract the weight sum by the current flows (I_I) through the inhibitory synaptic transistors.



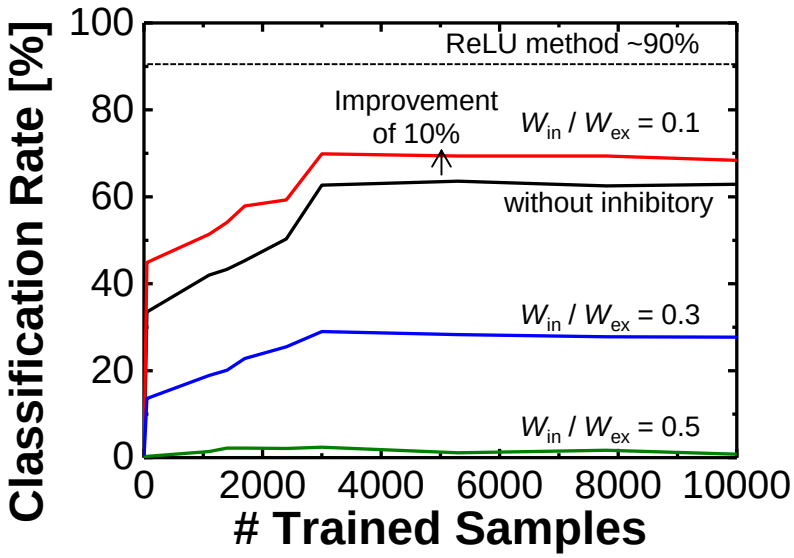
(a)



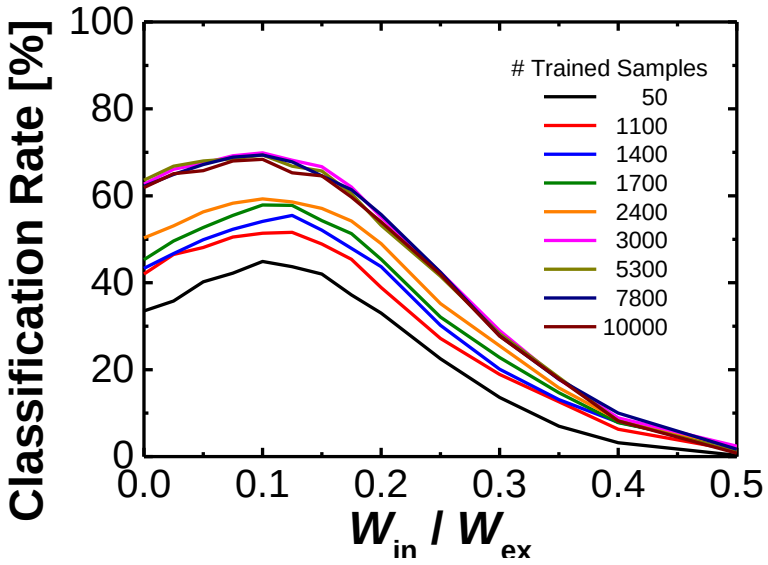
(b)

Fig. 5.6. Addition of inhibitory synaptic transistors. (a) Schematic circuit diagram of SNN with both excitatory and inhibitory synapse parts. (b) The illustration of the modified way SNN system classifies test samples using both parts.

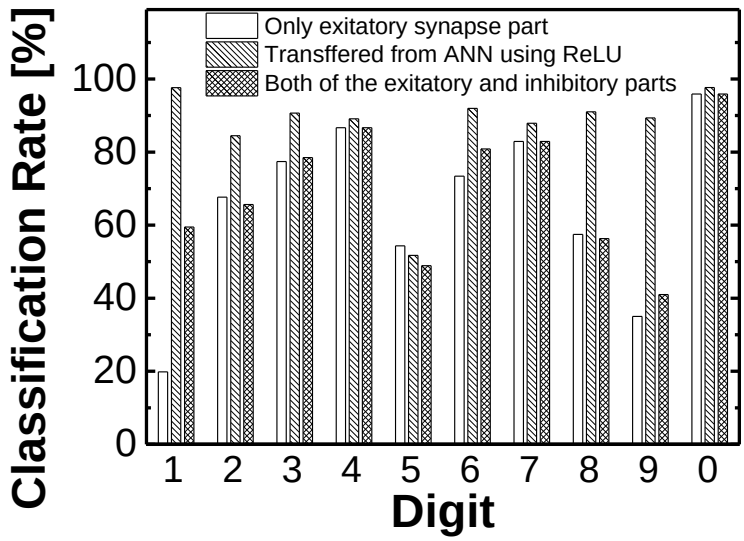
Figure 5.7(a) shows the accuracies as a function of the number of training samples for various ratios between the channel widths of the excitatory synaptic transistors (W_{ex}) and the inhibitory ones (W_{in}). The accuracy is improved up to 70% at W_{in}/W_{ex} = 0.1; however, it starts decreasing after that and reaches bottom (nearly 0%), when W_{in}/W_{ex} = 0.5. This result comes from the fact that the output nodes cannot fire with any testing sample if W_{in} is too wide compared to W_{ex} because the number of black pixels is larger than that of white pixels and I_I is higher than I_E in most cases. Therefore, it is very important to modulate W_{in}/W_{ex} to obtain the highest accuracy value. The accuracies depending on W_{in}/W_{ex} and the number of training samples are plotted in Fig. 5.7(b). In all cases, the accuracy has the highest value near W_{in}/W_{ex} = 0.1 (about 10% more compared to with the case having only the excitatory synapse part), and decrease to 0% for W_{in}/W_{ex} > 0.5. Figure 5.7(c) compares the classification rate of each digit for these two SNN systems. It is noteworthy that the accuracies of the digits which have a small number of white pixels like one (1) is significantly enhanced from 19% to 60%, while those of other digits keep the similar values. It is confirmed that the addition of the inhibitory synapse part can effectively solve the misclassified cases because of the overlapping pattern problems.


Classification Rate [%]
100
80
60
40
20
0
ReLU method ~90%
Improvement
of 10%
W_{in} / W_{ex} = 0.1
without inhibitory
W_{in} / W_{ex} = 0.3
W_{in} / W_{ex} = 0.5
0
2000
4000
6000
8000
10000
Trained Samples

(a)


Classification Rate [%]
100
80
60
40
20
0
Trained Samples
50
1100
1400
1700
2400
3000
5300
7800
10000
0.0
0.1
0.2
0.3
0.4
0.5
W_{in} / W_{ex}

(b)



(c)

Fig. 5.7. Improved classification accuracy with the inhibitory synapse part. (a) Classification rates after adding the inhibitory synapse part. (b) Dependence of classification rates on W_{in}/W_{ex}. (c) Comparison of the classification accuracy of each digit for three SNN systems.

Chapter 6. Conclusion

6.1. Review of Overall Work

In this dissertation, it is revealed that the asymmetric dual-gate structure can make it possible to connect the synaptic transistors to the neuron circuit without additional switches, the main cause of extra overheads of a hardware-based neuromorphic system. It allows the system composed of the synaptic transistors to transfer and receive signals at the same time.

Firstly, its mechanism of synaptic learning is studied with a TCAD device simulator. The transition from short-term to long-term memory occurs through positive feedback loop between body potential and accumulated excess holes, leading to hot carrier injection into the nitride layer. STDP characteristics are also obtained using triangular-shaped spikes.

The synaptic transistors are fabricated through the key process steps. A steep etch slope, almost close to perpendicular, is achieved using TDMR-AR87 as a PR. The fin channel is etched by using the MTO sidewall spacer as a hard mask, and the two gates are separated through the two-step CMP processes.

The electrical and synaptic characteristics of the fabricated devices are measured. It is confirmed that the two gates are electrically independent by checking the transfer curves for each gate, and the synaptic learning rules including the transition from short-term memory to long-term memory and STDP characteristics are measured. These characteristics are strongly dependent on the frequency of the spikes like a biological synapse. The measured data are analyzed statistically and modeled based on the hot carrier gate current model.

Finally, SNN systems are constructed using the device model and the functionality of pattern recognition with the MNIST data set is verified in three ways. SNN with only the excitatory synapse part learns the MNIST patterns under STDP rule using temporal coding method; however, it has poor classification ability with 60% of the total accuracy. It is dramatically improved to 87.6% in the case of SNN with the transferred synaptic weights from ANN using ReLU. The difference between those two systems is whether the region representing the unique features of each digit is potentiated well by the synaptic weights. The addition of the inhibitory synapse part improves the classification accuracy to 70% by reducing the overlapping pattern problem which comes from the fact that the output nodes having more white pixels tend to fire to unmatched training samples.

These results lead us to conclude that the synaptic transistor studied in this dissertation is a very strong candidate for a synaptic device in neuromorphic systems thanks to its direct connectability with neuron circuits and synaptic learning properties. In addition, the SNN systems and learning methods provide a framework for future studies about hardware-based neuromorphic systems.

6.2. Future Work

Related to this study, however, there are still remaining issues that need to be studied further. First, the array architecture of the synaptic transistors should be studied. In the present structure, there is a disadvantage in terms of memory density and array configuration. To be scaled down to under 10 nm regime, it is required to investigate the array structure with a vertical channel structure (source at the bottom and drain on the top) with a thin and long body.

Secondly, how to connect the synaptic transistors to neuron circuits should be investigated in terms of integration. Since the fabrication flow described here does not follow conventional CMOS process steps in sequence, other techniques including through-silicon via (TSV), which is supposed to pass completely through a synapse wafer and a neuron circuit wafer, should be considered.

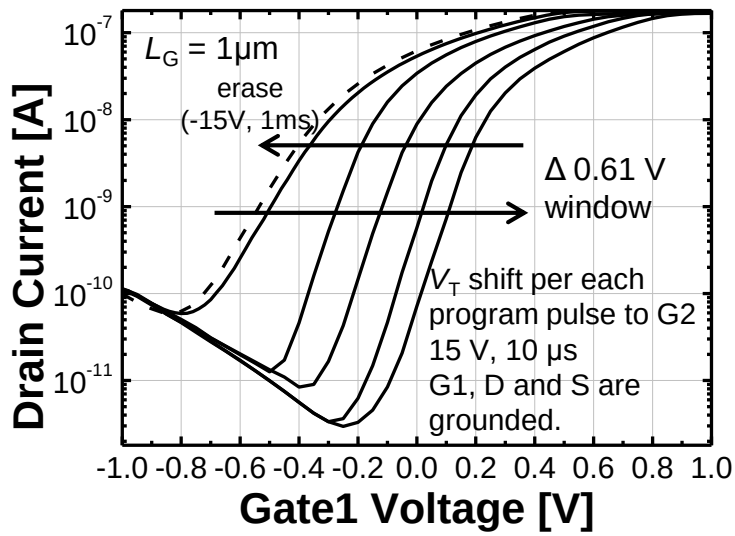
Finally, it is recommended that future work should focus on the effective in-situ learning method to obtain distinguishing synaptic weights for each digit. There is a lot to be improved compared to ANN. Manipulating the transferred synaptic weights from ANN can be an alternate method; however, it is the sort of the ex-situ method because the hardware-based system does not learn by itself. If a hardware-based neural network product having the in-situ learning method is realized, users can directly have the product learn and customize its functions. In order to do that, how to learn inhibitory synapse part should be investigated more. In a biological nervous system, inhibitory synapses are believed to have different learning rules compared to excitatory synapses; however, the synaptic weights of the inhibitory synaptic transistors studied here are updated under the same rule. This might be a decisive factor to achieve the in-situ learning method. Therefore, learning with the hardware-based system itself should be studied more.

Appendix A. Multi-Threshold Voltages in Ultra Thin Body Devices by Asymmetric Dual-Gate Structure

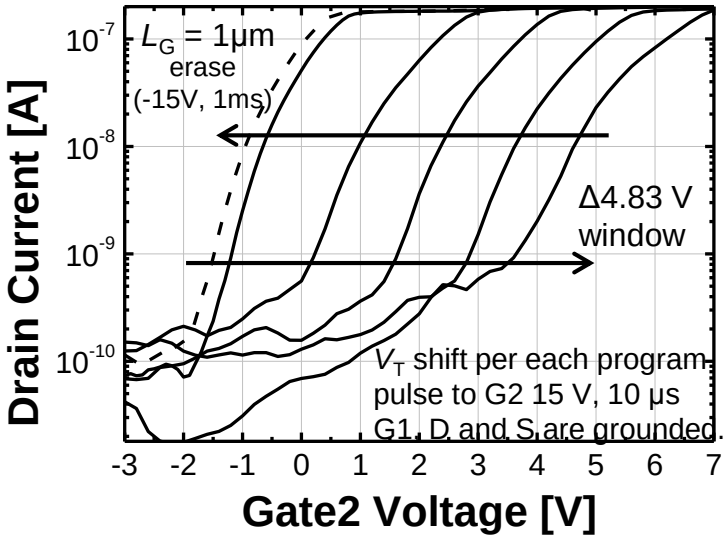
Ultra-thin body (UTB) devices are promising candidates in deep sub-100 nm regime because they provide good controllability of short-channel effect, suppression of random dopant fluctuation and high current drivability [96]–[104]. In addition, multi-threshold voltages (V_T) designs have been used for low-power CMOS applications [105]–[107]. Transistors with high-V_T are used in indecisive regions for low standby leakage current, while transistors with low-V_T are used in decisive regions for high operating current. However, it is hard to realize multi-V_T and dynamic V_T modulation with an undoped channel in UTB structure, so there are several techniques to control V_T such as ground plane, back-biasing and multiple gate materials [108]–[121]. And they requires BOX structure as a necessity because significantly high doping concentration for ground plane or back-gate bias is needed with thick BOX. In this Appendix A, we propose a new method to control V_T in UTB devices using asymmetric dual-gate structure. It is noteworthy that this method can

be used without ultra-thin BOX structure. The proposed method of V_T control is discussed. It is possible to modulate V_T by trapping charges in G2 stack.

The V_T modulation operation of the fabricated device is composed of several program pulse steps. Four programming pulses of 15 V and 10 μs are applied to G2 for trapping charges in the nitride layer by Fowler-Nordheim tunneling. Figure A.1 shows measured transfer characteristics of both G1 and G2 as a parameter of programing pulses. V_T for the G1 (V_{T1}) and the G2 (V_{T2}) are defined as V_{G1} and V_{G2} at drain current = 10^{-9} A, respectively. It is found that V_{T1} and V_{T2} are changed by the number of trapped charges in the nitride layer. Both of them are increased as the programming pulses are applied like a flash memory cell [122]−[124]. The programming pulses of 15 V with a width of 10 μs are applied to G2 and other electrodes grounded. The erase operation is performed by applying a single pulse of −15 V for 1 ms to G2. About 0.61 V of V_{T1} window is achieved by five programming pulses as shown in Fig. A.1(a), whereas about 4.83 V of V_{T2} window is achieved for same programming conditions as shown in Fig. A.1(b). V_T can be controlled by modulating conditions of programming pulses. Unlike a back-biasing scheme, this method does not require additional biasing for keeping V_T adjusted value [108]−[112].



(a)



(b)

Fig. A.1. Measured transfer curves of the device with V_D = 1 V as a function of programming pulses. (a) I_D-V_{G1}. (b) I_D-V_{G2}.

Figure A.2(a) shows the V_{T1} and V_{T2} change per each programming pulse. Both of them increase as the number of programming pulses increases. This is because the number of trapped electron in the nitride layer is increased. The amount of V_T change per each pulse is decreased because constant voltage pulses are used [125]. Figure A.2(b) shows the relationship between the subthreshold swing (SS) and V_{T1} controlled after programming pulses biased. It is found that the lower SS is obtained the higher V_{T1} is. It indicates that more trapped electrons in G2 stack effectively suppress the channel formation on G2 side.

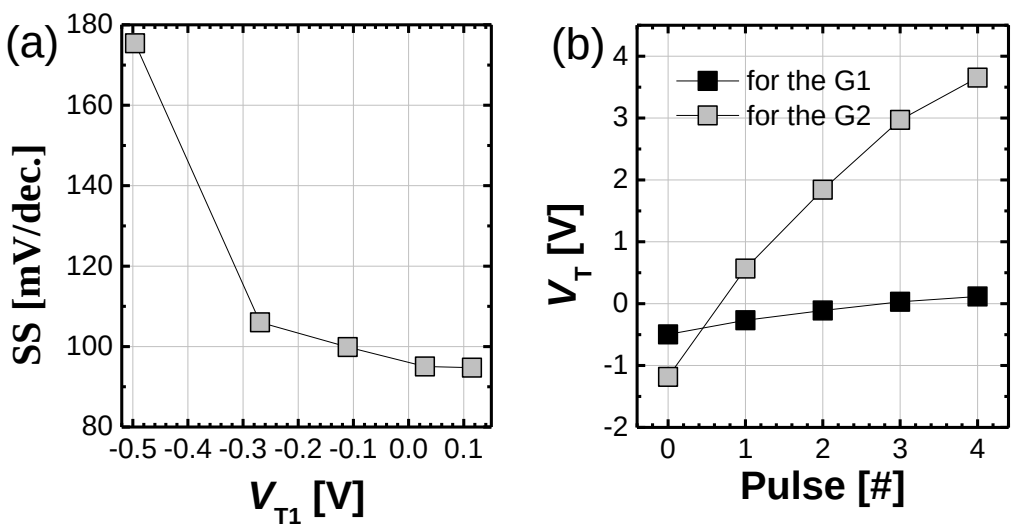


Fig. A.2. (a) V_T change for both gates per each programming pulse and (b) relationship between SS and V_{T1} controlled by the trapped charge in the G2.

Figure A.3 demonstrates the retention characteristics of both V_{T1} and V_{T2}. The extracted V_T windows at 85 ℃ up to 10 years are 0.46 V and 3.63 V for G1 and G2, respectively, which are 65% of the initial window. It means that accelerated charge loss at an elevated temperature of 85 ℃ is predicted to be 35% after 10 years by the extrapolation. This result corresponds to previous works in the point that the extracted charge losses at 85 ℃ after 10 years were 66% and 25% with 2 nm and 4.2 nm tunnel oxide, respectively [85], [124].

Coupling ratio between ΔV_{T1} and ΔV_{T2} per each pulse is plotted in the Fig. A.4(a). It remains almost constant value about 0.126 for every programming pulse. In order to understand how the amount of V_T window for G1 is determined and coupling ratio keeps nearly same value, capacitance network of the device in the top view is analyzed. Considering simple capacitance network model as shown in the inset of Fig. A.4(a), coupling ratio between V_T windows for two gates is given by $(C_{nitride}||C_{gox2}||C_{body})/C_{gox}$. It represents the effect of trapped charges on the channel potential and V_T for G1. The coupling ratio for the device is calculated as 0.131 which is almost the same as measured data 0.61 V/4.83 V. Figure A.4(b) shows coupling ratio as a function of body thickness. Higher coupling ratio is achieved when body thickness is thinner. To put it another way, this method is much more efficient in UTB structure with thinner body thickness.

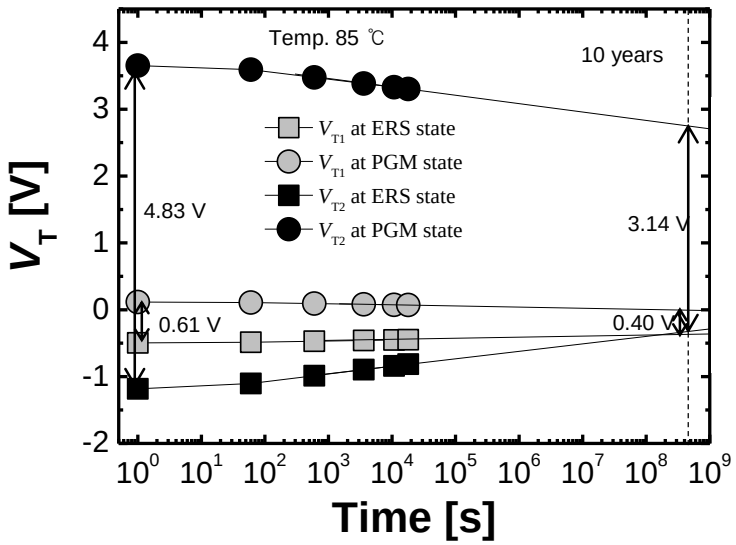


Fig. A.3. Retention characteristics of V_{T1} and V_{T2} at 85 ℃.

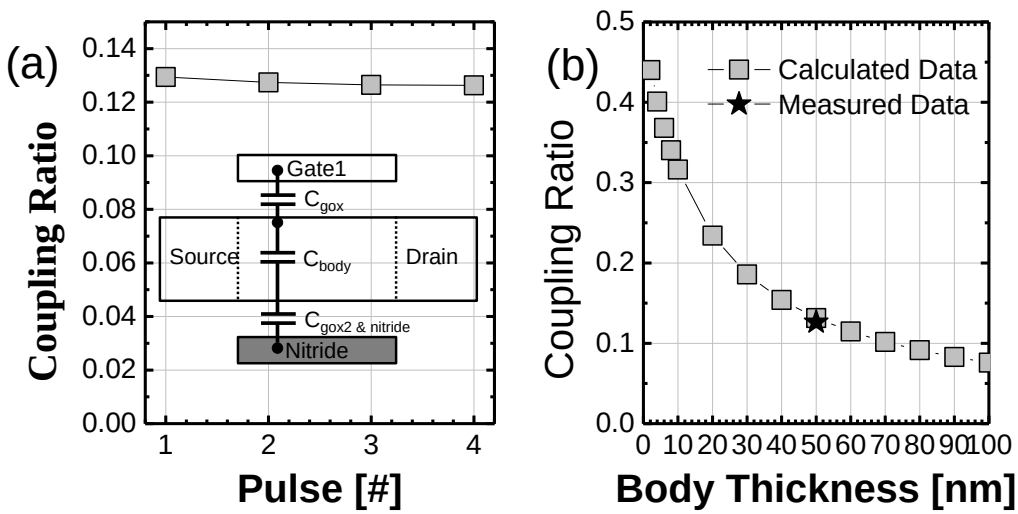
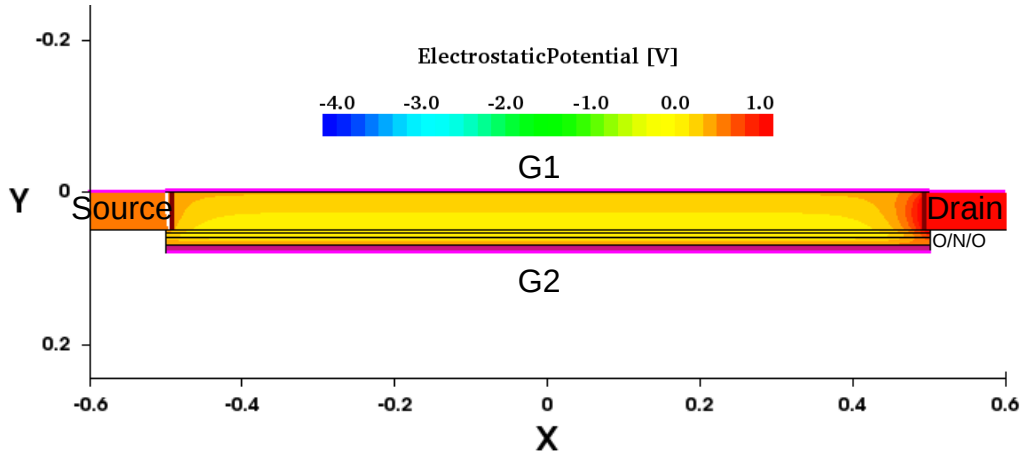


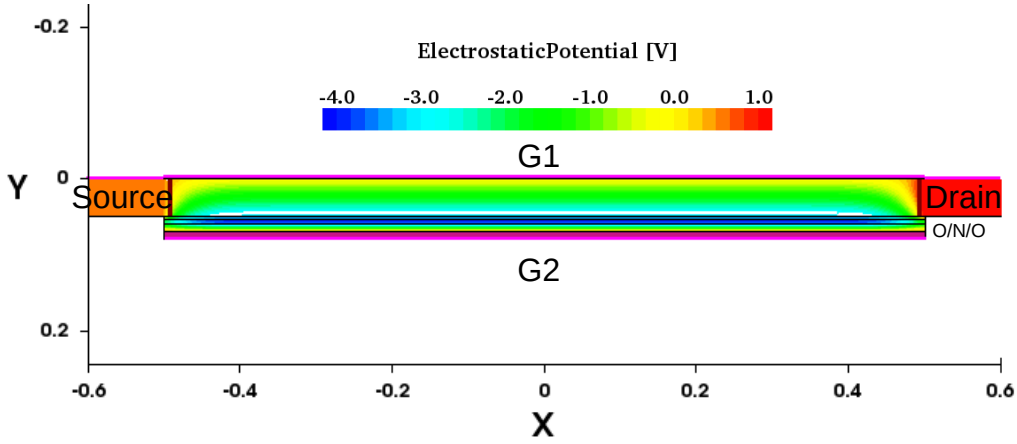
Fig. A.4. (a) Coupling ratio between ΔV_{T1} and ΔV_{T2} per each programming pulse. The inset shows a simple capacitance network model in the device. (b) Body thickness dependence on coupling ratio between ΔV_{T1} and ΔV_{T2}.

The effect of charges in the nitride layer on channel potential is investigated using simulated electrostatic potential contours for same structural conditions with the fabricated device as shown in Fig. A.5. The simulation is conducted using Synopsys Sentaurus Device TCAD tool (version J-2014.09) [126]. Understandably, channel potential is far more lessened when the nitride layer is negatively charged compared with when it is neutral. Consequently, this brings out increase of V_T for G1 because the charge of trapped electrons suppresses the formation of inversion channel on the G1 side.

In conclusion, a new method of V_T modulation in UTB devices was proposed. It has multi-V_T by trapping charges in G2 stack which is independent of G1 due to structural feature of asymmetric dual-gate. This method is more efficient at thinner body thickness by capacitance network. It is noteworthy that it can be used without ultra-thin BOX and additional biasing scheme and V_T can be modulated even after finishing fabrication process.



(a)



(b)

Fig. A.5. Simulated contours of electrostatic potential with V_G = 0 V and V_D = 1 V. (a) When nitride layer is neutral. (b) When nitride layer is negatively charged to increase V_T as much as 0.61 V.

Appendix B. Asymmetric Dual-Gate-Structured 1-T DRAM Cell for Retention Characteristics Improvement

Dynamic random access memory (DRAM) devices with high scalability and performance are essential to fulfill the needs of high memory capacity and high speed. As the conventional one-transistor (1T) one-capacitor DRAM cells are scaled down, reducing capacitor volume has become one of the major challenges [127]–[131]. According to International Technology Roadmap for Semiconductors 2013, the half-pitch size of DRAM cells will be less than 10 nm by 2025; however, it would be limited because capacitors with high-aspect-ratio trenches can lean into each other [132]. To solve this physical problem, many studies have focused on 1T-DRAM devices in recent years, which are DRAM cells composed of a single transistor on a silicon-on-insulator (SOI) wafer [133]–[142]. Moreover, 1T-DRAM cells have the advantage of high scalability because information is stored using the floating-body effect in SOI transistors, and thus a capacitor is not required. However, they do not meet the requirement of retention time for DRAM applications and

there remains a need for an efficient technique that can improve the retention characteristics of 1T-DRAM devices [143]−[152]. The purpose of this study is to describe and examine how sensing margin and data retention characteristics can be enhanced by asymmetric dual-gate structures. Han et al. (2010) reported the similar idea of the asymmetric gate oxide concept [153]; however, the improvement of sensing margin and data retention properties caused by charge trapping was negligible because of the poly-silicon channel [154]−[156]. A 1T-DRAM device with a buried back gate was previously reported, but it has several issues such as non-self-aligned gate and complicated fabrication method [142]. In this Appendix B, significantly better sensing margin and retention characteristics are obtained using a negative V_{G2} or trapped electrons in the nitride layer allowing for advanced performance without the negative V_{G2}. The approach described here could serve as a solution for overcoming poor retention characteristics of 1T-DRAM devices.

At first, the kink effect was obtained in the output characteristics as shown in Fig. B.1. This is because excess holes were generated and accumulated in the floating body in the large V_D region. Interestingly, this floating-body effect stood out as V_{G2} changed from 0 V to −3 V because a negative V_{G2} can keep more excess holes in the floating body owing to the high storage capacity of positive charges.

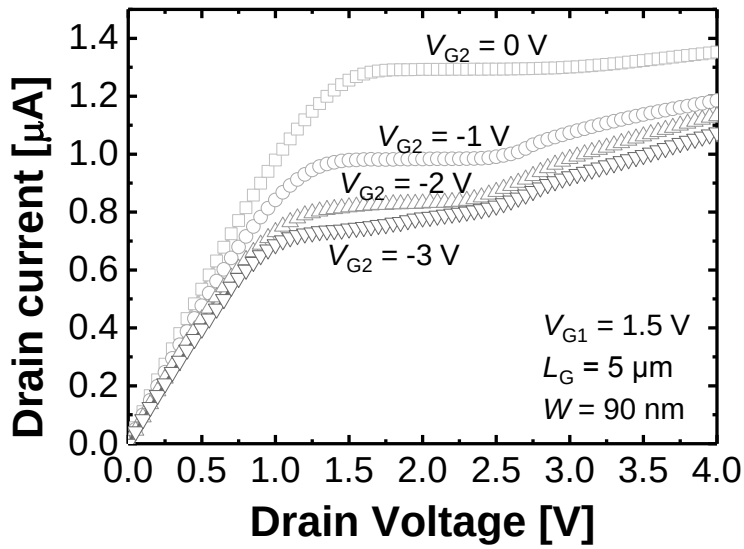


Fig. B.1. Measured output characteristics as a function of V_{G2}.

Figure B.2(a) shows the transient source current characteristics as a function of V_{G2} in the range from 0 V to −5 V. The source current in the read condition (V_{G1} = 1.5 V and V_D = 0.5 V) was increased because the accumulated holes were held in the floating body after the write "1" operation (V_{G1} = V_D = 1.5 V) for 20 ns, whereas it decreased after the write "0" operation (V_{G1} = 1.5 V and V_D = −1.5 V) for 20 ns because the negative V_D swept out stored excess holes from the floating body. Moreover, the source current difference between two states, defined as the sensing margin, was significantly increased as the V_{G2} decreased. This is consistent with previously discussed findings that the kink effect in the output characteristics becomes

significant for a negatively biased V_{G2}. Figure B.2(b) shows the read retention characteristics for various values of V_{G2} in the range from 0 V to −5 V. Here, the read retention time was defined as the time when the source current difference between the “1” and “0” states was decreased below 5%. The retention time was markedly increased from 1.8 μs to 1.5 s when V_{G2} was changed from 0 V to −3 V; however, it started to decrease when V_{G2} was decreased below −4 V and reached 0.13 s at V_{G2} = −5 V. This can be explained by considering the gate-induced-drain-leakage (GIDL) current near the G2 channel. When the negative V_{G2} is high enough to cause a tunneling current in the drain overlapped with G2, excess holes are generated by the GIDL mechanism and the retention properties of the “0” state become degraded. The sensing margin and read retention time were measured for 11 different single cells on the wafer. The statistical results are summarized in Figs. B.2(c) and B.2(d). On average, the sensing margin was improved by about 14 times as V_{G2} decreased from 0 V to −5 V; however, the retention time was increased by about 1.9×10^4 times when V_{G2} was decreased from 0 V to −3 V, and began to drop after when V_{G2} was below −4 V.

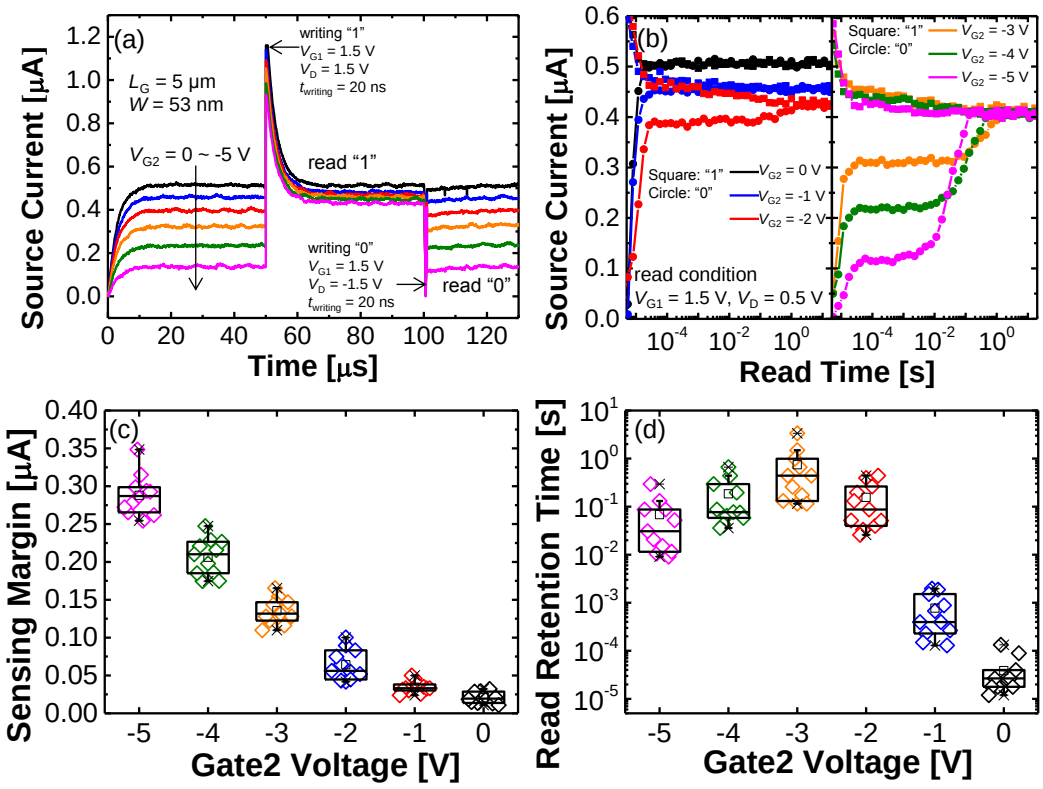


Fig. B.2. (a) Transient source current characteristics of writing and reading both the "1" and "0" states and (b) retention characteristics of each state with changing V_{G2} from 0 V to −5 V. The statistical distribution of (c) sensing margins and (d) read retention times with different V_{G2}.

The hold retention characteristics were also measured at room temperature and 85 ℃, as shown in Fig. B.3. Like the read retention time, the hold retention time was improved to over 1 s as V_{G2} decreasing to −3V, and then degraded when V_{G2} was below −4 V at both of the two temperatures. This is because a moderate negative V_{G2} enhances the hole capacity and retention characteristics of the device, but a high negative V_{G2} gives rise to an electric field, resulting in an increase in the electron−hole pair generation rate caused by band-to-band tunneling and accumulation of the excess holes generated in a short time. Figure B.4 shows the simulated contours of the band-to-band tunneling rate using a technology computer-aided design tool [126] and verifies that a higher V_{G2} generates electron−hole pairs at a faster rate.

Similar effects were observed when electrons were trapped in the nitride layer of the G2 stack. First, three programming pulses of 15 V and 10 μs were applied to G2 so as to trap charges in the nitride layer through Fowler-Nordheim tunneling. As a result, the threshold voltage of the device was increased as the parameter of programming pulses increased, and there were a total of four states with different numbers of trapped charges in the nitride layer as shown in Fig. B.5(a). The four states were the initial state, state 1 with one programming pulse biased, state 2 with two programming pulses biased, and state 3 with all of the three programming pulses biased. Then, the sensing

margin and retention time were measured for each state and divided by the number of trapped charges. Figure B.5(b) shows the transient source current characteristics for each state. Here, the bias conditions for all of the operations were the same as the previous ones except for V_{G2} = 0 V. The sensing margin was noticeably increased when charges were trapped in the nitride layer even without an applied V_{G2}. Sensing margins of 23.4 nA and 89.4 nA were obtained for the initial state and state 3, respectively. It is confirmed that the effect of the trapped electrons in the storage layer on the enhancement of accumulated holes in the floating body and on the sensing margin is very similar to that of the applied negative V_{G2}. The sensing margin and the read retention time for each state were also measured for 11 different single cells on the wafer. The results are summarized in Figs. B.5(c) and B.5(d). The average read retention time was improved to 0.52 s for state 2, but decreased to 40.3 ms for state 3, while the average sensing margin was increased by five times for state 3. These trends were in line with those of the applied negative V_{G2} because too many trapped electrons in the nitride layer can also induce tunneling current at the surface near G2. It is suggested that the sensing margin and retention characteristics can be also improved by trapped charges without additional biasing at G2.

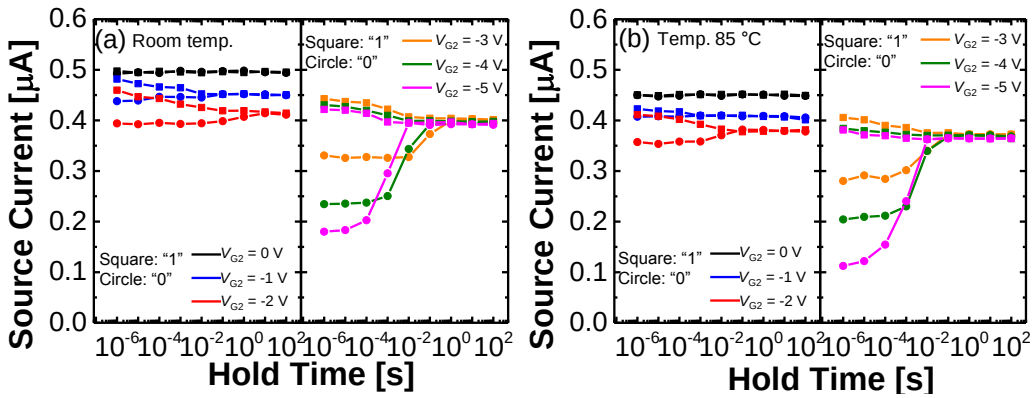


Fig. B.3. Hold retention characteristics as a function of V_{G2} at (a) room temperature and (b) 85 °C.

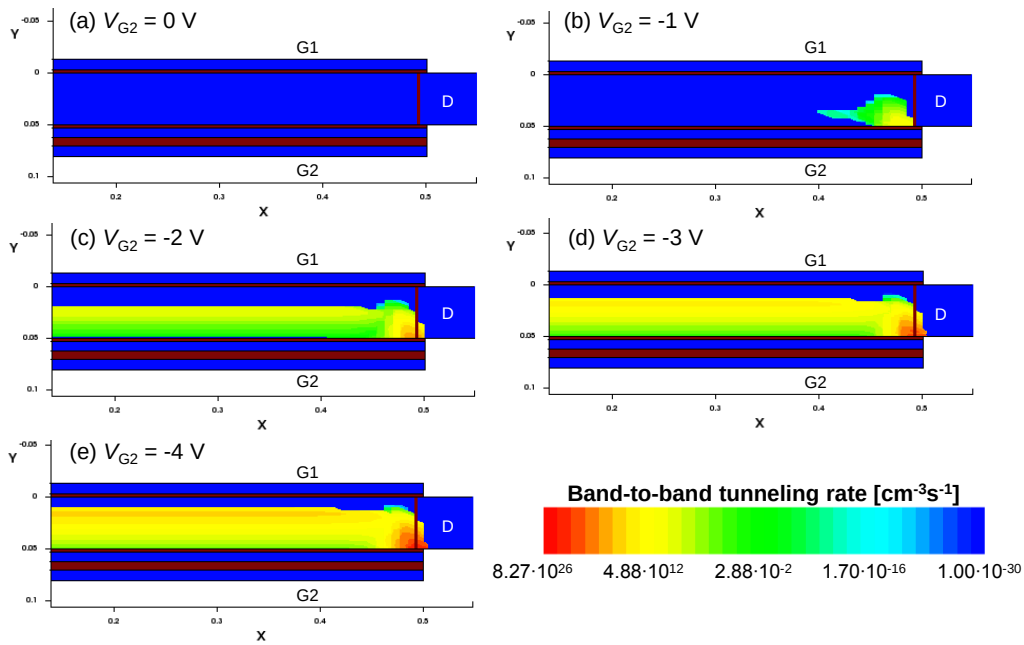


Fig. B.4. Simulated contours of band-to-band tunneling rate at the hold condition (V_{G1} = V_D = V_S = 0 V) with varying V_{G2}: (a) 0 V, (b) −1 V, (c) −2 V, (d) −3 V and (e) −4 V.

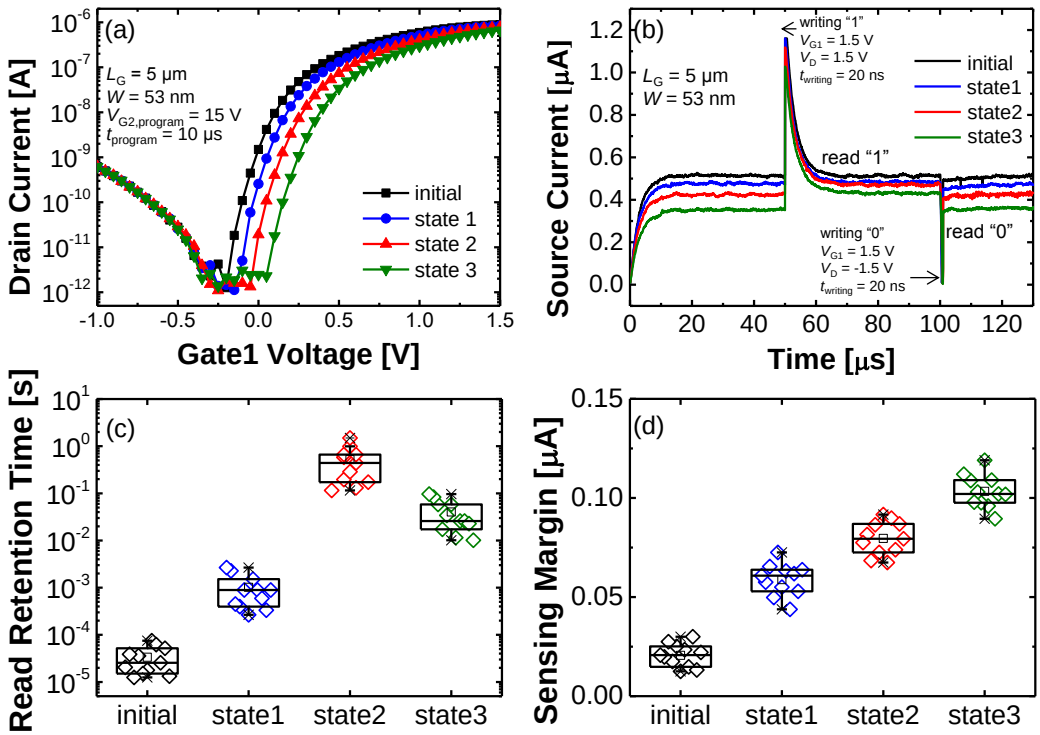


Fig. B.5. (a) Four split transfer curves and states by the number of trapped charges in the nitride layer and (b) transient source current characteristics for each state. The statistical distribution of (c) read retention times and (d) sensing margin for four states.

Additionally, the retention characteristics and soft programming issues of the G2 stack were measured. The accelerated charge loss at a high temperature of 85 ℃ was extracted to be 43% after 10 years as shown in Fig. B.6(a). The four states separated by charge trapping were predicted to be distinguishable over 10 years. The results of the soft programming issue during the 1T-DRAM operations are given in Fig. B.6(b). A dc stress (V_{G1} = V_D = 1.5 V, V_{G2} = 0 V), which is a programming condition of the 1T-DRAM by impact ionization, was applied in order to verify the immunity in the 1T-DRAM operations. It was found that the threshold voltages for four states were nearly unchanged under a dc stress for over 10^4 s. This is because the impact ionization during the 1T-DRAM operations occurred not near G2 but G1 and the two gates were apart by the channel width. These results indicate that the changes in the 1T-DRAM operation characteristics caused by the retention properties or soft programming of the G2 stack are negligible.

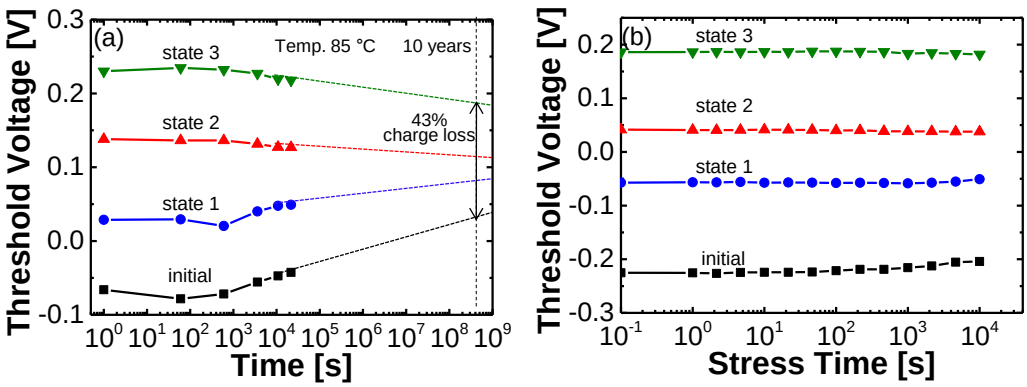


Fig. B.6. (a) Retention characteristics of four states at 85 ℃ and (b) soft programming issues under the dc stress (V_{G1} = V_D = 1.5 V, V_{G2} = 0 V, a programming condition of 1T-DRAM).

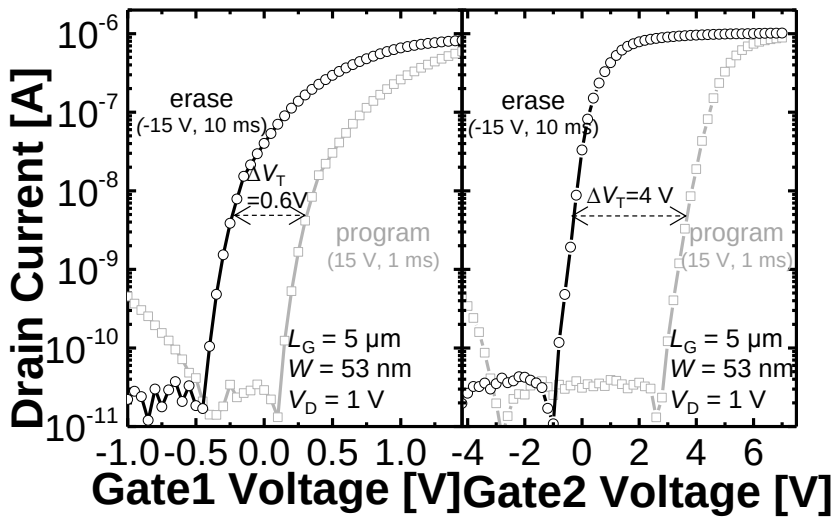
In conclusion, we attempted to enhance the sensing margin and retention characteristics of 1T-DRAM cells with two separated asymmetric gates by applying a negative V_{G2} and by trapping charges in the nitride layer. Both of the two methods could dramatically improve the sensing margin and retention time because of the increased storage capacity of holes in the floating body. However, a very high negative V_{G2} and a large amount of trapped charges reduced the retention characteristics of the "0" state because of the excess holes generated by the GIDL mechanism. This result indicates that two separated asymmetric gates would be a promising device structure for maximizing the sensing margin and retention properties of 1T-DRAM cells.

Appendix C. A Single Memory Cell with Voltaile and Non-Volatile Memory Functions

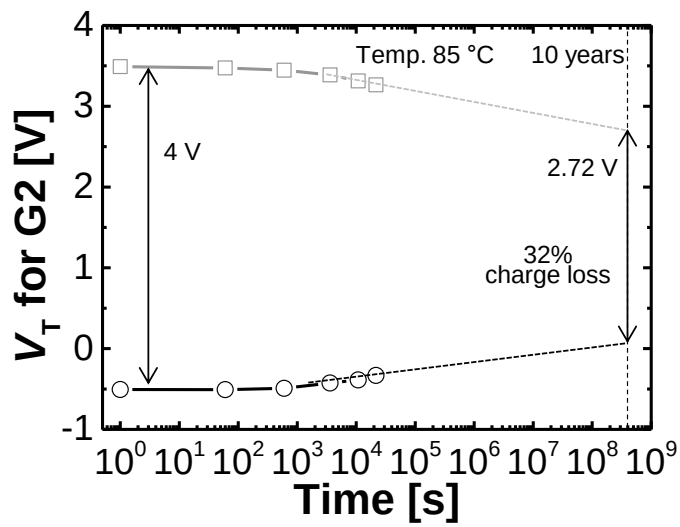
Combinations of volatile memory (VM) and non-volatile memory (NVM) functions in a single device are attracting widespread interest in fields such as embedded systems [153], [157]−[160]. Most of those approaches utilizing floating body effects have been based on partially depleted silicon-on-insulator (PDSOI) substrates in order to store excess generated holes instead of a capacitor for the VM function [160]−[163]. However, this results in a limitation on scaling down feature size and body thickness of a cell because PDSOI cannot effectively suppress leakage current such as drain-induced barrier lowering (DIBL), which is the major cause of poor retention properties of floating body effects [141], [164]−[165]. Han et al. (2010) reported the asymmetric gate dielectric stacks in a polysilicon channel device [153]; however, it also has trouble being scaled down because floating body effects of polysilicon devices mainly come from grain boundary traps in a polysilicon channel [154]−[156] and the random distributions of grain boundary traps can seriously degrade the device performance [166]−[168].

In this Appendix C, single-crystal silicon memory cell having both VM and NVM functions is demonstrated. The different gate stack structure of G1 G2 makes the device have both memory functions: the VM function by excess generated holes in the body and the NVM function by the charge trapping layer of G2. The floating body effects can be obtained even in a fully depleted body device thanks to the electric field effect of trapped electrons in the nitride layer. In addition, two writing methods, impact ionization (II) and gate-induced drain leakage (GIDL), for the VM function are compared in terms of soft-programming issues in the NVM function.

Figure C.1(a) shows the transfer characteristics of the two gates at the erased state and the programmed state for the NVM function. Each state was made by after applying V_{G2} of 15 V for 1 ms and -15 V for 10 ms, respectively. The V_T windows between the two states were 0.6 V for the G1 and 4 V for the G2, respectively. The ratio between these V_T windows is determined from the capacitance network as described in Appendix A. These findings indicate that the two gates are electrically independent and the two states for the NVM function are clearly distinguished. In addition, the retention characteristics of the NVM function at 85 ℃ were measured as shown in Fig. C.1(b). The V_T window and accelerated charge loss are extracted to be 2.74 V and only 32 %, respectively, even at a high temperature of 85 ℃ and even after 10 years.



(a)



(b)

Fig. C.1. Non-volatile memory function. (a) Transfer characteristics of the G1 and the G2 having V_T window as 0.6 V and 4 V, respectively. (b) Retention characteristics at 85 ℃.

Figures C.2(a) and C.2(b) show the transient source current characteristics when operating the device as a VM cell with two writing methods: II (V_{G1} = 1.5 V and V_D = 1.5 V for 20 ns) and GIDL (V_{G1} = -1.5 V and V_D = 1.5 V for 20 ns), respectively. The source currents of the "0" state in the read condition (V_{G1} = 1.5 V and V_D = 0.5 V) were almost the same, whereas those of the "1" state were quite different. When the GIDL method used, the read current of the "1" state was decreased at the erased state for the NVM function; on the other hand, the read current was increased at the programmed state compared with when the II method used. This comes from the fact that the trapped electrons in the nitride layer suppress II generation rates but enhance GIDL currents. The current density at II writing condition is decreased but the band-to-band tunneling current at the GIDL writing condition is increased at the programmed state due to the negative charges in the nitride layer.

In addition, the source current difference between the "1" state and the "0" state, defined as the sensing margin, was much larger at the programmed state than at the erased state regardless of which writing method used. At the erased state, there is very little room for excess holes storage because of the fully depleted body structure; however, the field-induced floating body effects can be obtained at the programmed state by the trapped electrons in the nitride

layer. The sensing margin was measured for 20 different single cells with the two writing methods and that statistical data are summarized in Fig. C.2(c). On average the sensing margin of the programmed cells is improved 1.5 times using the GIDL method and 18.6 times compared with the erased cells, which confirms that the GIDL method is more effective for large sensing margin at the programmed state. In order to verify the field-induced floating body effects, output characteristics of both two states were measured as shown in Fig. C.2(d). The kink effect was only obtained at the programmed state as expected, which is in line with the previously discussed results.

The hold retention characteristics depending on different writing methods and different NVM states were also measured and plotted in Fig. C.3. At the erased state, the sensing margin of under 0.05 μA was too little to distinguish two states and have enough retention time; here, the hold retention time of about 1 ms was obtained. In contrast, the enhanced sensing margin by the trapped electrons led to the improvement of the hold retention time which was obtained as about 10 s at the programmed state. The trapped electrons improve the performance of the VM operations in terms of the retention properties as well as the sensing margin. The difference in the hold retention time depending on the writing methods was insignificant because the retention time is dominantly determined by how long the “0” state stays at a low level.

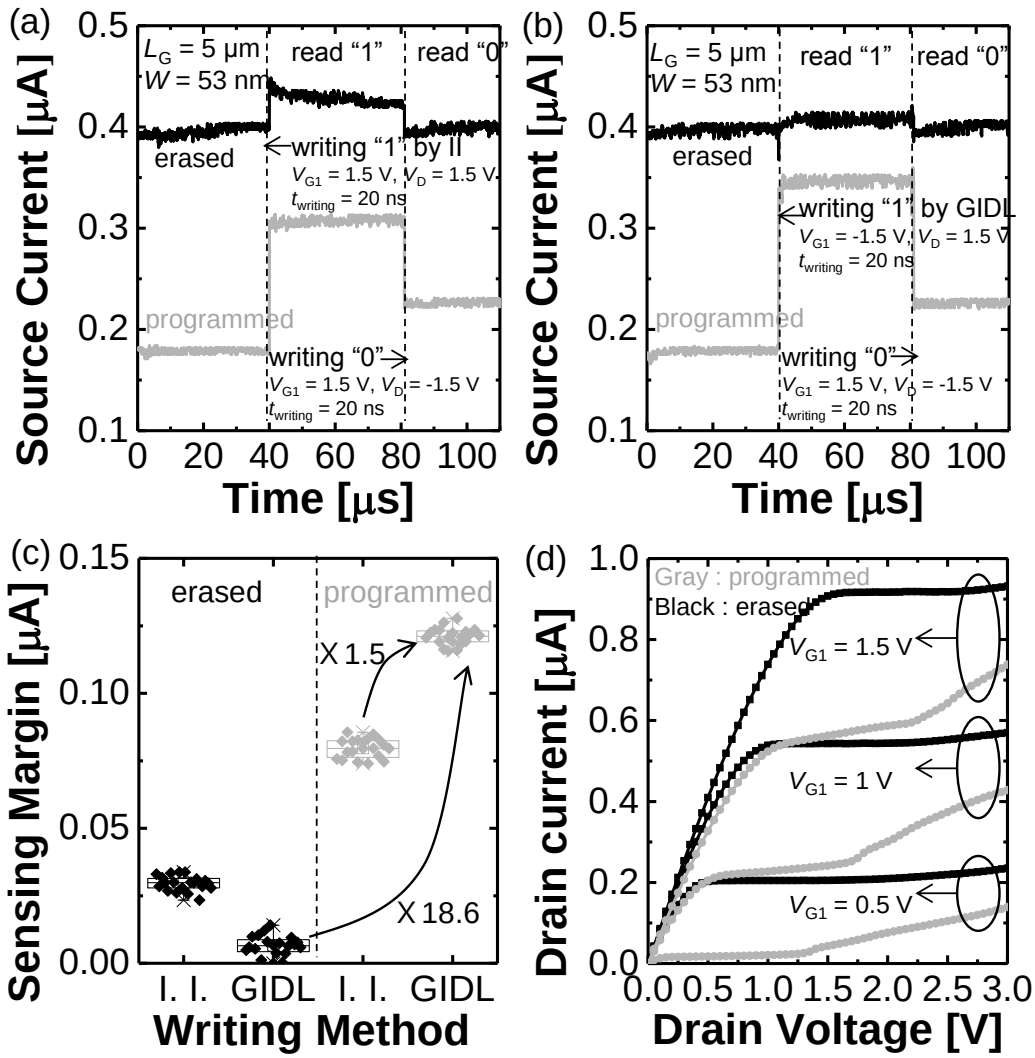


Fig. C.2. Volatile memory function. Transient source current characteristics using (a) II and (b) GIDL as a writing method. (c) Statistical distribution of the sensing margins. (d) Output characteristics depending on the state of NVM function.

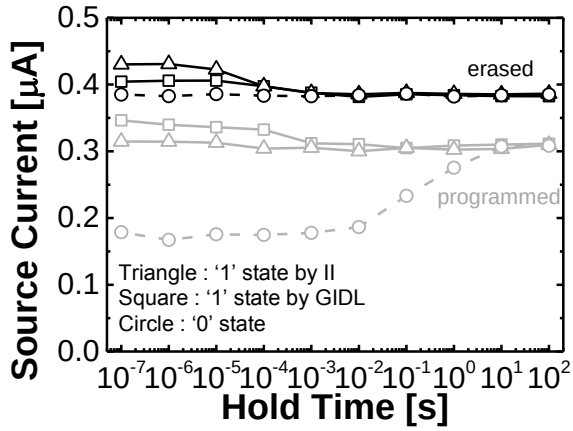


Fig. C.3. Hold retention characteristics with different programming methods and NVM states.

Lastly, the undesirable V_T disturbance properties depending on the writing methods of the VM operation are also investigated. Fig. C.4 shows the soft programming characteristics by the writing condition of the VM operation. The changes of V_T were measured under two dc stress conditions (II and GIDL), the same bias conditions as the writing methods of the VM function, over 4 hours. For the programmed cell, the V_T of the G1 was rarely changed regardless of which dc stress conditions used because the trapped charges prevent additional injections of electrons. In contrast, for the programmed cell, there was the V_T increase of 0.1 V when the II method used but little change when the GIDL method used. It seems that the GIDL method is an effective way to avoid undesirable changes of V_T when operating the device as a VM

cell.

In summary, a single memory cell with an asymmetric dual-gate structure is demonstrated to have VM and NVM functions and high scalability in terms of body thickness and feature size. The NVM function was obtained by trapping electrons in the nitride layer at the G2 and the VM function was achieved by the field-induced floating body effects even in a fully depleted body. The memory properties of both functions and their retention characteristics were measured and discussed. In addition, the GIDL writing method was effective in order not to unintentionally influence the NVM state when using the device as a VM cell. These results suggest that this device has potential as a multifunctional memory cell in embedded system applications.

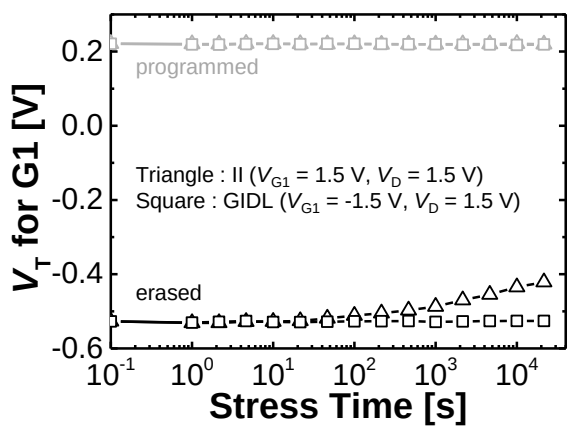


Fig. C.4. Soft programming characteristics under different dc stress conditions: II and GIDL methods.

Bibliography

[1] R. R. Schaller, "Moore's law: past, present and future," *IEEE Spectr.*, vol. 34, pp. 52-59, 1997.

[2] P. K. Bondyopadhyay, "Moore's law governs the silicon revolution," *Proc. IEEE,* vol. 86, pp. 78-81, 1998.

[3] C. A. Mack, "Fifty years of Moore's law," *IEEE Trans. Semicond. Manuf.,* vol. 24, pp. 202-207, 2011.

[4] L. B. Kish, "End of Moore's law: thermal (noise) death of integration in micro and nano electronics," *Phys. Lett. A,* vol. 305, pp. 144-149, 2002.

[5] S. E. Thompson and S. Parthasarathy, "Moore's law: the future of Si microelectronics," *Mater. Today,* vol. 9, pp. 20-25, 2006.

[6] M. Lundstrom, "Moore's law forever?," *Science,* vol. 299, pp. 210-211, 2003.

[7] M. M. Waldrop, "The chips are down for Moore’s law," *Nature*, vol. 530, p. 144-147, 2016.

[8] K. Ahmed and K. Schuegraf, "Transistor wars," *IEEE Spec.*, vol. 48, pp. 50-66, 2011.

[9] G. E. Moore, "No exponential is forever: but “Forever” can be delayed," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSC),* 2003, pp. 20-23.

[10] J. Von Neumann and M. D. Godfrey, "First draft of a report on the EDVAC," *IEEE Ann. Hist. Comput.*, vol. 15, pp. 27-75, 1993.

[11] A. Huang, "Architectural considerations involved in the design of an optical digital computer," *Proc. IEEE*, vol. 72, pp. 780-786, 1984.

[12] J. Backus, "Can programming be liberated from the von Neumann style?: a functional style and its algebra of programs," *Commun. ACM*, vol. 21, pp. 613-641, 1978.

[13] A. E. Pereda, "Electrical synapses and their functional interactions with chemical synapses," *Nat. Rev. Neurosci.*, vol. 15, p. 250-263, 2014.

[14] K. Krnjević, "Chemical nature of synaptic transmission in vertebrates," *Physiol. Rev.*, vol. 54, pp. 418-540, 1974.

[15] H. Kang and E. M. Schuman, "Long-lasting neurotrophin-induced enhancement of synaptic transmission in the adult hippocampus," *Science*, vol. 267, p. 1658-1662, 1995.

[16] R. Gray, A. S. Rajan, K. A. Radcliffe, M. Yakehiro, and J. A. Dani, "Hippocampal synaptic transmission enhanced by low concentrations of nicotine," *Nature*, vol. 383, pp. 713-716, 1996.

[17] T. V. Bliss and T. Lømo, "Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path," *J. Physiol.*, vol. 232, pp. 331-356, 1973.

[18] H. Barlow, "Temporal and spatial summation in human vision at different background intensities," *J. Physiol.*, vol. 141, p. 337-350, 1958.

[19] D. Oertel, "Use of brain slices in the study of the auditory system:

spatial and temporal summation of synaptic inputs in cells in the anteroventral cochlear nucleus of the mouse," *J. Acoust. Soc. Am.,* vol. 78, pp. 328-333, 1985.

[20] A. T. Gulledge, B. M. Kampa, and G. J. Stuart, "Synaptic integration in dendritic trees," *J. Neurobiol.,* vol. 64, pp. 75-90, 2005.

[21] E. R. Kandel, "The molecular biology of memory storage: a dialogue between genes and synapses," *Science,* vol. 294, pp. 1030-1038, 2001.

[22] R. D. Hawkins, E. R. Kandel, and C. H. Bailey, "Molecular mechanisms of memory storage in Aplysia," *Biol. Bull.,* vol. 210, pp. 174-191, 2006.

[23] G. Bi and M. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type," *J. Neurosci.,* vol. 18, pp. 10464-10472, 1998.

[24] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Science,* vol. 275, pp. 213-215, 1997.

[25] G. Bi and M. Poo, "Synaptic modification by correlated activity: Hebb's postulate revisited," *Annu. Rev. Neurosci.,* vol. 24, pp. 139-166, 2001.

[26] Y. Dan and M.-m. Poo, "Spike timing-dependent plasticity of neural circuits," *Neuron,* vol. 44, pp. 23-30, 2004.

[27] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron,* vol. 32, pp. 1149-1164, 2001.

[28] M. C. Van Rossum, G. Q. Bi, and G. G. Turrigiano, "Stable hebbian

learning from spike timing-dependent plasticity," *J. Neurosci.*, vol. 20, pp. 8812-8821, 2000.

[29] D. O. Hebb, *The organization of behavior: A neuropsychological approach*: John Wiley & Sons, 1949.

[30] M.-S. Rioult-Pedotti, J. P. Donoghue, and A. Dunaevsky, "Plasticity of the synaptic modification range," *J. Neurophysiol.*, vol. 98, pp. 3688-3695, 2007.

[31] M.-S. Rioult-Pedotti, D. Friedman, and J. P. Donoghue, "Learning-induced LTP in neocortex," *Science*, vol. 290, pp. 533-536, 2000.

[32] J. R. Whitlock, A. J. Heynen, M. G. Shuler, and M. F. Bear, "Learning induces long-term potentiation in the hippocampus," *Science*, vol. 313, pp. 1093-1097, 2006.

[33] C. Mead, "Neuromorphic electronic systems," *Proc. IEEE*, vol. 78, pp. 1629-1636, 1990.

[34] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, pp. 61-64, 2015.

[35] S. Ambrogio, S. Balatti, V. Milo, R. Carboni, Z.-Q. Wang, A. Calderoni, Zhong-Qiang Wang, Alessandro Calderoni, Nirmal Ramaswamy, and Daniele Ielmini, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM," *IEEE Trans. Electron Devices*, vol. 63, pp. 1508-1515, 2016.

[36] D. S. Jeong, K. M. Kim, S. Kim, B. J. Choi, and C. S. Hwang,

"Memristors for energy-efficient new computing paradigms," *Adv. Electron. Mater.,* vol.2, p. 1600090, 2016.

[37] S. Furber, "Large-scale neuromorphic computing systems," *J. Neural Eng.*, vol. 13, p. 051001, 2016.

[38] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H. S. P. Wong, "A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation," *Adv. Mater.,* vol. 25, pp. 1774-1779, 2013.

[39] A. Calimera, E. Macii, and M. Poncino, "The human brain project and neuromorphic computing," *Funct. Neurol.,* vol. 28, pp. 191-196, 2013.

[40] B. Rajendran, Y. Liu, J. Seo, K. Gopalakrishnan, L. Chang, D. J. Friedman, and M. B. Ritter, "Specifications of nanoscale devices and circuits for neuromorphic computational systems," *IEEE Trans. Electron Devices,* vol. 60, pp. 246-253, 2013.

[41] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip STDP learning," *Sci. Rep.*, vol. 6, p. 29545, 2016.

[42] I. Gupta, A. Serb, A. Khiat, R. Zeitler, S. Vassanelli, and T. Prodromakis, "Real-time encoding and compression of neuronal spikes by metal-oxide memristors," *Nat. Commun.*, vol. 7, p. 12805, 2016.

[43] S. Park, M. Chu, J. Kim, J. Noh, M. Jeon, B. H. Lee, H. Hwang, B. Lee, and B.-G. Lee, "Electronic system with memristive synapses for pattern recognition," *Sci. Rep.,* vol. 5, p. 10123, 2015.

[44] M. Chu, B. Kim, S. Park, H. Hwang, M. Jeon, B. H. Lee, and B.-G.

Lee, "Neuromorphic hardware system for visual pattern recognition with memristor array and CMOS neuron," *IEEE Trans. Ind. Electron.*, vol. 62, pp. 2410-2419, 2015.

[45] T. Serrano-Gotarredona, T. Prodromakis, and B. Linares-Barranco, "A proposal for hybrid memristor-CMOS spiking neuromorphic learning systems," *IEEE Circuits Syst. Mag.*, vol. 13, pp. 74-88, 2013.

[46] M. Hu, H. Li, Y. Chen, Q. Wu, G. S. Rose, and R. W. Linderman, "Memristor crossbar-based neuromorphic computing system: A case study," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 25, pp. 1864-1878, 2014.

[47] P. Zhang, C. Li, T. Huang, L. Chen, and Y. Chen, "Forgetting memristor based neuromorphic system for pattern training and recognition," *Neurocomputing*, vol.22, pp.47-56, 2017.

[48] S. Gaba, P. Sheridan, J. Zhou, S. Choi, and W. Lu, "Stochastic memristive devices for computing and neuromorphic applications," *Nanoscale*, vol. 5, pp. 5872-5878, 2013.

[49] Y. H. Liu, L. Q. Zhu, P. Feng, Y. Shi, and Q. Wan, "Freestanding artificial synapses based on laterally proton-coupled transistors on chitosan membranes," *Adv. Mater.*, vol. 27, pp. 5599-5604, 2015.

[50] S. Ramakrishnan, P. E. Hasler, and C. Gordon, "Floating gate synapses with spike-time-dependent plasticity," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, pp. 244-252, 2011.

[51] J. Shi, S. D. Ha, Y. Zhou, F. Schoofs, and S. Ramanathan, "A correlated nickelate synaptic transistor," *Nat. Commun.*, vol. 4, p. 2676, 2013.

[52] O. Bichler, W. Zhao, F. Alibart, S. Pleutin, D. Vuillaume, and C. Gamrat, "Functional model of a nanoparticle organic memory transistor for use as a spiking synapse," *IEEE Trans. Electron Devices*, vol. 57, pp. 3115-3122, 2010.

[53] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, pp. 1297-1301, 2010.

[54] X.-B. Yin, R. Yang, K.-H. Xue, Z. Tan, X.-D. Zhang, X.-S. Miao, and X. Guo, "Mimicking the brain functions of learning, forgetting and explicit/implicit memories with $SrTiO_3$-based memristive devices," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 31796-31802, 2016.

[55] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Mu, M. Barnell, G.-L. Li, H. L. Xin, R. S. Williams, Q. Xia, and J. J. Yang, "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," *Nat. Mater.*, p. 4756, 2016.

[56] Y.-F. Wang, Y.-C. Lin, I.-T. Wang, T.-P. Lin, and T.-H. Hou, "Characterization and modeling of nonfilamentary $Ta/TaO_x/TiO_2/Ti$ analog synaptic device," *Sci. Rep.*, vol. 5, p. 10150, 2015.

[57] M. Prezioso, F. M. Bayat, B. Hoskins, K. Likharev, and D. Strukov, "Self-adaptive spike-time-dependent plasticity of metal-oxide memristors," *Sci. Rep.*, vol. 6, p. 21331, 2016.

[58] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," *Nature Mater.*, vol. 10, pp. 591-595, 2011.

[59] Z. Q. Wang, H. Y. Xu, X. H. Li, H. Yu, Y. C. Liu, and X. J. Zhu, "Synaptic learning and memory functions achieved using oxygen ion migration/diffusion in an amorphous InGaZnO memristor," *Adv. Funct. Mater.*, vol. 22, pp. 2759-2765, 2012.

[60] M. D. Pickett, D. B. Strukov, J. L. Borghetti, J. J. Yang, G. S. Snider, D. R. Stewart, and R. S. Williams, "Switching dynamics in titanium dioxide memristive devices," *J. Appl. Phys.*, vol. 106, p. 074508, 2009.

[61] F. Miao, W. Yi, I. Goldfarb, J. J. Yang, M.-X. Zhang, M. D. Pickett, J. P. Strachan, G. Medeiros-Ribeiro, and R. S. Williams, "Continuous electrical tuning of the chemical composition of TaO_x-based memristors," *ACS Nano*, vol. 6, pp. 2312-2318, 2012.

[62] T. Chang, S.-H. Jo, and W. Lu, "Short-term memory to long-term memory transition in a nanoscale memristor," *ACS Nano*, vol. 5, pp. 7669-7676, 2011.

[63] S. Park, J. Noh, M.-l. Choo, A. M. Sheri, M. Chang, Y.-B. Kim, C. J. Kim, M. Jeon, B.-G. Lee, B. H. Lee, and H. Hwang, "Nanoscale RRAM-based synaptic electronics: toward a neuromorphic computing device," *Nanotechnology*, vol. 24, p. 384009, 2013.

[64] S. Ambrogio, S. Balatti, F. Nardi, S. Facchinetti, and D. Ielmini, "Spike-timing dependent plasticity in a transistor-selected resistive switching memory," *Nanotechnology*, vol. 24, p. 384012, 2013.

[65] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," *IEEE Trans. Electron Devices*, vol. 58, pp. 2729-2737, 2011.

[66] *Atlas User's Manual*, Ver. 5.20.2.R, Silvaco Inc., Santa Clara, CA,

2015.

[67] W. Shockley and W. Read Jr, "Statistics of the recombinations of holes and electrons," *Phys. Rev.,* vol. 87, pp. 835-842, 1952.

[68] R. N. Hall, "Electron-hole recombination in germanium," *Phys. Rev.,* vol. 87, p. 387, 1952.

[69] S. Selberherr, *Analysis and simulation of semiconductor devices*: Springer Science & Business Media, 1984.

[70] S. Tam, P.-K. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFET's," *IEEE Trans. Electron Devices,* vol. 31, pp. 1116-1125, 1984.

[71] S. Okhonin, M. Nagoga, and P. Fazan, "Principles of transient charge pumping on partially depleted SOI MOSFETs," *IEEE Electron Device Lett.,* vol. 23, pp. 279-281, 2002.

[72] J.-Y. Choi and J. G. Fossum, "Analysis and control of floating-body bipolar effects in fully depleted submicrometer SOI MOSFET's," *IEEE Trans. Electron Devices,* vol. 38, pp. 1384-1391, 1991.

[73] A. Wei, M. J. Sherony, and D. A. Antoniadis, "Effect of floating-body charge on SOI MOSFET design," *IEEE Trans. Electron Devices,* vol. 45, pp. 430-438, 1998.

[74] H. Shin, I.-S. Lim, M. Racanelli, W.-L. M. Huang, J. Foerstner, and B.-Y. Hwang, "Analysis of floating body induced transient behaviors in partially depleted thin film SOI devices," *IEEE Trans. Electron Devices,* vol. 43, pp. 318-325, 1996.

[75] J. Gautier and J.-C. Sun, "On the transient operation of partially depleted SOI NMOSFET's," *IEEE Electron Device Lett.,* vol. 16, pp.

497-499, 1995.

[76] P.-F. Lu, C.-T. Chuang, J. Ji, L. F. Wagner, C.-M. Hsieh, J. Kuang, L. L.-C. Hsu, M. M. Pelella, S.-F. S. Chu, and C. J. Anderson, "Floating-body effects in partially depleted SOI CMOS circuits," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1241-1253, 1997.

[77] *SmartSpice User's Manual*, Ver. 4.18.16.R, Silvaco Inc., Santa Clara, CA, 2016.

[78] *Athena User's Manual*, Ver. 5.20.0.R, Silvaco Inc., Santa Clara, CA, 2012.

[79] K. Izumi, M. Doken, and H. Ariyoshi, "CMOS devices fabricated on buried SiO2 layers formed by oxygen implantation into silicon," *Electron. Lett.*, vol. 14, pp. 593-594, 1978.

[80] Y. Kawai, A. Otaka, A. Tanaka, and T. Matsuda, "The effect of an organic base in chemically amplified resist on patterning characteristics using KrF lithography," *Jpn. J. Appl. Phys.*, vol. 33, pp. 7023-7027, 1994.

[81] M. Fayolle, J. Torres, G. Passemard, F. Fusalba, G. Fanget, D. Louis, M. Assous, O. Louveau, M. Rivoire, K. Haxaire, M. Mourier, S. Maitrejean, P. Besson, L. Broussous, L. Arnaud, and H. Feldis, "Integration of Cu/SiOC in Cu dual damascene interconnect for 0.1-μm technology," *Microelectron. Eng.*, vol. 64, pp. 35-42, 2002.

[82] A. A. Lamola, C. R. Szmanda, and J. W. Thackeray, "Chemically amplified resists," *Solid State Technol.*, vol. 34, pp. 53-61, 1991.

[83] K. R. Williams and R. S. Muller, "Etch rates for micromachining processing," *J. Microelectromech. Syst.*, vol. 5, pp. 256-269, 1996.

[84] H.-M. An, H.-D. Kim, Y. Zhang, Y. J. Seo, and T. G. Kim, "Substrate-bias assisted hot electron injection method for high-speed, low-voltage, and multi-bit flash memories," *Jpn J. Appl. Phys.*, vol. 50, p. 124201, 2011.

[85] Y. Kim, M. Kang, S. H. Park, and B.-G. Park, "Three-dimensional NAND flash memory based on single-crystalline channel stacked array," *IEEE Electron Device Lett.*, vol. 34, pp. 990-992, 2013.

[86] J.-G. Yun, G. Kim, J.-E. Lee, Y. Kim, W. B. Shim, J.-H. Lee, H. Shin, J. D. Lee, and B.-G. Park, "Single-crystalline Si stacked array (STAR) NAND flash memory," *IEEE Trans. Electron Devices*, vol. 58, pp. 1006-1014, 2011.

[87] K. Sonoda, M. Tanizawa, S. Shimizu, Y. Araki, S. Kawai, T. Ogura, S. Kobayashi, K. Ishikawa, T. Eimori, Y. Inoue, Y. Ohji, and N. Kotani, "Compact modeling of a flash memory cell including substrate-bias-dependent hot-electron gate current," *IEEE Trans. Electron Devices*, vol. 51, pp. 1726-1733, 2004.

[88] S. M. Bohte, J. N. Kok, and H. La Poutre, "Error-backpropagation in temporally encoded networks of spiking neurons," *Neurocomputing*, vol. 48, pp. 17-37, 2002.

[89] S. M. Bohte, H. La Poutré, and J. N. Kok, "Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks," *IEEE Trans. Neural Netw.*, vol. 13, pp. 426-435, 2002.

[90] T. Masquelier, R. Guyonneau, and S. J. Thorpe, "Competitive STDP-based spike pattern learning," *Neural Comput.*, vol. 21, pp. 1259-1276, 2009.

[91] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol.

521, pp. 436-444, 2015.

[92] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013.

[93] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807-814.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012.

[95] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, 2015.

[96] Z. Zhang, S. Zhang, and M. Chan, "Self-align recessed source drain ultrathin body SOI MOSFET," *IEEE Electron Device Lett.*, vol. 25, pp. 740-742, 2004.

[97] Y.-K. Choi, D. Ha, T.-J. King, and C. Hu, "Nanoscale ultrathin body PMOSFETs with raised selective germanium source/drain," *IEEE Electron Device Lett.*, vol. 22, pp. 447-448, 2001.

[98] K. Uchida and S. Takagi, "Carrier scattering induced by thickness fluctuation of silicon-on-insulator film in ultrathin-body metal–oxide–semiconductor field-effect transistors," *Appl. Phys. Lett.*, vol. 82, pp. 2916-2918, 2003.

[99] S. Xiong, T.-J. King, and J. Bokor, "A comparison study of

symmetric ultrathin-body double-gate devices with metal source/drain and doped source/drain," *IEEE Trans. Electron Devices,* vol. 52, pp. 1859-1867, 2005.

[100] S. Jin, M. V. Fischetti, and T.-W. Tang, "Modeling of surface-roughness scattering in ultrathin-body SOI MOSFETs," *IEEE Trans. Electron Devices,* vol. 54, pp. 2191-2203, 2007.

[101] H. Fang, S. Chuang, K. Takei, H. S. Kim, E. Plis, C.-H. Liu, S. Krishna, Y.-L. Chueh, and A. Javey, "Ultrathin-body high-mobility InAsSb-on-insulator field-effect transistors," *IEEE Electron Device Lett.,* vol. 33, pp. 504-506, 2012.

[102] N. Agrawal, Y. Kimura, R. Arghavani, and S. Datta, "Impact of transistor architecture (bulk planar, trigate on bulk, ultrathin-body planar SOI) and material (silicon or III–V Semiconductor) on variation for logic and SRAM applications," *IEEE Trans. Electron Devices,* vol. 60, pp. 3298-3304, 2013.

[103] K. Takei, S. Chuang, H. Fang, R. Kapadia, C.-H. Liu, J. Nah, H. S. Kim, E. Plis, S. Krishna, Y.-L. Chueh, and A. Javey, "Benchmarking the performance of ultrathin body InAs-on-insulator transistors as a function of body thickness," *Appl. Phys. Lett.,* vol. 99, p. 103507, 2011.

[104] K. Alam, S. Takagi, and M. Takenaka, "Analysis and comparison of L-valley transport in GaAs, GaSb, and Ge ultrathin-body ballistic nMOSFETs," *IEEE Trans. Electron Devices,* vol. 60, pp. 4213-4218, 2013.

[105] M.-C. Chang, C.-S. Chang, C.-P. Chao, K.-I. Goto, M. Ieong, L.-C. Lu, and C. H. Diaz, "Transistor-and circuit-design optimization for low-power CMOS," *IEEE Trans. Electron Devices,* vol. 55, pp. 84-95, 2008.

[106] J.-P. Noel, O. Thomas, M. Jaud, O. Weber, T. Poiroux, C. Fenouillet-Beranger, P. Rivallin, P. Scheiblin, F. Andrieu, M. Vinet, O. Rozeau, F. Boeuf, O. Faynot, and A. Amara, "Multi-UTBB FDSOI device architectures for low-power CMOS circuit," *IEEE Trans. Electron Devices,* vol. 58, pp. 2473-2482, 2011.

[107] D. Jacquet, F. Hasbani, P. Flatresse, R. Wilson, F. Arnaud, G. Cesana, T. D. Gilio, C. Lecocq, T. Roy, A. Chhabra, C. Grover, O. Minez, J. Uginet, G. Durieu, C. Adobati, D. Casalotto, F. Nyer, P. Menut, A. Cathelin, I. Vongsavady, and P. Magarshack, "A 3 GHz dual core processor ARM cortex TM-A9 in 28 nm UTBB FD-SOI CMOS with ultra-wide voltage range and energy efficiency optimization," *IEEE J. Solid-State Circuits,* vol. 49, pp. 812-826, 2014.

[108] T. Numata, T. Mizuno, T. Tezuka, J. Koga, and S. Takagi, "Control of threshold-voltage and short-channel effects in ultrathin strained-SOI CMOS devices," *IEEE Trans. Electron Devices,* vol. 52, pp. 1780-1786, 2005.

[109] S. Burignat, D. Flandre, M. M. Arshad, V. Kilchytska, F. Andrieu, O. Faynot, and J.-P. Raskin, "Substrate impact on threshold voltage and subthreshold slope of sub-32 nm ultra thin SOI MOSFETs with thin buried oxide and undoped channel," *Solid-State Electron.,* vol. 54, pp. 213-219, 2010.

[110] C. Fenouillet-Beranger, O. Thomas, P. Perreau, J.-P. Noel, A. Bajolet, S. Haendler, L. Tosti, S. Barnola, R. Beneyto, C. Perrot, C. de Buttet, F. Abbate, F. Baron, B. Pernet, Y. Campidelli, L. Pinzelli, P. Gouraud, M. Cassé, C. Borowiak, O. Weber, F. Andrieu, K. K. Bourdelle, B. Y. Nguyen, F. Boedt, S. Denorme, F. Boeuf, O. Faynot, and T. Skotnicki, "Efficient multi-V_T FDSOI technology with UTBOX for low power circuit

design," in *Proc. Symp. VLSI Tech.*, 2010, pp. 65-66.

[111] R. Tsuchiya, M. Horiuchi, S. Kimura, M. Yamaoka, T. Kawahara, S. Maegawa, T. Ipposhi, Y. Ohji, and H. Matsuoka, "Silicon on thin BOX: A new paradigm of the CMOSFET for low-power high-performance application featuring wide-range back-bias control," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2004, pp. 631-634.

[112] K. Ikeda, Y. Moriyama, M. Ono, Y. Kamimuta, T. Irisawa, Y. Kamata, A. Sakai, and T. Tezuka, "First demonstration of threshold voltage control by sub-1 V back-gate biasing for thin body and buried-oxide (TBB) Ge-on-insulator (GOI) MOSFETs for low-power operation," in *Proc. IEEE Int. SOI Conf.*, 2012.

[113] R. Muralidhar, J. Cai, D. J. Frank, P. Oldiges, D. Lu, and I. Lauer, "Meeting the challenge of multiple threshold voltages in highly scaled undoped FinFETs," *IEEE Trans. Electron Devices*, vol. 60, pp. 1276-1278, 2013.

[114] C. Fenouillet-Beranger, S. Denorme, P. Perreau, C. Buj, O. Faynot, F. Andrieu, L. Tosti, S. Barnola, T. Salvetat, X. Garros, M. Casse, F. Allain, N. Loubet, L. Pham-Nguyen, E. Deloffre, M. Gros-Jean, R. Beneyton, C. Laviron, M. Marin, C. Leyris, S. Haendler, F. Leverd, P. Gouraud, P. Scheiblin, L. Clement, R. Pantel, S. Delenibus, and T. Skotnicki, "FDSOI devices with thin BOX and ground plane integration for 32 nm node and below," *Solid-State Electron.*, vol. 53, pp. 730-734, 2009.

[115] M.-C. Sun, H. W. Kim, H. Kim, S. W. Kim, G. Kim, J.-H. Lee, H. Shin, B.-G. Park, "V_T-modulation of planar tunnel field-effect transistors with ground-plane under ultrathin body and bottom oxide," *J. Semicond. Technol. Sci.*, vol. 14, pp. 139-145, 2014.

[116] M. M. Arshad, S. Makovejev, S. Olsen, F. Andrieu, J.-P. Raskin, D. Flandre, and V. Kilchytska, "UTBB SOI MOSFETs analog figures of merit: Effects of ground plane and asymmetric double-gate regime," *Solid-State Electron.*, vol. 90, pp. 56-64, 2013.

[117] C. Fenouillet-Beranger, P. Perreau, S. Denorme, L. Tosti, F. Andrieu, O. Weber, S. Monfray, S. Barnola, C. Arvet, Y. Campidelli, S. Haendler, R. Beneyton, C. Perrot, C. de Buttet, P. Gros, L. Pham-Nguyen, F. Leverd, P. Gouraud, F. Abbate, F. Baron, A. Torres, C. Laviron, L. Pinzelli, J. Vetier, C. Borowiak, A. Margain, D. Delprat, F. Boedt, K. Bourdelle, B.-Y. Nguyen, O. Faynot, and T. Skotnicki, "Impact of a 10 nm ultra-thin BOX (UTBOX) and ground plane on FDSOI devices for 32 nm node and below," *Solid-State Electron.*, vol. 54, pp. 849-854, 2010.

[118] O. Weber, F. Andrieu, J. Mazurier, M. Casse, X. Garros, C. Leroux, F. Martin, P. Perreau, C. Fenouillet-Béranger, S. Barnola, R. Gassilloud, C. Arvet, O. Thomas, J.-P. Noel, O. Rozeau, M.-A. Jaud, T. Poiroux, D. Lafond, A. Toffoli, F. Allain, C. Tabone, L. Tosti, L. Brevard, P. Lehnen, U. Weber, P. K. Baumann, O. Boissiere, W. Schwarzenbach, K. Bourdelle, B.-Y. Nguyen, F. Breuf, T. Skotnicki, and O. Faynot, "Work-function engineering in gate first technology for multi-V_T dual-gate FDSOI CMOS on UTBOX," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2010, pp. 58-61.

[119] H. Kim, M.-C. Sun, H. W. Kim, S. W. Kim, G. Kim, and B.-G. Park, "Study on threshold voltage control of tunnel field-effect transistors using V_T-control doping region," *IEICE Trans. Electron.*, vol. 95, pp. 820-825, 2012.

[120] C. M. Lai, C. T. Lin, L. W. Cheng, C. H. Hsu, J. T. Tseng, T. F. Chiang, C. H. Chou, Y. W. Chen, C. H. Yu, S. H. Hsu, C. G. Chen, Z. C.

Lee, J. F. Lin, C. L. Yang, G. H. Ma, and S. C. Chien, "A novel “hybrid” high-k/metal gate process for 28 nm high performance CMOSFETs," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2009, pp. 655-658.

[121] K. Mistry, C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, M. Buehler, A. Cappellani, R. Chau, C.-H. Choi, G. Ding, K. Fischer, T. Ghani, R. Grover, W. Han, D. Hanken, M. Hattendorf, J. He, J. Hicks, R. Huessner, D. Ingerly, P. Jain, R. James, L. Jong, S. Joshi, C. Kenyon, K. Kuhn, K. Lee, H. Liu, J. Maiz, B. Mcintyre, P. Moon, J. Neirynck, S. Pae, C. Parker, D. Parsons, C. Prasad, L. Pipes, M. Prince, P. Ranade, T. Reynolds, J. Sandford, L. Shifren, J. Sebastian, J. Seiple, D. Simon, S. Sivakumar, P. Smith, C. Thomas, T. Troeger, P. Vandervoorn, S. Willams, and K. Zawadzki, "A 45 nm logic technology with high-k+ metal gate transistors, strained silicon, 9 Cu interconnect layers, 193 nm dry patterning, and 100% Pb-free packaging," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2007, pp. 247-250.

[122] T. Chan, K. Young, and C. Hu, "A true single-transistor oxide-nitride-oxide EEPROM device," *IEEE Electron Device Lett.*, vol. 8, pp. 93-95, 1987.

[123] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, K. C. Park, and K. Kim, "A novel SONOS structure of SiO_2/SiN/Al_2O_3 with TaN metal gate for multi-giga bit flash memories," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2003, pp. 613-616.

[124] S. Cho, W. B. Shim, Y. Kim, J.-G. Yun, J. D. Lee, H. Shin J.-H. Lee, and B.-G. Park, "A charge trap folded nand flash memory device with band-gap-engineered storage node," *IEEE Trans. Electron Devices*, vol. 58, pp. 288-295, 2011.

[125] K.-D. Suh, B.-H. Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, Y.-N. Koh, S.-S. Lee, S.-C. Kwon, B.-S. Choi, J.-S. Yum, J.-H. Choi, J.-R. Kim, and H.-K. Lim, "A 3.3 V 32 Mb NAND flash memory with incremental step pulse programming scheme," *IEEE J. Solid-State Circuits*, vol. 30, pp. 1149-1156, 1995.

[126] *Sentaurus Device User Guide*, Ver. K-2015.06, Synopsys Inc., Mountain View, CA, 2015.

[127] S. W. Lee, J. H. Han, S. Han, W. Lee, J. H. Jang, M. Seo, S. K. Kim, C. Dussarrat, J. Gatineau, and Y.-S. Min, "Atomic layer deposition of SrTiO3 thin films with highly enhanced growth rate for ultrahigh density capacitors," *Chem. Mater.*, vol. 23, pp. 2227-2236, 2011.

[128] B. Kaczer, S. Clima, K. Tomida, B. Govoreanu, M. Popovici, M.-S. Kim, J. Swerts, A. Belmonte, W.-C. Wang, V. Afanas'ev, A. S. Verhulst, G. Pourtois, G. Groeseneken, and M. Jurczak, "Considerations for further scaling of metal–insulator–metal DRAM capacitors," *J. Vac. Sci. Technol. B*, vol. 31, p. 01A105, 2013.

[129] M. Popovici, J. Swerts, A. Redolfi, B. Kaczer, M. Aoulaiche, I. Radu, S. Clima, J.-L. Everaert, S. Van Elshocht, and M. Jurczak, "Low leakage Ru-strontium titanate-Ru metal-insulator-metal capacitors for sub-20 nm technology node in dynamic random access memory," *Appl. Phys. Lett.*, vol. 104, p. 082908, 2014.

[130] M. Pešić, S. Knebel, K. Cho, C. Jung, J. Chang, H. Lim, N. Kolomiiets, V. V. Afanas'ev, T. Mikolajick, and U. Schroeder, "Conduction barrier offset engineering for DRAM capacitor scaling," *Solid-State Electron.*, vol. 115, pp. 133-139, 2016.

[131] B. Kaczer, L. Larcher, L. Vandelli, H. Reisinger, M. Popovici, S.

Clima, Z. Ji, S. Joshi, J. Swerts, A. Redolfi, V. V. Afanas'ev, and M. Jurczak, "$SrTiO_x$ for sub-20 nm DRAM technology nodes—characterization and modeling," *Microelectron. Eng.*, vol. 147, pp. 126-129, 2015.

[132] D. Kim, J. Y. Kim, M. Huh, Y. S. Hwang, J. Park, D. Han, D. I. Kim, M. H. Cho, B. H. Lee, H. K. Hwang, J. W. Song, N. J. Kang, G. W. Ha, S. S. Song, M. S. Shim, S. E. Kim, J. M. Kwon, B. J. Park, H. J. Oh, H. J. Kim, D. S. Woo, M. Y. Jeong, Y. I. Kim, Y. S. Lee, J. C. Shin, J. W. Seo, S. S. Jeong, K. H. Yoon, T. H. Ahn, J. B. Lee, Y.W. Hyung, S. J. Park, W. T. Choi, G. Y. Jin, Y. G. Park, and K. Kim, "A mechanically enhanced storage node for virtually unlimited height (MESH) capacitor aiming at sub 70 nm DRAMs," in *Proc. IEEE Int. Electron Devices Meet. (IEDM)*, 2004, pp. 69-72.

[133] S. Okhonin, M. Nagoga, J. Sallese, and P. Fazan, "A capacitor-less 1T-DRAM cell," *IEEE Electron Device Lett.*, vol. 23, pp. 85-87, 2002.

[134] E. Yoshida and T. Tanaka, "A capacitorless 1T-DRAM technology using gate-induced drain-leakage (GIDL) current for low-power and high-speed embedded memory," *IEEE Trans. Electron Devices*, vol. 53, pp. 692-697, 2006.

[135] H. Jeong, K.-W. Song, I. H. Park, T.-H. Kim, Y. S. Lee, S.-G. Kim, J. Seo, K. Cho, K. Lee, H. Shin, J. D. Lee, and B.-G. Park, "A new capacitorless 1T DRAM cell: Surrounding gate MOSFET with vertical channel (SGVC cell)," *IEEE Trans. Nanotechnol.*, vol. 6, pp. 352-357, 2007.

[136] K.-S. Shim, I.-Y. Chung, and Y. J. Park, "A BJT-based heterostructure 1T-DRAM for low-voltage operation," *IEEE Electron*

Device Lett., vol. 33, pp. 14-16, 2012.

[137] T.-H. Kim and J.-G. Park, "Capacitor-less memory cell fabricated on nano-scale strained Si on a relaxed SiGe layer-on-insulator," *Semicond. Sci. Technol.*, vol. 28, p. 045001, 2013.

[138] A. Biswas, N. Dagtekin, W. Grabinski, A. Bazigos, C. Le Royer, J.-M. Hartmann, C. Tabone, M. Vinet, and A. M. Ionescu, "Investigation of tunnel field-effect transistors as a capacitor-less memory cell," *Appl. Phys. Lett.*, vol. 104, p. 092108, 2014.

[139] N. Rodriguez, S. Cristoloveanu, and F. Gamiz, "Novel capacitorless 1T-DRAM cell for 22-nm node compatible with bulk and SOI substrates," *IEEE Trans. Electron Devices*, vol. 58, pp. 2371-2377, 2011.

[140] Z. Zhou, J. G. Fossum, and Z. Lu, "Physical insights on BJT-based 1T DRAM cells," *IEEE Electron Device Lett.*, vol. 30, pp. 565-567, 2009.

[141] C. Kuo, T.-J. King, and C. Hu, "A capacitorless double gate DRAM technology for sub-100-nm embedded and stand-alone memory applications," *IEEE Trans. Electron Devices*, vol. 50, pp. 2408-2416, 2003.

[142] K.-H. Park, S. Cristoloveanu, M. Bawedin, Y. Bae, K.-I. Na, and J.-H. Lee, "Double-gate 1T-DRAM cell using nonvolatile memory function for improved performance," *Solid-State Electron.*, vol. 59, pp. 39-43, 2011.

[143] K. R. A. Sasaki, T. Nicoletti, L. Almeida, S. dos Santos, A. Nissimoff, M. Aoulaiche, E. Simoen, C. Claeys, and J. A. Martino, "Improved retention times in UTBOX nMOSFETs for 1T-DRAM

applications," *Solid-State Electron.,* vol. 97, pp. 30-37, 2014.

[144] G. Kim, S. W. Kim, K.-C. Ryoo, J.-H. Oh, M.-C. Sun, H. W. Kim, D. W. Kwon, J. S. Jang, S. Jung, J. H. Kim, and B.-G. Park, "Split-gate-structure 1T DRAM for retention characteristic improvement," *J. Nanosci. Nanotechnol.,* vol. 11, pp. 5603-5607, 2011.

[145] T. Nicoletti, M. Aoulaiche, L. M. Almeida, S. D. Santos, J. A. Martino, A. Veloso, M. Jurczak, E. Simoen, and C. Claeys, "The dependence of retention time on gate length in UTBOX FBRAM with different source/drain junction engineering," *IEEE Electron Device Lett.,* vol. 33, pp. 940-942, 2012.

[146] J. S. Shin, H. Choi, H. Bae, J. Jang, D. Yun, E. Hong, D. H. Kim, and D. M. Kim, "Vertical-gate Si/SiGe double-HBT-based capacitorless 1T DRAM cell for extended retention time at low latch voltage," *IEEE Electron Device Lett.,* vol. 33, pp. 134-136, 2012.

[147] A. Nissimoff, J. A. Martino, M. Aoulaiche, A. Veloso, L. J. Witters, E. Simoen, and C. Claeys, "Spike anneal peak temperature impact on 1T-DRAM retention time," *IEEE Electron Device Lett.,* vol. 35, pp. 639-641, 2014.

[148] A. Pal, A. Nainani, S. Gupta, and K. C. Saraswat, "Performance improvement of one-transistor DRAM by band engineering," *IEEE Electron Device Lett.,* vol. 33, pp. 29-31, 2012.

[149] M. Aoulaiche, E. Simoen, C. Caillat, L. Witters, K. Bourdelle, B.-Y. Nguyen, J. Martino, C. Claeys, P. Fazan, and M. Jurczak, "Understanding and optimizing the floating body retention in FDSOI UTBOX," *Solid-State Electron.,* vol. 117, pp. 123-129, 2016.

[150] J.-T. Lin, P.-H. Lin, Y.-C. Eng, and Y.-R. Chen, "Novel vertical SOI-

based 1T-DRAM with trench body structure," *IEEE Trans. Electron Devices*, vol. 60, pp. 1872-1877, 2013.

[151] J.-T. Lin, P.-H. Lin, S. W. Haga, Y.-C. Wang, and D.-R. Lu, "Transient and thermal analysis on disturbance immunity for 4 surrounding gate 1T-DRAM with wide trenched body," *IEEE Trans. Electron Devices*, vol. 62, pp. 61-68, 2015.

[152] M. Aoulaiche, A. Bravaix, E. Simoen, C. Caillat, M. Cho, L. Witters, P. Fazan, G. Groeseneken, and M. Jurczak, "Endurance of one transistor floating body RAM on UTBOX SOI," *IEEE Trans. Electron Devices*, vol. 61, pp. 801-805, 2014.

[153] J.-W. Han, S.-W. Ryu, D.-H. Kim, and Y.-K. Choi, "Polysilicon channel TFT with separated double-gate for Unified RAM (URAM)—unified function for nonvolatile SONOS flash and high-speed capacitorless 1T-DRAM," *IEEE Trans. Electron Devices*, vol. 57, pp. 601-607, 2010.

[154] M. Hack and A. G. Lewis, "Avalanche-induced effects in polysilicon thin-film transistors," *IEEE Electron Device Lett.*, vol. 12, pp. 203-205, 1991.

[155] M. Hack, A. Lewis, and J. Shaw, "Influence of traps on the characteristics of thin film transistors," *Journal of Non-Cryst. Solids*, vol. 137, pp. 1229-1232, 1991.

[156] M. Valdinoci, L. Colalongo, G. Baccarani, G. Fortunato, A. Pecora, and I. Policicchio, "Floating body effects in polysilicon thin-film transistors," *IEEE Trans. on Electron Devices*, vol. 44, pp. 2234-2241, 1997.

[157] C. W. Oh, N. Y. Kim, H. J. Song, S. I. Hong, S. H. Kim, Y. L. Choi,

H. J. Bae, D. U. Choi, Y. S. Lee, D.-W. Kim, D. Park, and B.-I. Ryu, "Floating body DRAM characteristics of silicon-On-ONO (SOONO) devices for system-on-Chip (SoC) applications," in *Proc. Symp. VLSI Tech.*, 2007, pp. 168-169.

[158] J.-K. Park and W.-J. Cho, "Dual read method by capacitance coupling effect for mode-disturbance-free operation in channel-recessed multifunctional memory," *IEEE Electron Device Lett.*, vol. 33, pp. 1708-1710, 2012.

[159] J.-W. Han, S.-W. Ryu, C.-J. Kim, S. Kim, M. Im, S. J. Choi, J. S. Kim, K. H. Kim, G. S. Lee, J. S. Oh, M. H. Song, Y. C. Park, J. W. Kim, and Y.-K. Choi, "Partially depleted SONOS FinFET for unified RAM (URAM)—Unified function for high-speed 1T DRAM and nonvolatile memory," *IEEE Electron Device Lett.*, vol. 29, pp. 781-783, 2008.

[160] B.-H. Lee, D.-C. Ahn, M.-H. Kang, S.-B. Jeon, and Y.-K. Choi, "Vertically integrated nanowire-based unified memory," *Nano Lett.*, vol. 16, pp. 5909-5916, 2016.

[161] S.-W. Ryu, J.-W. Han, C.-J. Kim, and Y.-K. Choi, "Investigation of isolation-dielectric effects of PDSOI FinFET on capacitorless 1T-DRAM," *IEEE Trans. Electron Devices*, vol. 56, pp. 3232-3235, 2009.

[162] J.-T. Lin and P.-H. Lin, "Multifunction behavior of a vertical MOSFET with trench body structure and new erase mechanism for use in 1T-DRAM," *IEEE Trans. Electron Devices*, vol. 61, pp. 3172-3178, 2014.

[163] M. Lee, T. Moon, and S. Kim, "Floating body effect in partially depleted silicon nanowire transistors and potential capacitor-less one-transistor DRAM applications," *IEEE Trans. Nanotechnol.*, vol. 11, pp.

355-359, 2012.

[164] T. Shino, T. Ohsawa, T. Higashi, K. Fujita, N. Kusunoki, Y. Minami, Morikado, H. Nakajima, K. Inoh, T. Hamamoto, and A. Nitayama, "Operation voltage dependence of memory cell characteristics in fully depleted floating-body cell," *IEEE Trans. Electron Devices*, vol. 52, pp. 2220-2226, 2005.

[165] M.-S. Kim and W.-J. Cho, "Characteristics of fully depleted strained-silicon-on-insulator capacitorless dynamic random access memory cells," *IEEE Electron Device Lett.*, vol. 30, pp. 1356-1358, 2009.

[166] M. Kimura, S. Inoue, T. Shimoda, and T. Eguchi, "Dependence of polycrystalline silicon thin-film transistor characteristics on the grain-boundary location," *J. Appl. Phys.*, vol. 89, pp. 596-600, 2001.

[167] P.-Y. Wang and B.-Y. Tsui, "A novel approach using discrete grain-boundary traps to study the variability of 3-D vertical-gate NAND flash memory cells," *IEEE Trans. Electron Devices*, vol. 62, pp. 2488-2493, 2015.

[168] H. Oh, J. Kim, J. Lee, T. Rim, C.-K. Baek, and J.-S. Lee, "Effects of single grain boundary and random interface traps on electrical variations of sub-30nm polysilicon nanowire structures," *Microelectron. Eng.*, vol. 149, pp. 113-116, 2016.

초 록

현재의 폰 노이만 구조의 계산 시스템은 나노크기의 전자소자에서의 심각한 누설 전류 문제를 겪고 있다. 신경계 모방 시스템은 생물학적 시스템을 모방함으로써 이러한 근본적인 문제점들을 해결할 수 있을 것으로 생각되어왔다. 시냅스 모방 소자는 신경계 모방 시스템에서 가장 중요한 부분으로 여겨지는데 생물학적 시냅스가 신호 전달과 기억 기능을 담당하고 있기 때문이다. 하지만 시냅스 모방 소자의 가장 강력한 후보인 멤리스터(memristor)는 하나의 전극으로 신호 전달 및 수신 기능을 모두 수행하여야 하기 때문에 스위치 요소를 필요로 하는데, 이는 신경계 모방 시스템의 가장 큰 장점인 병렬적 구성을 저해하는 추가 오버헤드가 된다.

본 논문에서는, 비대칭 듀얼게이트 구조의 실리콘 기반 시냅스 트랜지스터를 연구하였다. 이러한 구조적 특징은 시냅스 트랜지스터가 시냅스 전, 시냅스 후 뉴런 회로들과 직접적인 연결이 가능하도록 만든다. 소자 시뮬레이터와 회로 시뮬레이터를 이용하여 소자의 학습 특성과 원리를 근본적으로 확인하였다. 공정 시뮬레이터를 통해 모든 공정 순서를 규명한 후 시냅스 트랜지스터들은 2-단 CMP 공정을 포함한 핵심 공정 기술을 통해 제작되었다. 제작된 소자의 전기적, 시냅스적 특성들은 반도체 분석 장비를

통해 측정되었고 측정 결과를 바탕으로 소자 모델을 만들었다. 이 소자 모델을 이용하여 시냅스 트랜지스터들로 구성된 발화 신경망(spiking neural network)을 시스템적으로 검증하였다.

시뮬레이션 연구와 전기적 측정을 통해 단기기억에서 장기기억으로의 전환 및 뾰족 타이밍 의존 가소성(spike-timing dependent plasticity)을 포함한 시냅스 학습 규칙을 시냅스 트랜지스터를 통해 관측하였다. 또한, 시냅스 트랜지스터들로 구성된 발화 신경망의 패턴 인식 기능을 MNIST 데이터를 이용하여 확인하였다. 784 개의 입력 노드와 10 개의 출력 노드를 갖는 하드웨어 기반 신경망의 전체 인식률은 억제성 시냅스 부분을 추가함으로써 70%까지 향상되었다.

이러한 결과들은 본 논문에서 연구된 시냅스 트랜지스터가 신경계 모방 시스템 내에서 뉴런 회로들과의 직접 연결가능성 및 시냅스 학습 특성으로 인해 시냅스 소자로 활용될 수 있음을 나타낸다.

주요어 : 비대칭 듀얼게이트 구조, 발화 신경망, 시냅스 학습, 신경계 모방 시스템, 실리콘 기반 시냅스 모방 트랜지스터, 패턴 인식.

학 번 : 2012-30934