Ph.D. DISSERTATION

Synaptic Array Architectures Based on NAND Flash Cell Strings

낸드 플래시 셀 스트링 기반의 시냅틱 어레이 아키텍처

by

SUNG-TAE LEE

August 2021

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

Synaptic Array Architectures Based on NAND Flash Cell Strings

낸드 플래시 셀 스트링 기반의 시냅틱 어레이 아키텍처

지도교수 이 종 호

이 논문을 공학박사 학위논문으로 제출함

2021 년 8 월

서울대학교 대학원

전기정보공학부

이 성 태

이성태의 공학박사 학위논문을 인준함

2021 년 8 월

위 원 장 : 박 병 국

부위원장 : 이 종 호

위 원 : 유 승 주

위 원 : 김 재 준

위 원 : 김 재 하

Synaptic Array Architectures Based on NAND Flash Cell Strings

by

Sung-Tae Lee

Advisor: Jong-Ho Lee

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Electrical and Computer Engineering)

in Seoul National University

August 2021

Doctoral Committee:

Professor Byung-Gook Park, Chair

Professor Jong-Ho Lee, Vice-Chair

Professor Sungjoo Yoo

Professor Jae-Joon Kim

Professor Jaeha Kim

ABSTRACT

Neuromorphic computing using synaptic devices has been proposed to efficiently process vector-matrix multiplication (VMM) which is a significant task in DNN. Until now, resistive RAM (RRAM) was mainly used as synaptic devices for neuromorphic computing. However, a number of limitations still exist for RRAMs to implement a large-scale synaptic device array due to device nonideality such as variation, endurance and monolithic integration of RRAMs and CMOS peripheral circuits. Due to these problems, SRAM cells, which are mature silicon memory, have been proposed as synaptic devices. However, SRAM occupies large area (~150 F^2 per bitcell) and on-chip SRAM capacity (~a few MB) is insufficient to accommodate a large number of parameters.

In this dissertation, synaptic architectures based on NAND flash cell strings are proposed for off-chip learning and on-chip learning. A novel synaptic architecture based on NAND cell strings is proposed as a high-density synapse capable of XNOR operation for binary neural networks (BNNs) in off-chip learning.

By changing the threshold voltage of NAND flash cells and input voltages in complementary fashion, the XNOR operation is successfully demonstrated. The large on/off current ratio (~7×10^5) of NAND flash cells can implement high-density and highly reliable BNNs without error correction codes. We propose a novel synaptic architecture based on a NAND flash memory for highly robust and high-density quantized neural networks (QNN) with 4-bit weight. Quantization training can minimize the degradation of the inference accuracy compared to post-training quantization. The proposed operation scheme can implement QNN with higher inference accuracy compared to BNN.

On-chip learning can significantly reduce time and energy consumption during training, compensate the weight variation of synaptic devices, and can adapt to changing environment in real time. On-chip learning using the high-density advantage of NAND flash memory structure is of great significance. However, the conventional on-chip learning method used for RRAM array cannot be utilized when using NAND flash cells as synaptic devices because of the cell string structure of NAND flash memory. In this work, a novel synaptic array architecture enabling

forward propagation (FP) and backward propagation (BP) in the NAND flash memory is proposed for on-chip learning. In the proposed synaptic architecture, positive synaptic weight and negative synaptic weight are separated in different array to enable weights to be transposed correctly. In addition, source-lines (SL) are separated, which is different from conventional NAND flash memory, to enable both the FP and BP in the NAND flash memory. By applying input and error input to bit-lines (BL) and string-select lines (SSL) in NAND cell array, respectively, accurate vector-matrix multiplication is successfully performed in both FP and BP eliminating the effect of pass cells. The proposed on-chip learning system is much more robust to weight variation compared to the off-chip learning system. Finally, superiority of the proposed on-chip learning architecture is verified by circuit simulation of a neural network.

Keywords: hardware-based neural network, NAND flash memory, neuromorphic system, in-memory computing, binary neural network, on-chip learning.

Student number: 2016-20951

CONTENTS

List of Figures

Chapter 1
Introduction

1.1 Background

Recently, deep neural networks (DNNs) have achieved remarkable fulfillment for various intelligent tasks, such as speech recognition, computer vision, and natural language processing [1]-[3]. However, recent state-of-the-art DNNs demand a large neural network size and a huge volume of parameters, which need very fast graphic processing units (GPUs), enormous memory-storage and large computational power [4], [5]. In addition, the von Neumann bottleneck results in enormous energy and time consumption when performing VMM operations due to the large amount of moving data between the memory and processor. Neuromorphic systems have been actively investigated as a solution to the von Neumann bottleneck utilizing in-memory computing with a synaptic array architecture. When an input voltage is applied to a synaptic array, the multiplication of the input voltage and the conductance of the synaptic device gives the current, and the currents from

multiple synapses connected to one bit-line are summed up by Kirchhoff's current law (KCL). Each current in each bit-line in the array is summed simultaneously. Therefore, a synaptic device array can perform VMM in a single time step, which is orders of magnitude more efficient than the conventional von Neumann architecture [6].

Quantized neural networks (QNNs) significantly reduce the computing resources and memory storage by quantizing the weight and activation [7]-[17]. Instead of a high-precision floating-point weight and activation, they enable a low-bit weight in synaptic devices and low-bit activation in neuron circuits, providing a promising solution to the implementation of neuromorphic systems [8], [12]. In addition, recent studies have shown that QNNs could achieve a satisfying classification accuracy on representative image datasets, such as MNIST (mixed national institute of standards and technology), CIFAR-10 (Canadian institute for advanced research) and ImageNet [9]-[12].

In previous studies, RRAMs were commonly utilized as a synaptic device in neuromorphic systems [14], [15]. However, RRAMs require further research in

terms of parametric variability, stochastic programming, reliability, and integration of selectors for large-scale integration [16]. Moreover, the IR drop of a metal wire can result in an inaccurate VMM operation in an RRAM crossbar array [17]. In addition, the small on/off current ratio of RRAM causes an error in the sum of currents from many devices [18], [19].

Recent high performance DNN algorithms commonly require a vast parameter size and large network size. To accommodate immense parameters, NAND flash memory cells can be used as synaptic devices which have a huge advantage in cell density and an enormous storage capacity per chip. NAND flash memory technology has been well known as one of the most competitive solutions for immense data storage. In addition, NAND flash memory has been demonstrated as a technologically mature and cost-competitive technology among the various nonvolatile memory technologies [20]-[22]. However, it is hard to utilize NAND flash memory consisting of cell strings as synaptic architecture in neuromorphic computing systems due to the characteristics of the string structure.

A novel synaptic architecture based on NAND cell strings is proposed as a

high-density synapse capable of XNOR operation for binary neural networks (BNNs) for the first time. By changing the threshold voltage of NAND flash cells and input voltages in complementary fashion, the XNOR operation is successfully demonstrated. The large on/off current ratio (~7×10^5) of NAND flash cells can implement high-density and highly-reliable BNNs without error correction codes. It is shown that without conventional ISPP scheme, only 1 erase or program pulse can achieve sufficiently low bit-error rate. Finally, the estimated synapse density of VNAND memory with 128 stacks is ~100 times that of 2T2R synapse in RRAMs.

We propose a novel synaptic architecture based on a NAND flash memory for highly robust and high-density quantized neural networks (QNN) with 4-bit weight and binary neuron activation, for the first time. The proposed synaptic architecture is fully compatible with the conventional NAND flash memory architecture by adopting a differential sensing scheme and a binary neuron activation of (1, 0). A binary neuron enables using a 1-bit sense amplifier, which significantly reduces the burden of peripheral circuits and power consumption and enables bitwise communication between the layers of neural networks. Operating NAND cells in

the saturation region eliminates the effect of metal wire resistance and serial resistance of the NAND cells. With a read-verify-write (RVW) scheme, low-variance conductance distribution is demonstrated for 8 levels. Vector-matrix multiplication (VMM) of a 4-bit weight and binary activation can be accomplished by only one input pulse, eliminating the need of a multiplier and an additional logic operation. In addition, quantization training can minimize the degradation of the inference accuracy compared to post-training quantization. Finally, the low-variance conductance distribution of the NAND cells achieves a higher inference accuracy compared to that of resistive random access memory (RRAM) devices by 2~7 % and 0.04~0.23 % for CIFAR 10 and MNIST datasets, respectively.

A novel synaptic array architecture enabling forward propagation (FP) and backward propagation (BP) in the NAND flash memory is proposed for the first time for on-chip learning. In the proposed synaptic architecture, positive synaptic weight and negative synaptic weight are separated in different array to enable weights to be transposed correctly. In addition, source-lines (SL) are separated, which is different from conventional NAND flash memory, to enable both the FP

and BP in the NAND flash memory. By applying input and error input to bit-lines (BL) and string-select lines (SSL) in NAND cell array, respectively, accurate vector-matrix multiplication is successfully performed in both FP and BP eliminating the effect of pass cells. At a read voltage of 2 V, inference accuracy of 95.58 % which is comparable to that of 95.81 % obtained with perfect linear device is achieved. The proposed on-chip learning system is much more robust to weight variation compared to the off-chip learning system. Finally, superiority of the proposed on-chip learning architecture is verified by circuit simulation of a neural network.

Chapter 2
Binary neural networks based on NAND flash memory

2.1 Synaptic architecture for BNN

A novel 2T2S (two transistors and two NAND cell strings) synaptic string structure for XNOR operation is proposed in Fig. 2.1. Two NAND cell strings are used for one synapse string consisting of serially connected synaptic cells with two input transistors of which two input voltages are applied to each gate. The two input transistors can be replaced by two NAND cells having gates (word-lines) isolated from each other. For each synapse consisting of adjacent two NAND cells in two cell strings, synaptic weight of +1 can be represented by the two cells of which the left cell is on-state ($V_{th,low}$) and the right cell is off-state ($V_{th,high}$). For an input value, the input value of +1 can be represented by complementary input voltages where V_{in1} is turn-on voltage (V_{on}) and V_{in2} is turn-off voltage (V_{off}). Fig. 2.1 (a) and (b) represent the cases when the input value is +1 and -1. Fig. 2.2 (a) and (b) explain

the read operation scheme where a read bias (V_{read}) is applied to a selected word-line (WL) and a pass bias (V_{PASS}) is applied to unselected WLs. In the proposed scheme, the synaptic device in the k^{th} row of synaptic string in Fig. 2.2 (a) is the synapse connected to the k^{th} post-synaptic neuron of the BNN in Fig. 2.2 (c). The output for each post-synaptic neuron is generated when the read bias (V_{read}) is applied to WL sequentially along the synapse string as shown in Fig. 2.2 (b). In this scheme, the CSA sensing the current of synapse string is reused for all synapses in the synapse string, which reduces the burden of circuit and increases the integration density. In addition, the proposed 2T2S design is entirely digital, avoiding the need of large area operational amplifiers or analog-to-digital converters (ADC) which are needed in analogue vector-matrix multiplication. In Fig. 2.3, we propose block diagrams for peripheral circuits and a circuit diagram of a synapse array architecture consisting of 2T2S synapses. The encoded input vector is applied to the input vector switch matrix, and WL decoder applies a read bias (V_{read}) to a selected WL and applies V_{PASS} to unselected WLs. The adder sums the number of +1s in the XNOR operation outputs and the counted sum is passed to a digital comparator to generate

1-bit neuron output (+1 or -1). Therefore, for example, as read bias is applied to k^{th} WL, the output for k^{th} neuron in post-synaptic neuron layer is produced. Using this proposed operation scheme, the adder and the comparator are reused for all neurons in the neuron layer, thereby reducing the burden of CMOS circuits compared to the previous study [19]. In addition, energy consumption is significantly reduced because no multipliers are required.

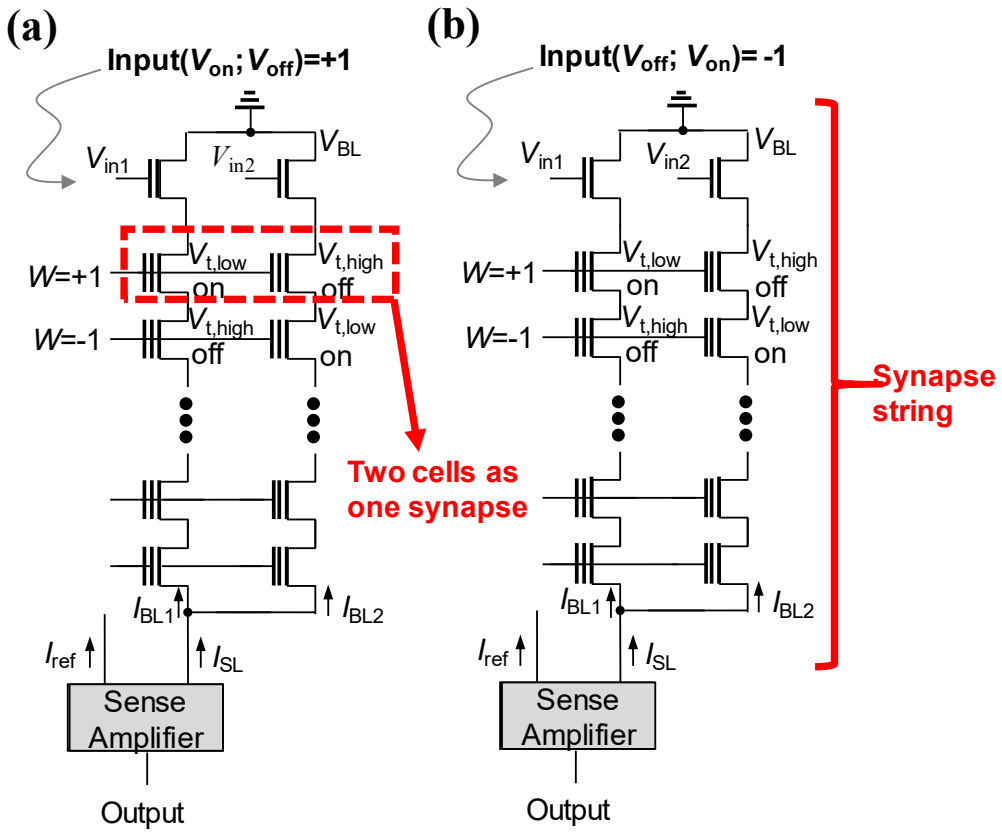

Fig. 2.1. 2T2S (two input transistors and two NAND strings) synapse string structure for XNOR operation.

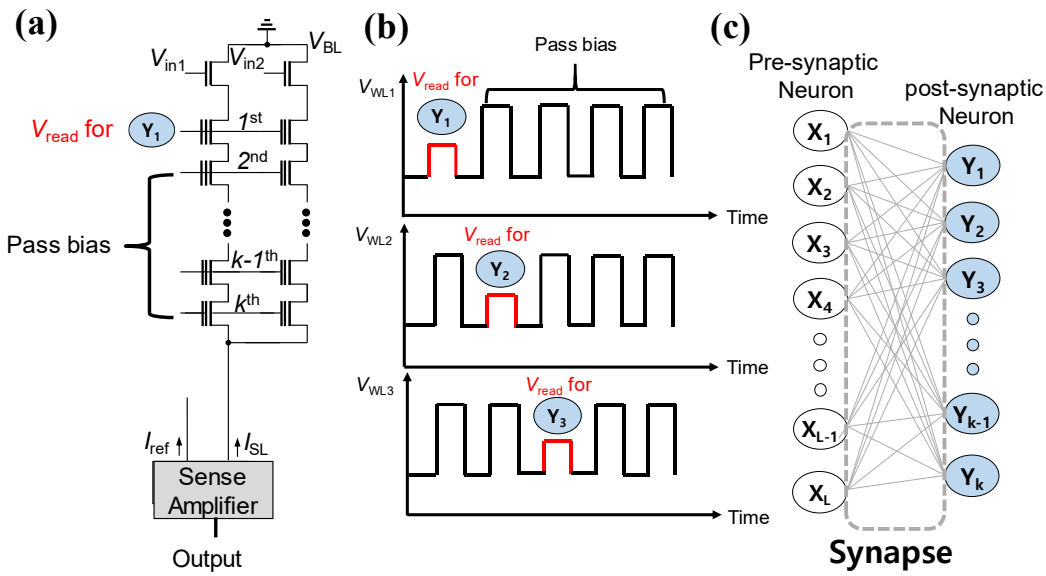

Fig. 2.2. (a) Read operation scheme for synaptic string. (b) Read operation scheme as a function of time. (c) Schematic diagram of binary neural networks.

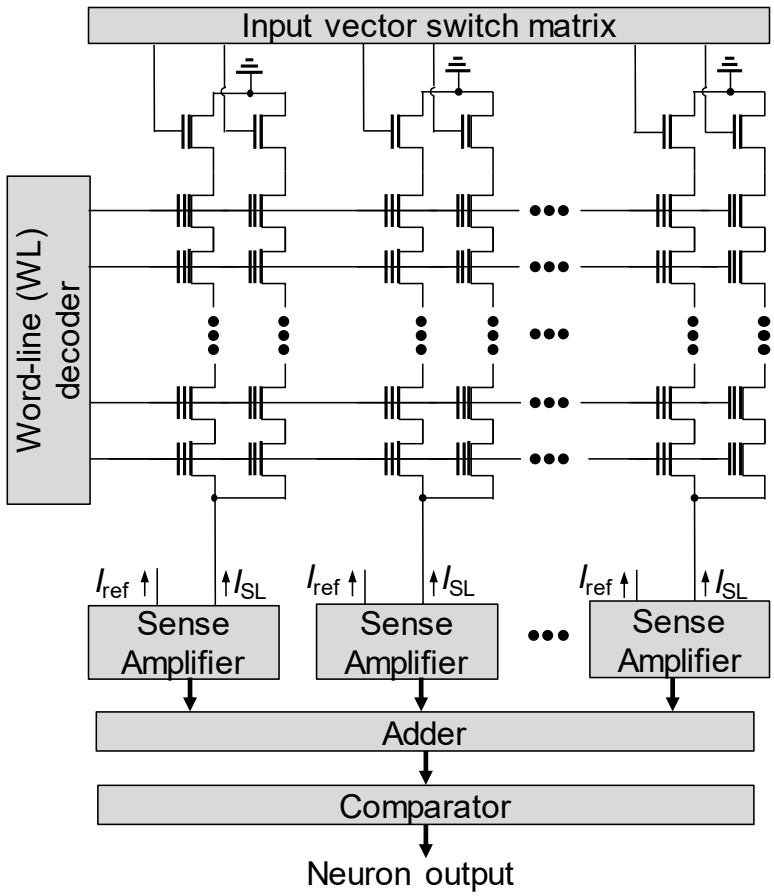

Fig. 2.3. Schematic diagram of 2T2S synaptic array architecture.

2.2 Measurement results

The 2-D NAND flash memory in this work was fabricated in the industry using 26 nm technology. One cell string consists of 64 cells and channel length and width are 26 and 20 nm, respectively. Fig. 2.4 (a) and (b) show the DC bit-line current (I_{BL}) –WL bias (V_{WL}) and I_{BL}-V_{BL} curves measured in the NAND cell as a parameter of programming voltage (V_{PGM}) when V_{PASS} is 6 V, respectively. As shown in Fig. 2.4 (b), the on/off current ratio which affects the bit-error rate and margin of the sense amplifier can be modulated by V_{PGM}. Fig. 2.5 (a), (b) and (c) show DC I_{BL}-V_{WL} curves measured from the NAND cells in the array as a parameter of V_{PGM}. Fig. 2.6 shows a cumulative distribution of XNOR output states measured in two adjacent NAND flash cells as a parameter of V_{PGM}. Even at 14 V V_{PGM}, the +1 and -1 states differ by 4 orders of magnitude. Fig. 2.7 (a) and (b) show normalized counts and Gaussian fittings of the XNOR outputs of +1 and -1 states measured in NAND flash cells. In digital BNN systems, the bit-error rate is an important factor as it affects the inference accuracy. The estimated bit error rates of NAND cells are about 4.2×10^{-8} % and 2.3×10^{-7} % when the V_{PGM} is 16 V and 14 V, respectively, and

the estimation is based on the statistical parameters from the measurement data and the assumption of Gaussian distribution as shown in Fig. 2.7. The estimated bit-error rate of NAND cells is sufficiently low compared to that of RRAM devices. Fig. 2.8 shows the program and erase windows when the P/E (program/erase) cycle is repeated up to 3×10^3. The $\Delta V_{th,PGM}$ (ΔV_{th} by V_{PGM}) increases and $\Delta V_{th,ERS}$ (ΔV_{th} by V_{ERS}) decreases as the number of P/E cycles increases when the V_{PGM}s are 16V and 18 V. Since the decrement of $\Delta V_{th,ERS}$ is larger than the increment of $\Delta V_{th,PGM}$, therefore, the memory window decreases with increasing P/E cycles. But the V_{PGM} of 14 V has no effect on the windows. *SS* increases by increasing the number of P/E cycles only when the V_{PGM} is 18 V as shown in Fig 2.9. Fig. 2.10 shows the I_{BL} behavior when only one P or E pulse is applied to write the weight during P/E cycles under different P/E conditions. The on-current (I_{on}) decreases due to the increase of V_{th} and *SS* as the number of P/E cycles increases only when V_{PGM} is 18 V. When V_{PGM} is below 16V, highly reliable BNN can be implemented regardless of P/E cycling number without using conventional ISPP method. Therefore, the time and energy can be greatly reduced compared to those of the ISPP method. If the ISPP

method is used, it is confirmed that the I_{BL} is almost the same regardless of the number of P/E cycles even when V_{PGM} is 18V. Fig. 2.11 shows the retention characteristic of fresh cell and 3×10^3 cycled cell. The I_{on} and I_{off} of the 3×10^3 cycled cell do not change until 10^4 s. The I_{BL} increases as the V_{PASS} increases from 3 V to 8 V as shown in Fig. 2.12. Unlike in NAND flash memory operation, a V_{PASS} of 6 V is also appropriate because a lower I_{BL} is allowed in the BNN. Fig. 2.13 (a) and (b) show *SS* and I_{BL} variations with the V_{PASS} stress, respectively. SS and I_{BL} exhibit ignorable variations regardless of the number of V_{PASS} stress (6 V) as shown in Fig. 2.13. Thanks to lower V_{PASS} than that of NAND flash memory, BNN is more robust to V_{PASS} disturbance.

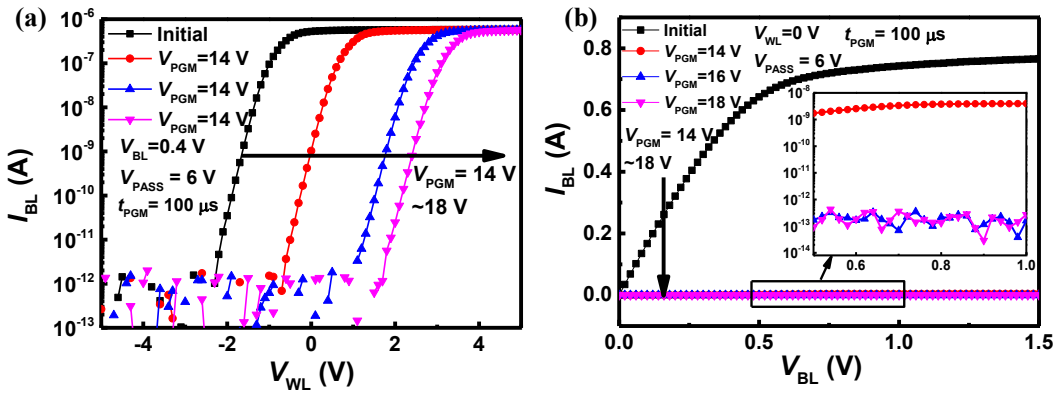

Fig. 2.4. (a) I_{BL}-V_{WL} curves as a parameter of V_{PGM} at V_{BL}=0.4 V, V_{PASS}=6 V and t_{PGM}=100 µs. (b) I_{BL}-V_{BL} curves as a parameter of V_{PGM} at V_{WL} = 0 V, V_{PASS}=6 V and t_{PGM}=100 µs. On/off current ratio which affects bit-error rate and margin of the sense amplifier can be modulated by programming voltage.

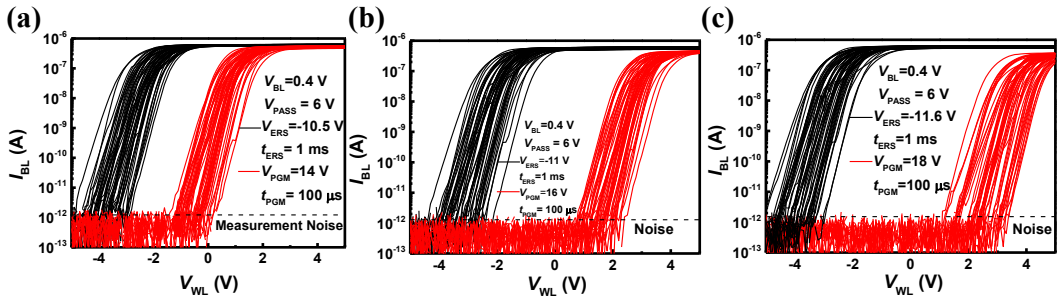

Fig. 2.5. (a) I_{BL}-V_{WL} curves of NAND flash memory cells in an array when V_{PGM} =14 V and V_{ERS}= -10.5 V. (b) The results when V_{PGM} =16 V and V_{ERS}= -11 V. (c) The results when V_{PGM} =18 V and V_{ERS}= -11.6 V. t_{PGM} and t_{ERS} are fixed at 100 µs and 1 ms, respectively.

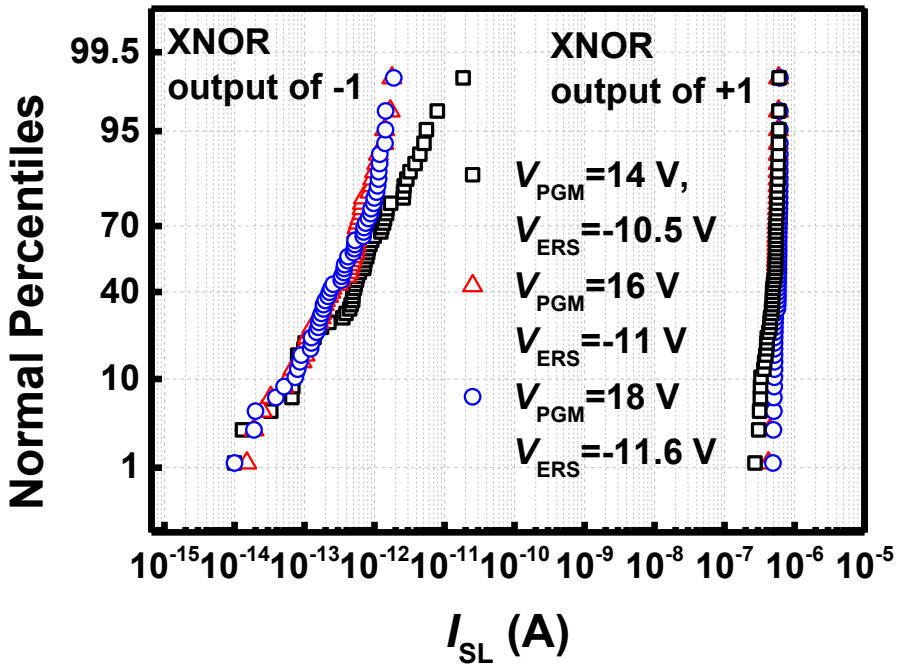

Fig. 2.6. Cumulative distribution of XNOR output of -1 and +1 as a parameter of different V_{PGM} and V_{ERS} at fixed t_{PGM}=100 μs and t_{ERS}=1 ms.

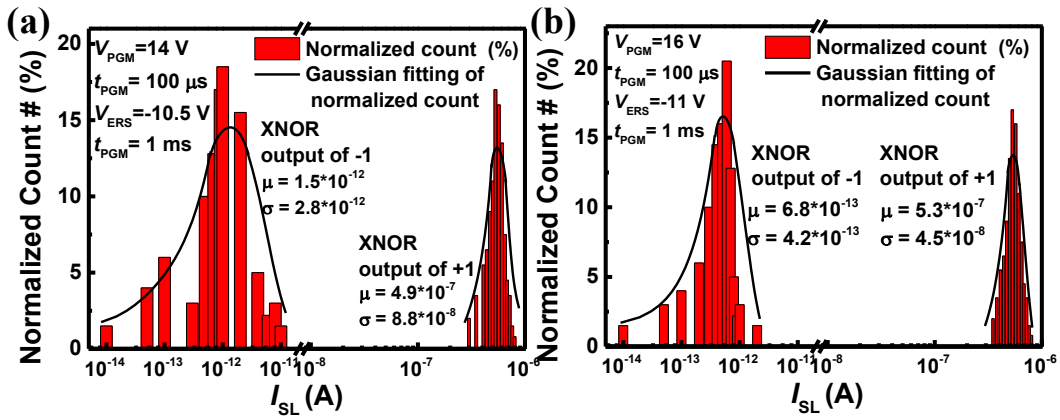

Fig. 2.7. Normalized counts of XNOR outputs of +1 and -1 measured in NAND flash cells and their Gaussian fitting at (a) V_{PGM}=14 V and V_{ERS}=-10.5 V. (b) V_{PGM}=16 V and V_{ERS}=-11 V. t_{PGM} and t_{ERS} are fixed at 100 μs and 1 ms, respectively.

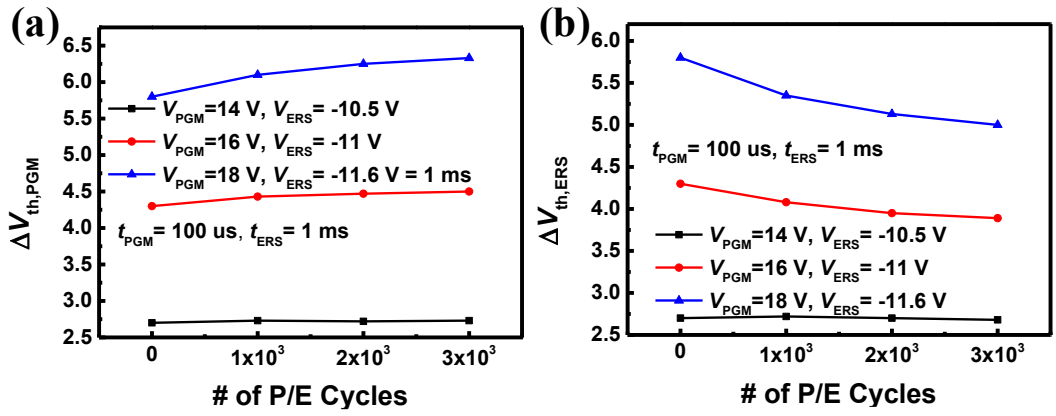

Fig. 2.8. Program and erase window ($\Delta V_{th,PGM}$, $\Delta V_{th,ERS}$) when P/E (program/erase) cycles are repeated until 3×10^3. The $\Delta V_{th,PGM}$ (ΔV_{th} by V_{PGM}) increases and $\Delta V_{th,ERS}$ (ΔV_{th} by V_{ERS}) decreases as the number of P/E cycles increases when the V_{PGM} is 16V and 18 V. $\Delta V_{th,PGM}$ and $\Delta V_{th,ERS}$ do not change with the cycles at V_{PGM}=14 V.

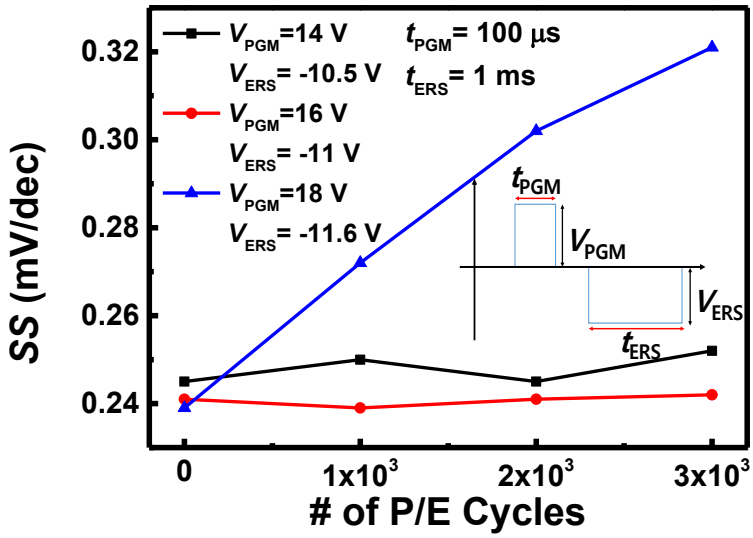

Fig. 2.9. *SS* change with P/E cycling for various P/E conditions. *SS* increases when the V_{PGM} is 18 V, but does not increase when the V_{PGM} is 16 V or less.

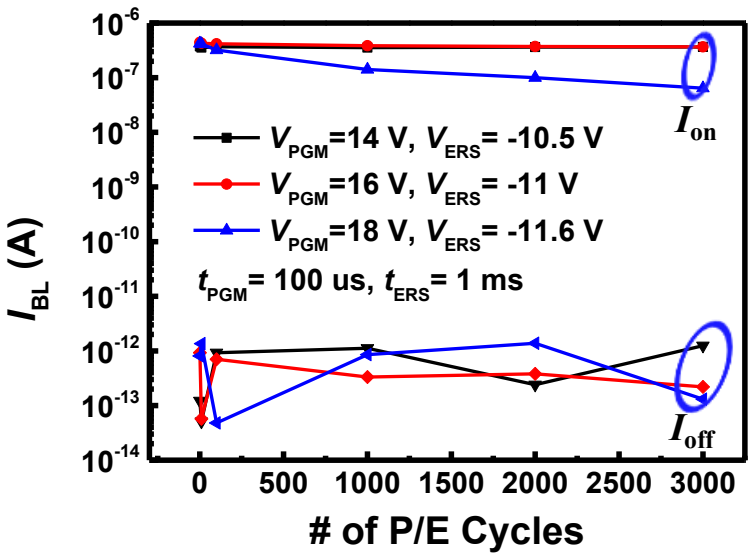

Fig. 2.10. Bit-line current (I_{BL}) with P/E cycles under three P/E conditions. On-current decreases only when V_{PGM} is 18 V.

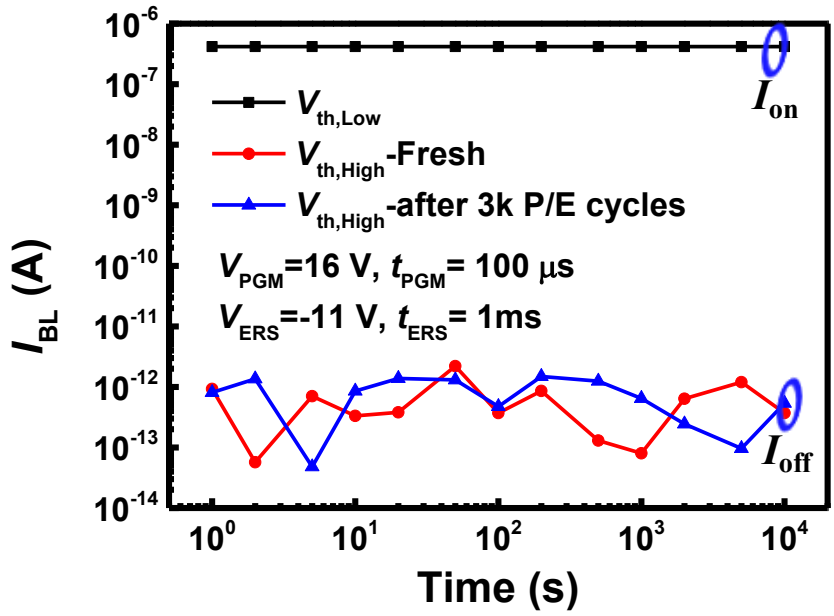

Fig. 2.11. Retention characteristic of fresh and 3×10^3 cycled cells. Off-current of 3×10^3 cycled cell does not increase until 10^4 s.

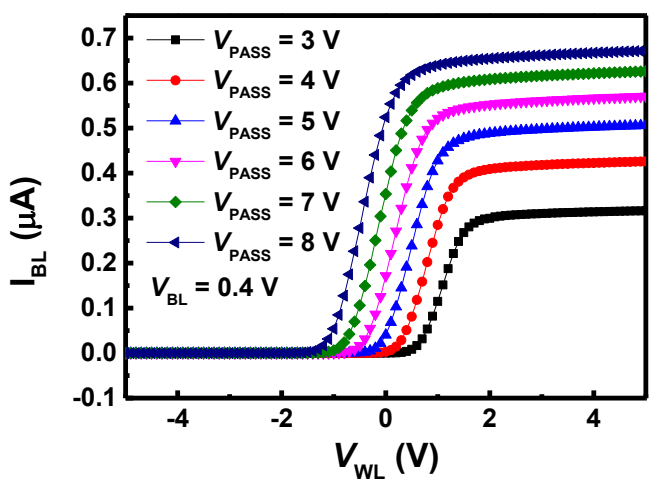

Fig. 2.12. Bit-line current (I_{BL}) with word-line bias (V_{WL}) as a parameter of V_{PASS} at a fixed V_{BL} of 0.4 V. I_{BL} increases as the V_{PASS} increases.

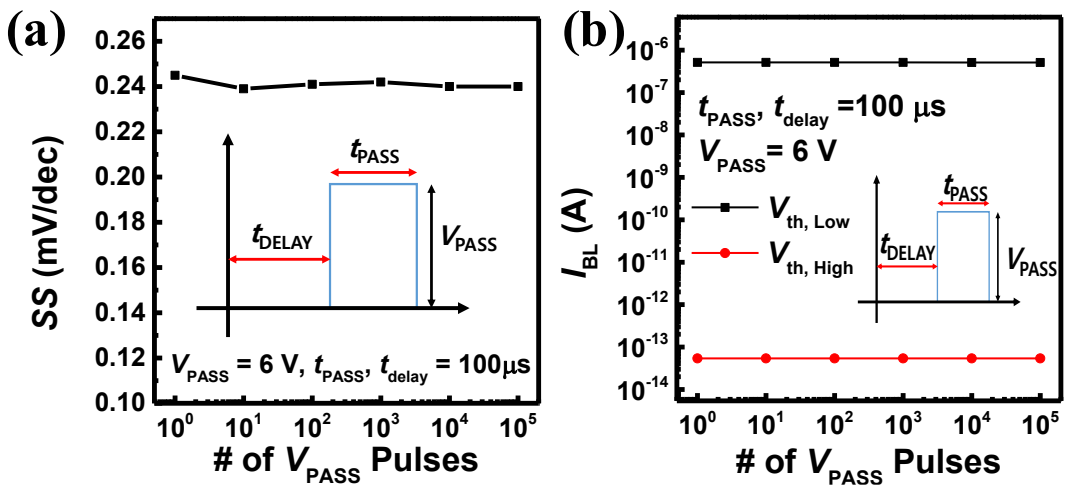

Fig. 2.13. (a) *SS* and (b) Bit-line current (I_{BL}) with the V_{PASS} stress. I_{BL} and SS exhibits negligible variations regardless of the number of V_{PASS}. Thanks to lower V_{PASS} than that of conventional NAND flash memory, BNN is more robust to V_{PASS} disturbance.

2.3 Binary neuron circuit

Fig. 2.14 shows a current-latch based CSA circuit [23] for BNNs. The circuit is simulated using a 20-nm FinFET based BSIM-CMG model [24] and an I_{SL} (composed of I_{BL}s) measured in this work. Fig. 2.15 (a) and (b) show transient waveforms for the XNOR output of +1 and -1, respectively, when the I_{on} and I_{off} of NAND cell are 590 nA and 0.1 pA. When the XNOR output is +1, the CSA senses the I_{on} of the NAND cell and the read access time is 2 ns. On the other hand, when the XNOR output is -1, the CSA senses the I_{REF} and the read access time is 12ns. XNOR operation is successfully demonstrated in 2T2S synapse based on two adjacent NAND cell strings as shown in Fig. 2.16. Fig. 2.16 (a) shows the circuits for measuring I_{SL}s for weights of +1 and -1. Fig. 2.16 (b) and (c) show the I_{SL} waveforms measured in synaptic strings with weights of +1 and -1 respectively. Note that the time scale is long because input and read pulses are supplied from the HP4145A.

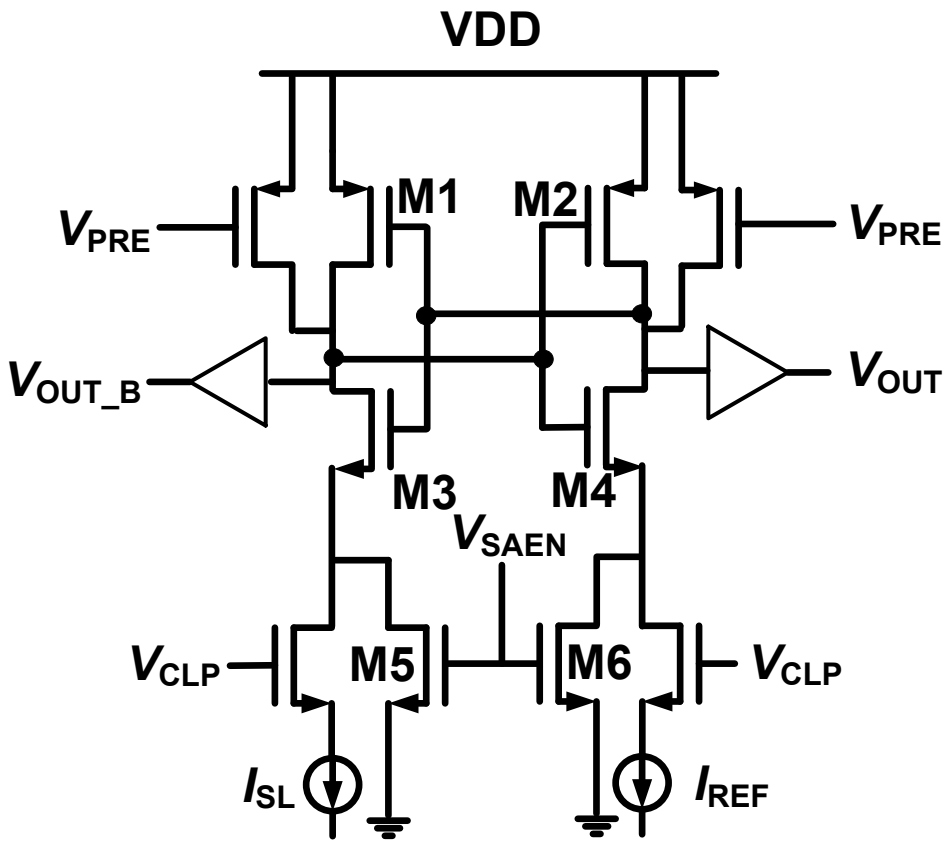

Fig. 2.14. A current-latch based CSA (current sense amp.) in BNNs.

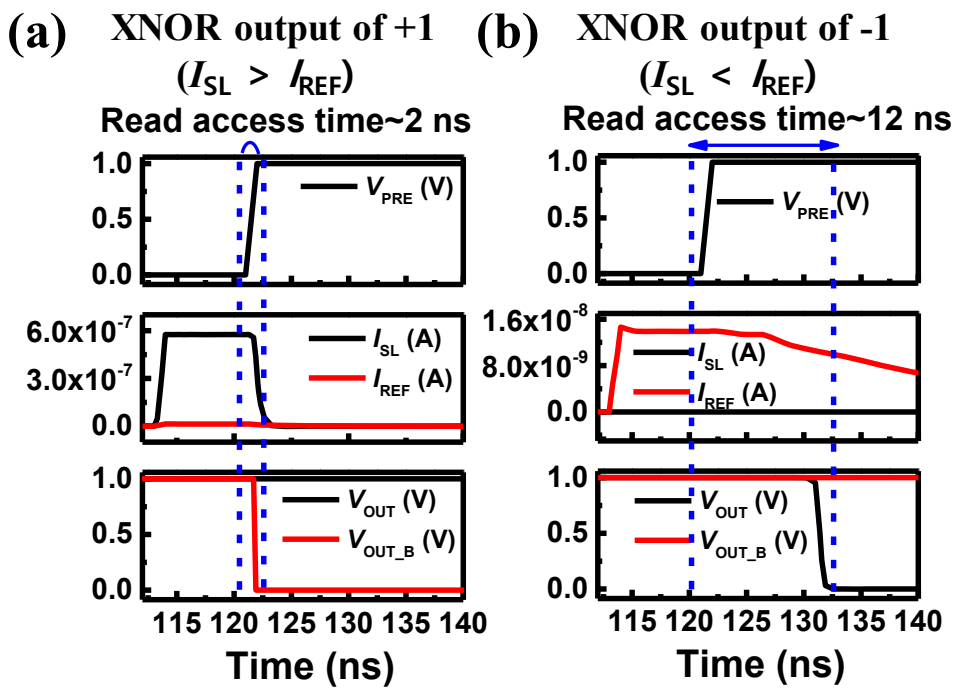

Fig. 2.15. Simulated transient results of CSA when XNOR output of (a) +1 (I_{SL}>I_{REF}) and (b) -1 (I_{SL}<I_{REF}). The I_{on} and I_{off} of the synapse are 590 nA and 0.1 pA, respectively.

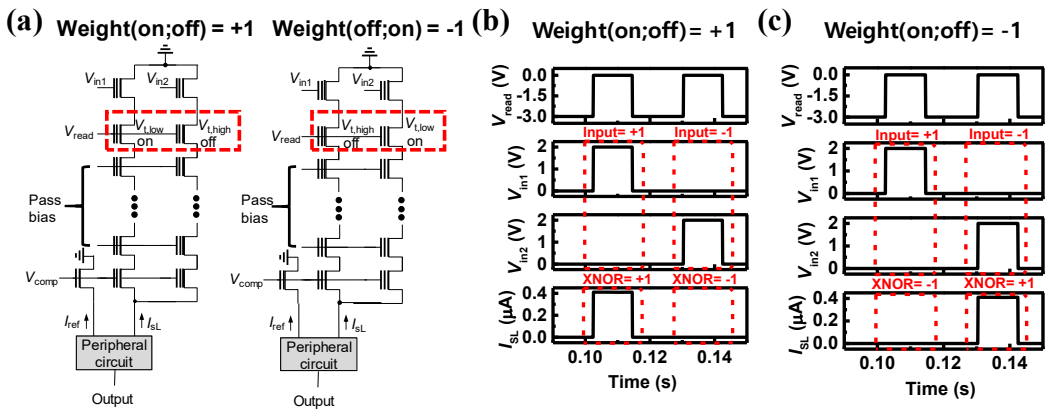

Fig. 2.16. Demonstration of XNOR operation in synapse based on NAND flash cell strings. (a) Circuits for measuring the I_{SL}s for weights of +1 and -1. Measured I_{SL}s which are the results of XNOR operation when the weight is (b) +1 and (c) -1. XNOR operation has been successfully demonstrated in 2T2S Synapse based on NAND flash cell strings.

2.4 Simulation results

Fig. 2.17 shows simulated inference accuracies for MNIST and CIFAR 10 patterns with the number of epochs. Fig. 2.18 compares the inference accuracy versus bit error rate for the proposed NAND flash memory synapses and reported RRAM synapses. The bit-error rate of our work is much lower than those of RRAM synapses, while keeping much higher synapse density. Our work provides highly-reliable BNNs that do not require error correction codes (ECC), which can reduce the enormous time, energy and complex decoding circuitry required for the ECC. Although cell density is high in 2-D NAND flash, cell density can be much higher in vertical NAND (VNAND) flash than in 2-D case. Fig. 2.19 shows the effective area per VNAND synapse and synapse density ratio of the VNAND synapse to 2T2R synapse with increasing number of stacks. Here, control is the area of one 2T2R-based synapse in RRAMs. The area occupied by the 2T2R synapse is calculated to be 24300 nm^2 by assuming that two 22nm FinFETs under two RRAMs determine the area of one synapse. As the number of stacks increases, the effective area of one synapse in VNAND memory becomes smaller. The synapse density of

VNAND at a stack number of 128 is about ~100 times higher than that of control in RRAMs.

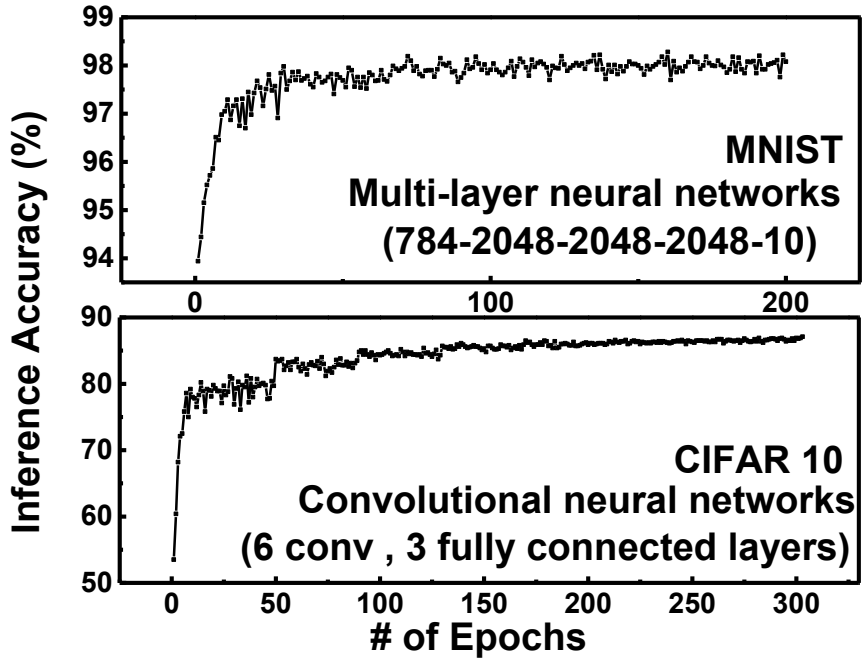

Fig. 2.17. Simulated inference accuracies for MNIST and CIFAR 10 patterns with the number of epochs. The final accuracies for both patterns are 98.12% and 87.11%. MNIST and CIFAR10 are trained on binarized multi-layer and convolutional neural networks, respectively.

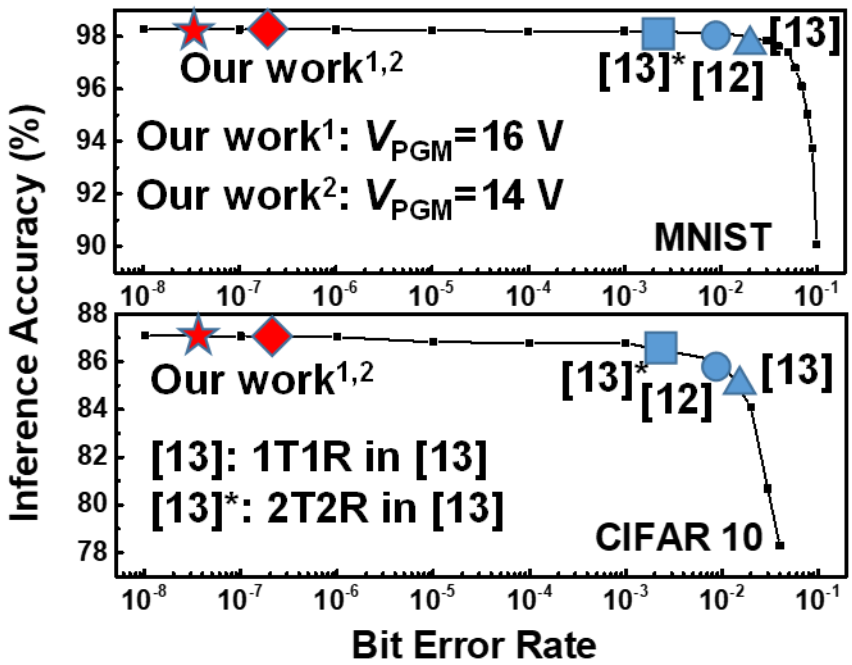

Fig. 2.18. Inference accuracy with bit error rate. Our work[1,2] indicate the cases when V_{PGM}s are 16 V and 14 V, respectively. Our work[1,2] have much lower bit-error rate than RRAMs.

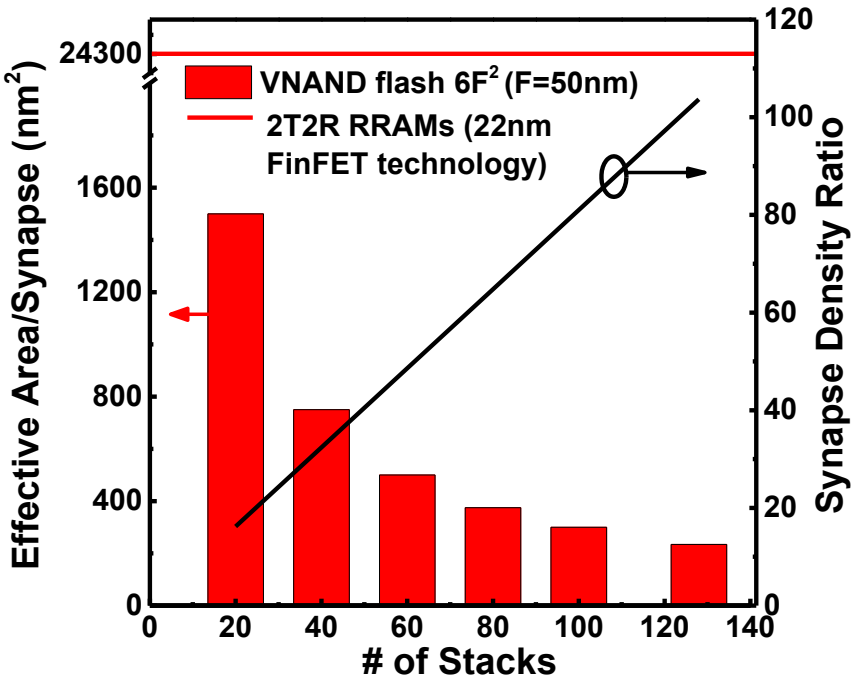

Fig. 2.19. Effective area per synapse and synapse density ratio with the number of stacks. Here, control is the area of one synapse in RRAMs. The synapse density at a stack number of 128 is about 100 times higher than that of control in RRAMs.

2.5 Differential scheme

2.5.1 Differential synaptic architecture

A novel 4T2S (four transistors and two NAND strings) synaptic string structure for XNOR operations and a differential sensing scheme merged with the NAND string are proposed in Fig.2.20. Two NAND strings are used for one synaptic string consisting of serially connected synaptic cells with four input transistors of which two input voltages are applied to each gate. In fact, the four input transistors are merged with one sense amplifier, which is simpler than the synaptic string and sense amplifier in Fig. 2.1. Fig. 2.20 (a) shows the synapse string connected to a sense amplifier. As shown in Fig. 2.20 (b), the differential current sense amplifier consists of two precharge PMOSFETs, a cross-coupled inverter pair, and four input transistors. The four input transistors, of which V_{in1} and V_{in2} are applied to each gate, are reused for all synapses in one synaptic string, thereby reducing the number of input transistors compared to the scheme in a previous study [19]. The differential current sense amplifier compares the two bit-line currents (I_{BL1}, I_{BL2}) of the two NAND flash cells to generate an *XNOR* output. Due to its

differential structure, this scheme has an intrinsically reduced bit-error rate by ~5.5 times than that using a fixed reference current in Fig. 2.1. In addition, the differential scheme does not need circuits for a fixed reference current source which is needed in the scheme shown in Fig. 2.1. To take advantage of the lower bit-error rate and avoid using circuitry for a fixed reference current, we only consider the differential sensing scheme in this work from now on. In addition, it is also possible to extend the functionality of the sense amplifier to contribute to performing the logic operation, and thus to reduce the CMOS overhead. The differential current sense amplifier reads the current of the NAND flash cells, and at the same time, performs an XNOR operation.

Fig. 2.21 shows the operation rule of the XNOR implementation. For each synapse consisting of two adjacent NAND cells, a synaptic weight of +1 can be represented by the two cells of which the left cell is the on-state ($V_{t,low}$) and the right cell is the off-state ($V_{t,high}$). In contrast, a synaptic value of -1 can be represented by the reverse pattern of the states of the two NAND cells. For an input value, the input value of +1 can be represented by complementary input voltages where V_{in1} is the

turn-on voltage (V_{on}) and V_{in2} is the turn-off voltage (V_{off}). In contrast, the input value of -1 can be represented by the reverse pattern of the two input voltages. By using the above scheme, the *XNOR* output of the current sense amplifier is determined by the combination of the complementary input voltages and the states of the two adjacent NAND flash cells. Fig. 2.21 (a) represents the case when the input value is +1. In this case, because I_{BL1} is larger than I_{BL2} for a weight of +1, the voltage at node Q_B drops to the trip-point voltage faster than node Q, raising node Q toward VDD. As a result, *XNOR* remains at VDD while *XNOR_B* drops to zero. In contrast, because I_{BL2} is larger than I_{BL1} for a weight of -1, the voltage at node Q drops to the trip-point voltage faster than node Q_B, raising node Q_B toward VDD. As a result, *XNOR_B* remains at VDD while *XNOR* drops to zero. Fig. 2.21 (b) represents the case when the input value is -1.

Fig. 2.22 shows a sequential operation scheme and a proposed parallel operation scheme for a NAND flash-based synaptic architecture. Fig. 2.22. (a) shows a schematic diagram of a binary neural network consisting of two neuron layers as an example. Fig. 2.22. (b) shows a schematic diagram of the sequential

operation scheme for a synaptic architecture. Synaptic devices in the k^{th} ($1 \leq k \leq L$, L: the number of neuron) row of the synaptic array in Fig. 2.22 (b) represent synapses connected to the k^{th} neuron in the 1^{st} neuron layer in Fig. 2.22 (a). Therefore, the output for each neuron (Output 1 ~ Output L) in the 1^{st} neuron layer is generated sequentially when the read bias (V_{read}) is applied to the word-line sequentially along the synaptic string as shown in Fig. 2.22 (b). The adder sums the number of +1s in the XNOR operation outputs and the counted sum is passed to a digital comparator to generate the 1-bit neuron output (+1 or -1). Using the sequential operation scheme, the adder and the comparator are reused for all synaptic devices in the synaptic string, thereby reducing the burden of the CMOS circuits compared to the sequential read scheme in a previous study [19].

On the other hand, Fig. 2.22. (c) shows a schematic diagram of the proposed parallel read operation for a synaptic architecture. For the RRAM-based synaptic devices, the currents of synaptic devices are summed through the bit-line, which easily enables parallel read operation [19]. However, in NAND flash memory structure, currents of NAND cells connected in the same cell string cannot be

summed due to the string structure of NAND flash memory. To enable parallel operation in NAND flash cell array, we propose a new architecture and its read operation scheme. Synaptic devices in the 1^{st} row of the synaptic array in Fig. 2.22 (c) represent all of the synapses connected between the 0^{th} neuron layer and 1^{st} neuron layer in Fig. 2.22 (a). The input vector from the 0^{th} neuron layer is divided from input 1 to input M and it is passed to the input-vector (IV) switch matrix. A read bias is applied to the 1^{st} row of all synaptic arrays, and the pass bias is applied to the unselected rows of all synaptic arrays. Then, the adder counts the number of +1s in the XNOR operation outputs and the counted sum is passed to a digital comparator to generate the 1-bit neuron output (+1 or -1) for the 1^{st} neuron layer. By using the parallel read operation, all of the neuron outputs (output 1~ output L) of the 1^{st} neuron layer are produced at the same time as shown in Fig. 2.22 (c). Therefore, the proposed parallel read scheme resolves the difficulty of applying NAND flash to neural networks due to the specificity of the string structure. It significantly reduces the read-out latency compared to the sequential read scheme in Fig. 2.22 (b). As an example, let's consider a case where there are N neuron layers

in a neural network and L neurons on average in each neuron layer. When t_{read} is the time taken to obtain an output by applying a read voltage to a word-line in a synaptic string, the processing time of one image in the sequential read scheme in Fig. 2.22 (b) is $t_{read} \times N \times L$. On the other hand, when the parallel read scheme in Fig. 2.22 (c) is applied, the time to process one image decreases to $t_{read} \times N$. Estimating t_{read} requires further research, including measurement and circuit simulation of NAND flash memory, which will be further explored in the future. In addition, registers, which are required in the sequential read scheme to store sequentially generated neuron outputs, are not needed in this parallel read scheme. Therefore, we can reduce the required hardware resources compared to the sequential read scheme [19].

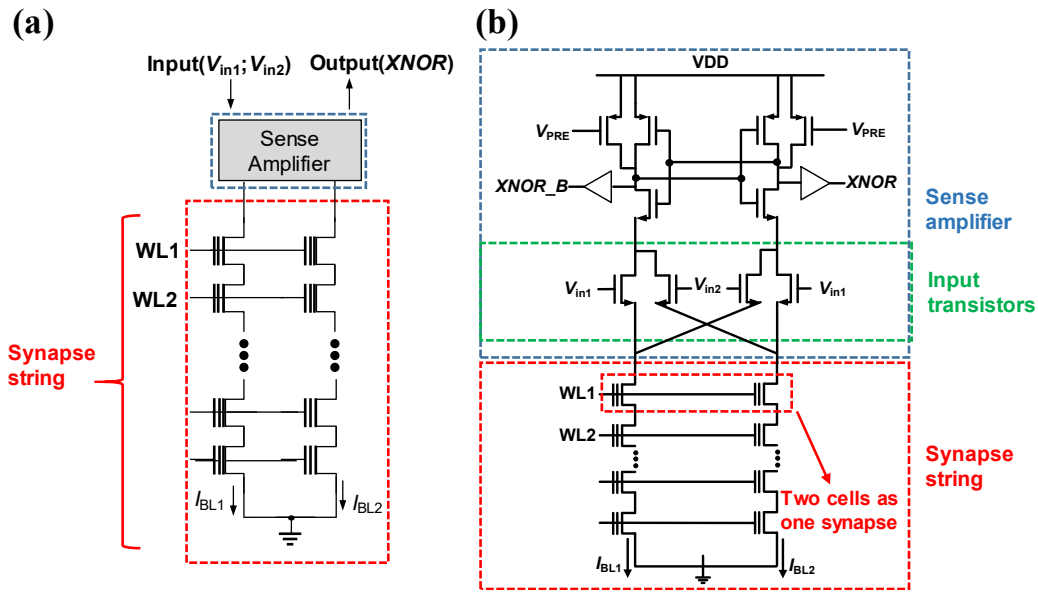

Fig. 2.20. 4T2S (four input transistors and two NAND strings) synaptic string structure with a sense amplifier for a differential sensing scheme. Four input transistors are merged with the sense amplifier.

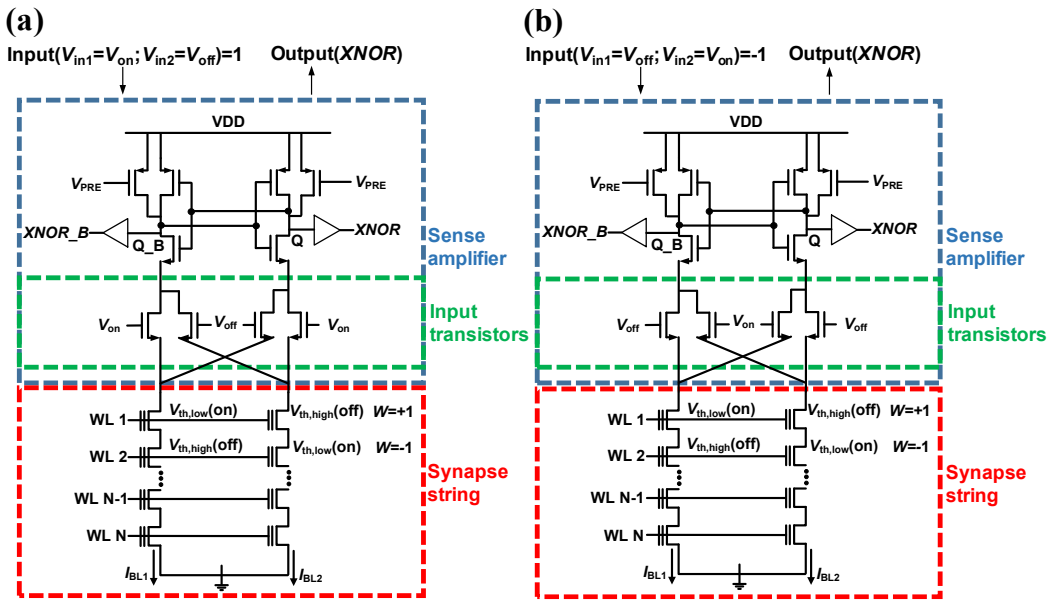

Fig. 2.21. Operation rule of the XNOR implementation. (a) The case when the input value is +1 (b) The case when the input value is -1.

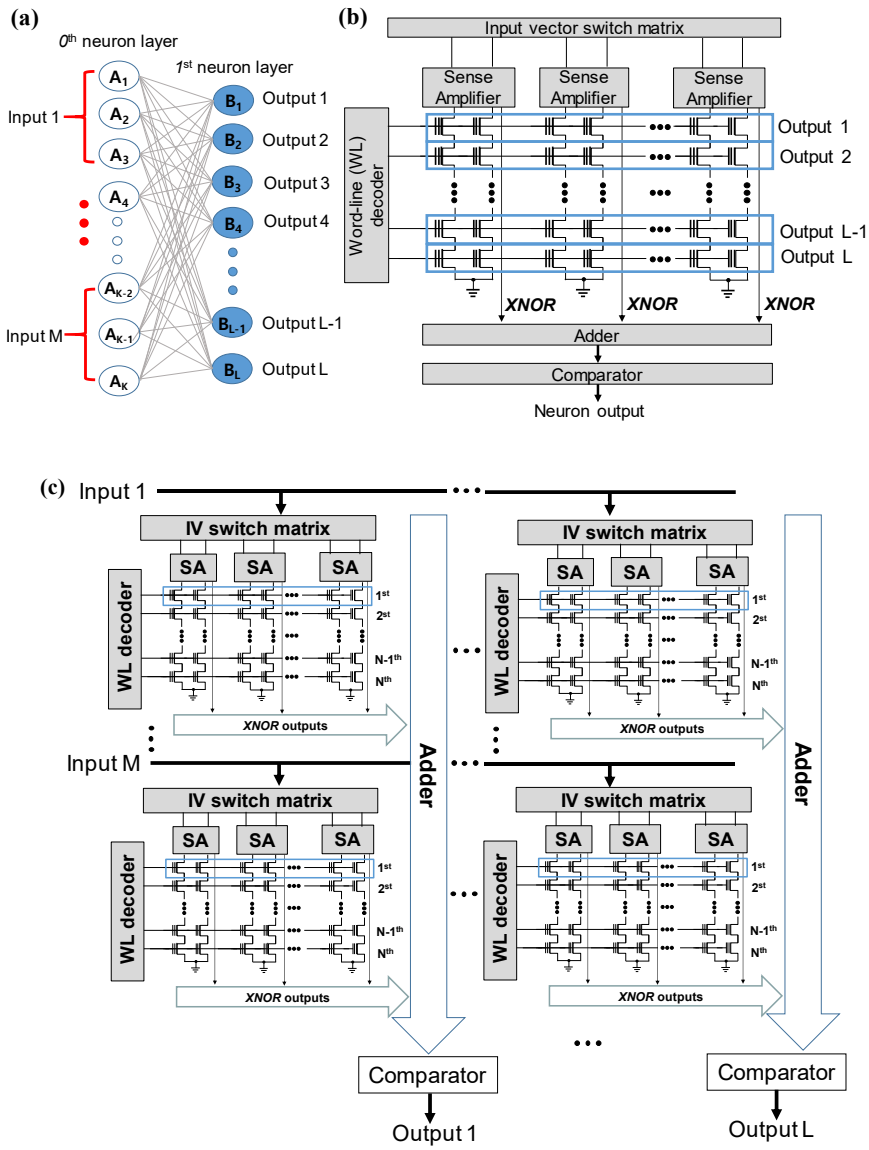

Fig. 2.22. (a) Schematic diagram of a neural network consisting of two neuron layers. (b) Schematic diagram of the sequential read operation. (c) Schematic diagram of the parallel read operation.

2.5.2 Simulation results

Fig. 2.23 shows the simulated inference accuracy for MNIST with the number of epochs as a parameter of the size of the binary multi-layer neural networks. As shown in Fig. 2.23, the inference accuracy increases as the depth and width of the binary neural networks increase. Therefore, an enormous number of synaptic devices are needed to implement binary neural networks with a high inference accuracy. As a way to accommodate this, NAND flash memory, which has a great advantage in cell density and a large storage capacity per chip, appears to be very promising in terms of providing dense and reliable synaptic devices.

Fig. 2.24 shows the simulated inference accuracy for the MNIST and CIFAR 10 patterns with the number of epochs. Note that, the proposed synaptic architecture in this work utilizes an off-chip learning scheme. In off-chip learning scheme, trained weights obtained using programming language are transferred to synaptic devices. In this process, various flash transition layer (FTL) designs [25]-[27] can be used to effectively find its physical address in the NAND flash memory based on a logical address in a system file. The MNIST and CIFAR10 patterns are trained

on binarized multi-layer and convolutional neural networks, respectively. The final accuracy of both patterns is 98.16% and 87.2%, respectively. In addition, the effect of bit-error rate on the inference accuracy is investigated. Fig. 16 compares the inference accuracy versus the bit error rate using the proposed NAND flash memory synapses and the reported RRAM synapses [9], [14] for MNIST and CIFAR10 patterns. The weights obtained by off-chip learning in Fig. 15 are transferred to the synaptic array. This transfer is done by first erasing all cells with one erase pulse and applying one pulse for programming to selected cells of which V_{th} needs to be increased. Bit-errors can occur during this transfer process. To evaluate the effect of the bit error rate, different bit error rates are applied to the simulation of the neural networks. Black squares in Fig. 16 represent inference accuracy with respect to bit error rates. In Fig. 16, our work[1] has a bit error rate of 4.2×10^{-8} % when the sensing scheme using the fixed reference current shown in Fig. 1 is applied. Our work[2] further reduces the bit error rate to 7.6×10^{-9} % by using the differential sensing scheme shown in Fig. 2. The bit-error rates of our work[1,2] are much lower than those of the RRAM synapses in [12], [13], while keeping a much higher

synapse density. For MNIST dataset, the inference accuracy with our work[2] is 98.16 % while the inference accuracies with RRAM synapses ([13]*, [12], [13]) are 98.04, 97.95 and 97.87 %. For CIFAR 10 dataset, the inference accuracy with our work[2] is 87.2 % while the inference accuracies with RRAM synapses ([13]*, [12], [13]) are 86.78, 86.02 and 85.2 %. Error correction codes (ECC) are not needed because the bit-error rate is sufficiently low as shown in Fig. 16. ECC decoding requires much more CMOS overhead. It demands logic circuits to detect whether an error occurred and complex circuits to detect the location of the error to revise it, requiring quite a large number of logic gates. Therefore, it puts a heavy burden of circuits on the neuromorphic system, because ECC decoders would need to be duplicated for each memory array in the system. By contrast, our method only uses sense amplifier circuits that has no additional complexity compared to the current sense amplifier used in conventional NAND flash memory. Therefore, our work provides highly reliable BNNs that do not require ECC, which can reduce the enormous time, energy and complex decoding circuitry required for the ECC.

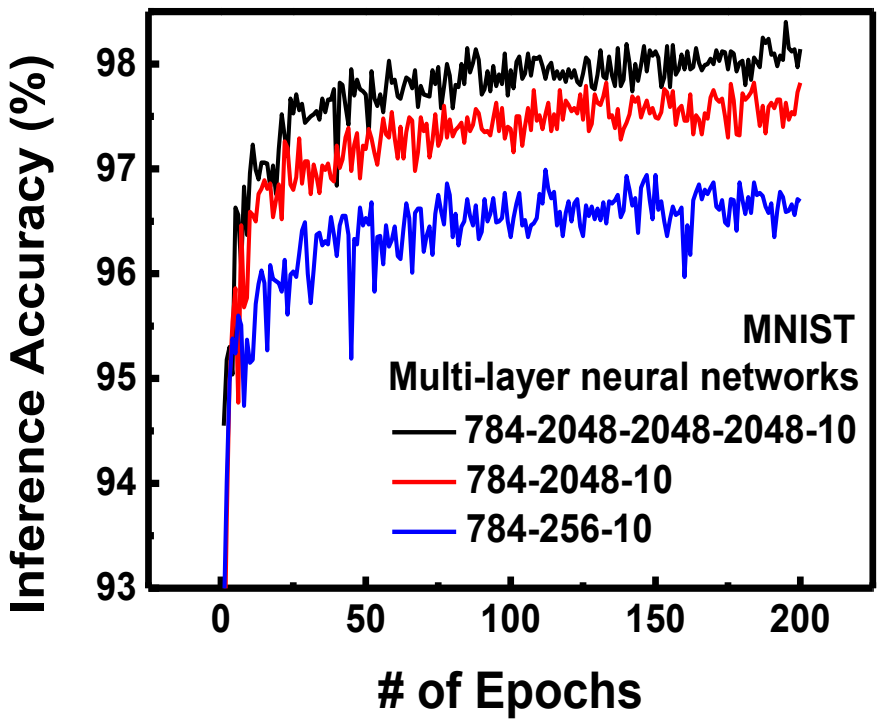

Fig. 2.23. Simulated inference accuracy for MNIST patterns with the number of epochs as a parameter of the size of the neural networks. MNIST patterns are trained on binarized multi-layer neural networks.

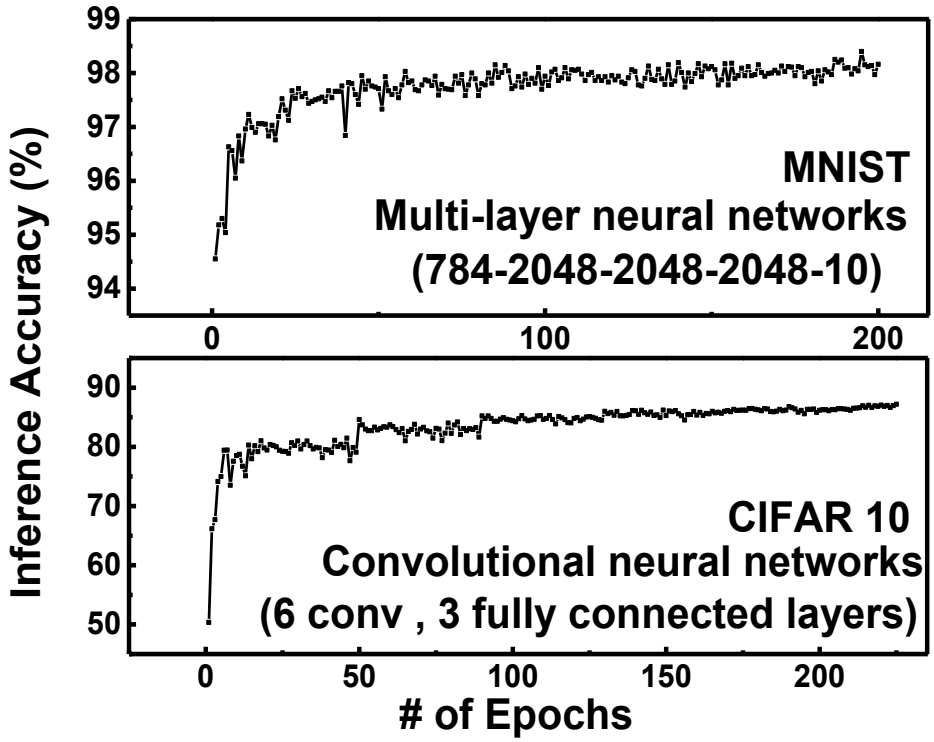

Fig. 2.24. Simulated inference accuracy for the MNIST and CIFAR 10 patterns with the number of epochs.

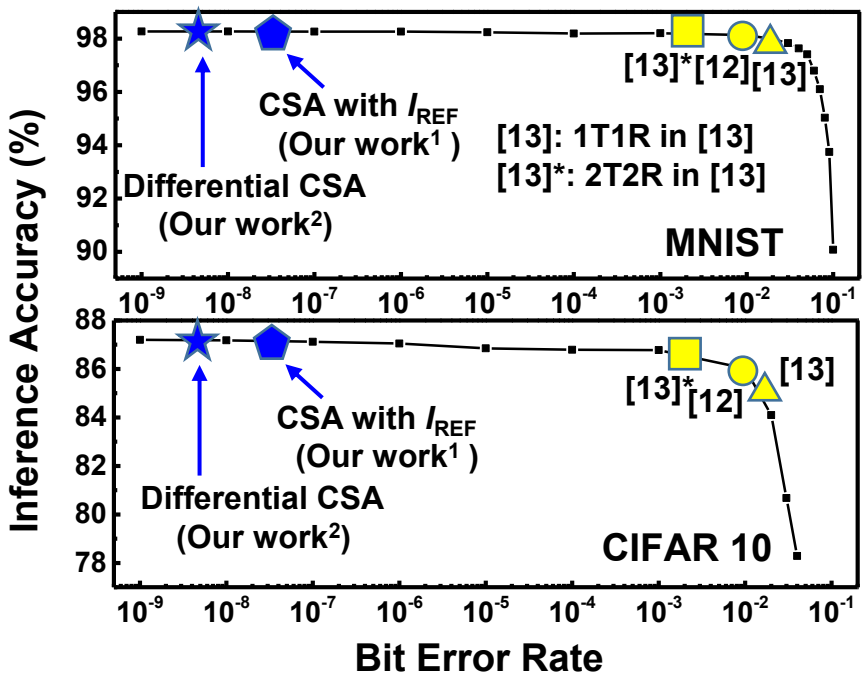

Fig. 2.25. Inference accuracy for the MNIST and CIFAR 10 patterns with respect to bit error rates. The pentagon and star symbols represent our work 1 and 2, respectively.

Chapter 3
Quantized neural networks based on NAND flash memory

3.1 Synaptic architecture for QNN

Fig. 3.1 (a) shows a proposed synaptic architecture based on NAND flash memory for quantized neural networks with differential sense amplifier (DSA). Note that a floating-gate or a charge trap layer are used in the NAND flash memory to store the charge, and the proposed synaptic architecture can be applied to both structures. The synaptic architecture proposed in this work is preferably applied to a vertical NAND flash memory composed of cells with a charge-trap layer. The input voltages from the DSA circuits are applied to the string-select lines (SSL), and the string current is summed through the bit-lines (BL) as shown in Fig. 3.1 (a). In neuromorphic system, VMM is implemented in a hardware synaptic array by mapping the input and weight in the DNN model to the input voltage and the conductance of the synaptic device, respectively. Since the input voltage is applied

to the SSL device, the *I-V* characteristics of the device need to be considered. Fig. 3.2 (a) and (b) show the measured BL current (I_{BL}) versus the SSL voltage (V_{SSL}) curve in log scale and linear scale, respectively, which represents that I_{BL} has a nonlinear relationship with V_{SSL}. However, in the weighted sum equation of the DNN model, the weighted sum output is linearly proportional to the input value. Therefore, the amplitude of the input in the DNN model cannot be encoded as an analogue amplitude of the input voltage in the neuromorphic system. We adopt a binary activation of (1, 0) which can be applied to the nonlinear *I-V* curve by assuming that the input of 1 and 0 correspond to the turn-on (V_{on}) and turn-off voltage (V_{off}) of the SSL device, respectively. In addition, binary activation can reduce the burden of peripheral circuits and QNN with binary activation achieves satisfying accuracy on various recognition tasks. Fig. 3.1 (b) and (c) show a schematic diagram of quantized neural networks and a read pulse scheme as a function of time, respectively. The NAND cells connected to the k^{th} WL in Fig. 3.1 (a) correspond to the synapses in the k^{th} synapse layer shown in Fig. 3.1 (b).

The read bias (V_{read}) is applied to a selected word-line (WL) sequentially, while

a pass bias (V_{PASS}) is applied to unselected WLs as shown in Fig. 3.1 (c). Weights stored in cells connected to a selected WL where the read bias (V_{read}) is applied determine the string current. In this scheme, input voltages are simultaneously applied to all string-select lines (SSL) to sum the currents from multiple strings connected to the BL. It is different from the operation of a conventional NAND flash memory where the turn-on voltage is sequentially applied to each SSL to read the states of the memory cell. Therefore, our scheme can reduce the latency compared to the operation scheme of the conventional NAND flash memory. Therefore, the output for a post-synaptic neuron layer is produced each time a read-bias (V_{read}) is applied to word-line sequentially along the synapse string.

Fig. 3.3 explains the VMM operation using a unit synaptic string array with a DSA circuit. To represent the negative weight value, two adjacent NAND cells in the synaptic string are used as one synapse. The weight is represented by

$$W_k(i,j) = G_k^+(i,j) - G_k^-(i,j) \tag{1}$$

where subscript k represents the k^{th} weight layer, j represents the j^{th} post-synaptic neuron, and i represents the i^{th} synapse connected to the j^{th} post-synaptic

neuron. G^+ and G^- represent the positive and negative weights, respectively. Input is 0 or 1 and if the input is 1, then the DSA circuit applies V_{on} to the SSLs, which contributes to the bit-line current (I_{BL}) sum. On the other hand, if the input is 0, the DSA circuit applies V_{off} to the SSLs, which does not contribute to bit-line current (I_{BL}) sum. Synaptic devices which have a negative weight (G^-) are connected to an odd BL, and synaptic devices which have a positive weight (G^+) are connected to an even BL. In a post-synaptic neuron, the DSA circuit compares the current from the odd BL (I_{ODD}) with the current from the even BL (I_{EVEN}), and then produces a binary output of (1, 0). If I_{EVEN} is larger than $I_{ODD,}$ the DSA circuit produces VDD which corresponds to a binary output of 1. If I_{ODD} is larger than $I_{EVEN,}$ the DSA circuit produces 0 V which corresponds to a binary output of 0. Therefore, this scheme can perform VMM considering positive and negative synaptic weights with only a single input pulse, without additional logic operations. In addition, adopting a differential sensing scheme combined with a neuron activation of (1, 0) instead of (1, -1) is appropriately compatible with existing NAND flash memory architecture to efficiently implement a QNN with binary activation without changing the

memory architecture. The DSA circuits are reused for each neuron layer, which is effective for reducing the area of the peripheral circuits. In addition, the function of the sense amplifier can be extended to perform the neuron activation function, thereby reducing the CMOS circuit overhead. In a neuromorphic system, analogue current needs to be converted to digital outputs using an operational amplifier or an analogue-to-digital converter (ADC) [28]. On the other hand, binary neuron activation enables adopting a differential sense amplifier as a neuron circuit, significantly reducing the burden of neuron circuits and power consumption compared to an ADC. In addition, the binary activation reduces the number of bits in digital communication between the network`s layers compared to multi-bit activation, reducing the burden of peripheral circuits and memory storage.

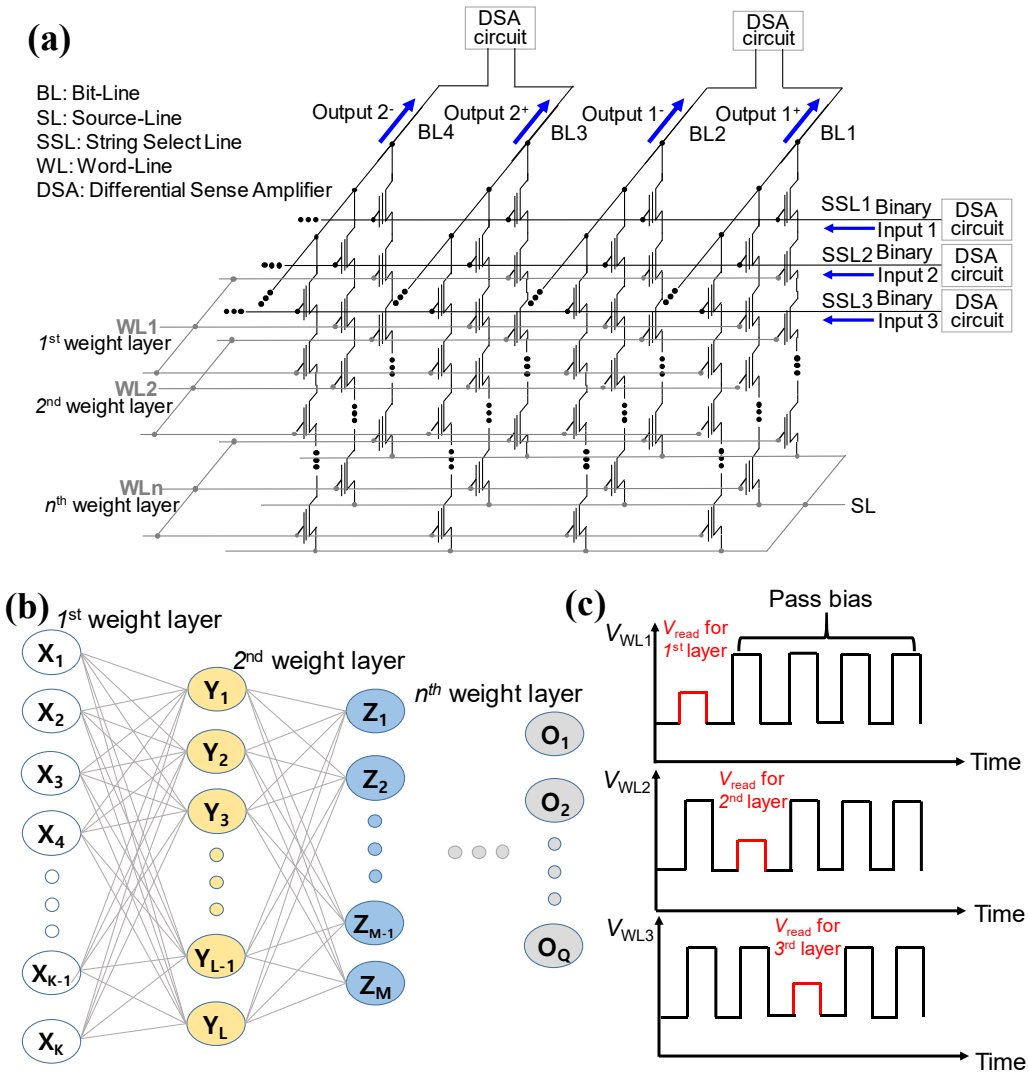

Fig. 3.1. (a) Operation scheme of vector matrix multiplication in the proposed architecture using 3D NAND flash memory architecture. Binary inputs are applied to corresponding SSLs. (b) Schematic diagram of a neural network. (c) Read pulse scheme as a function of time.

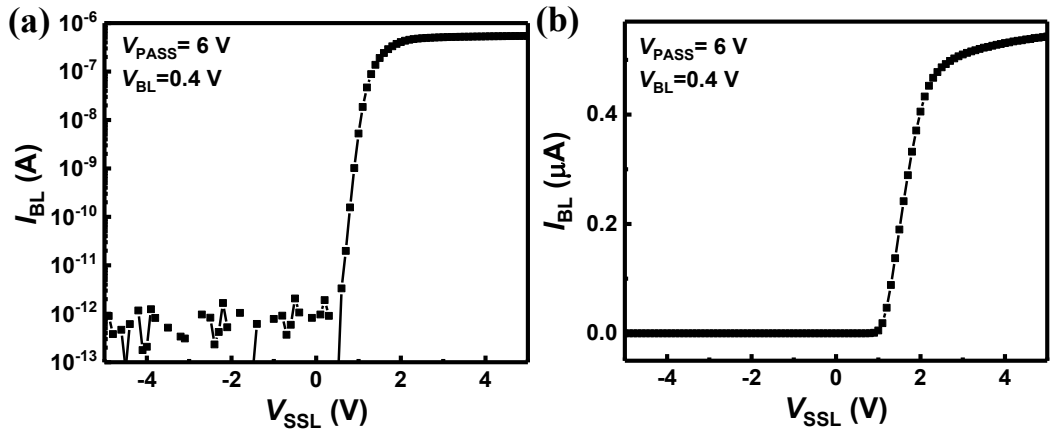

Fig. 3.2. Measured I_{BL}-V_{SSL} curves in (a) log scale and (b) linear scale.

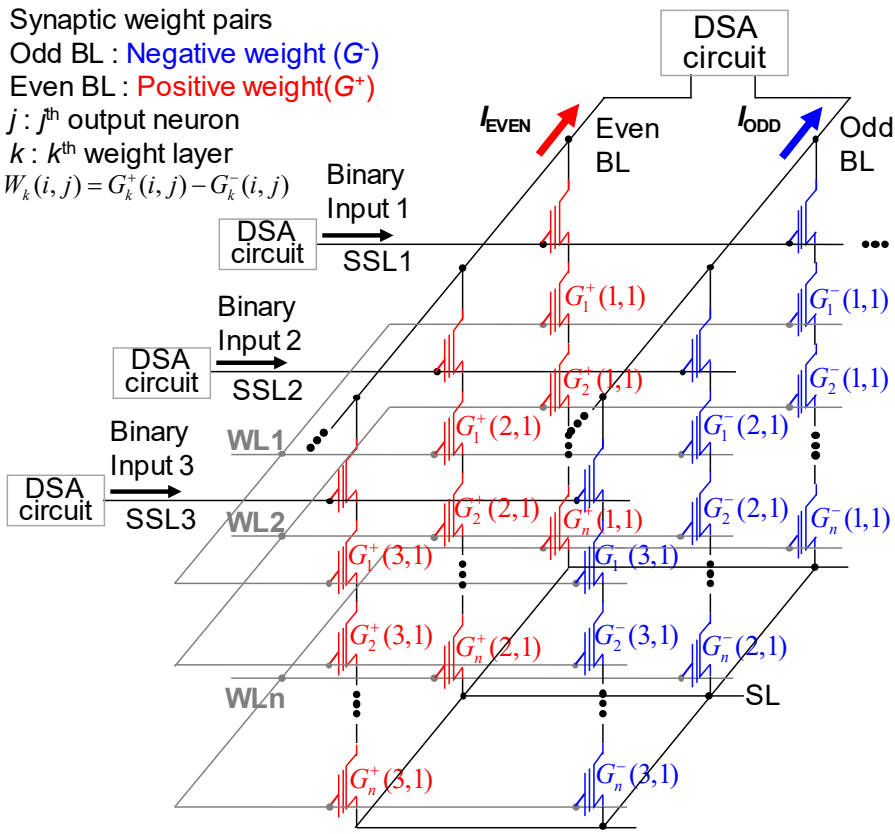

Fig. 3.3. Schematic diagram of the unit synaptic string array consisting of positive weights (G^+) and negative weights (G^-).

3.2 Measurement results

To investigate the characteristics of the NAND cells, floating-gate 2D-NAND flash memory cells fabricated at 26nm technology are measured. Fig. 3.4 represents the BL current (I_{BL}) versus BL voltage (V_{BL}) curves as a parameter of the weight level at a WL voltage (V_{WL}) of 0 V and V_{PASS} of 6 V. Positive and negative weight cells have 8 levels (3-bit) with target currents from 0 to 1.4 μA, resulting in a 4-bit weight. When the input voltage is applied to the synaptic device array, the effective voltage applied to the synaptic devices can be reduced due to the IR drop of the metal wire, resulting in an inaccurate vector-matrix multiplication. Moreover, in the cell string of the NAND flash memory, the channel resistance of the pass cells connected in series to a selected cell where a read bias is applied decreases the effective bias across the drain and source of the selected cell at a given BL bias. Operating NAND cells in the saturation region can eliminate the effect of the IR drop in the metal wire, the noise of the input and output voltages, and the serial parasitic resistance from unselected pass cells to achieve an accurate weight sum. Fig. 3.5 represents the cumulative distribution of I_{BL} measured in the NAND string

array as a parameter of the weight level utilizing the RVW method. The RVW method repeats the cycle of reading, verifying, and writing the conductance of the synaptic devices to ensure that the weights of the pre-trained model are correctly mapped to the synaptic devices. The current of the NAND cell is read by the read voltage (V_{read}) after each program pulse is applied to the device to check that the current of the device is within the target current range. If the measured current is outside of the target current range, then an incremental program voltage pulse is applied to the device. This process is repeated until the current of the device is within the target current range. In this work, the program pulse starts from 11 V with a width of 100 μs and ~40 pulses are needed to tune the conductance of each cell precisely on average. As shown in Fig. 3.5, 8 levels are clearly distinguished by the RVW scheme. In this measurement, 26-nm floating-gate NAND cells are measured in the odd WLs (WL1, WL3, WL5, …) to prevent interference between adjacent cells. Among the 8 levels, the W2 and W3 levels have the largest and smallest device variation, respectively.

Fig. 3.6 (a) and (b) show the effect of the V_{PASS} and V_{read} disturbance,

respectively. In every read operation, V_{PASS} is applied to the unselected word-lines, so the V_{PASS} disturbance needs to be investigated. In Fig. 3.6 (a), the solid square symbols represent an I_{BL}-V_{WL} curve measured in a fresh cell device, and the curve measured after applying a V_{PASS} of 6 V to the cell 64×10^4 times is represented by open circle symbols. Since these two I_{BL}-V_{WL} curves are almost the same, the disturbance caused by V_{PASS} appears to be negligible. The I_{BL}-V_{WL} curves measured after the application of 10 V_{PGM} pulses with amplitudes of 12 and 13 V for the programming of the fresh cell device are represented by solid triangle and diamond symbols, respectively. The I_{BL} exhibits a negligible change with a 64×10^4 V_{PASS} disturbance compared to V_{PGM}. In the read operation, it is necessary to investigate the influence of hot carriers because hot carriers that can be generated in the channel of the NAND cell operating in the saturation region can change the cell's threshold voltage. In Fig. 6 (b), the solid square symbols represent an I_{BL}-V_{WL} curve measured in a fresh cell device, and the curve measured after applying a V_{read} of 0 V to the cell 64×10^4 times is represented by open circle symbols. Since these two I_{BL}-V_{WL} curves are almost the same, the hot carrier effect appears to be negligible. In

addition, as shown in Fig. 3.6, the off-current (I_{off}) is below 10 pA and the on/off current ratio is more than 10^5, which provides a high bandwidth to sum the currents in parallel from much more cells compared to the cells in RRAM [18], [19]. Fig. 3.7 shows the measured retention characteristics as a parameter of the weight level at T=300 K. I_{BL}s in 8 levels exhibit excellent retention characteristics up to 10^4 s.

In the NAND cell string, since multiple cells are serially connected between the bit-line and the source-line, the pass cells act as resistors in the read operation. As a result, the effective voltage across the control gate and source of the selected cell at a given read WL voltage depends on the location of the selected cell in the cell string. Therefore, under the assumption that all cells in the string have the same threshold voltage, a cell close to SL has a higher BL current than a cell close to BL when the same WL bias is applied to the control gate for a read operation. Consequently, the different BL current depending on the position of the cell results in an inference error in the neuromorphic system. To address this problem, the read-write-verify method is adopted. The method enables the same saturation current regardless of the cell's position within the cell string by adjusting the threshold

voltage of the cells. Fig. 3.8 shows the eight I_{BL}-V_{BL} curves measured from two cells at different positions in the cell string. WL 0 represents the cell near the SL and WL 60 represents the cell near the BL. As shown in Fig. 3.8, the currents of WL0 and WL60 show almost the same value at the 8 levels of the target currents by using the RVW method. In this way, the cell device in the NAND cell string can be adjusted to have a target current regardless of its position.

We investigate device variation that affects the inference accuracy of neural networks. Fig. 3.9 shows the I_{BL} distribution measured from cells in a NAND string array and Gaussian fitting curves at the 2nd weight level (W2) and 3rd weight level (W3). The W2 and W3 levels have the largest and smallest device variation, respectively, among the 8 weight levels shown in Fig. 3.5. The estimated device variation (σ_w/μ_w) of the W2 and W3 levels are 3.04% and 1.88% respectively, based on the measured data with the assumption of a Gaussian distribution [29]-[35]. Fig. 3.10 shows a demonstration of the VMM operation using a synaptic string. V_{read} is applied to the first WL where the control electrodes of two cells are connected, and the target currents of the left and right cells controlled by the first WL are the 1st

weight level (0.2 μA) and 2nd weight level (0.4 μA), respectively. When SSL1 turns on and SSL2 turns off, the I_{BL} is 0.21 μA. On the other hand, when SSL1 turns off and SSL2 turns on, the I_{BL} is 0.405 μA. When both SSL1 and SSL2 turn on, the I_{BL} is 0.615 μA, as shown in Fig. 3.10 (b). Therefore, the VMM operation is successfully demonstrated in the proposed synaptic string array.

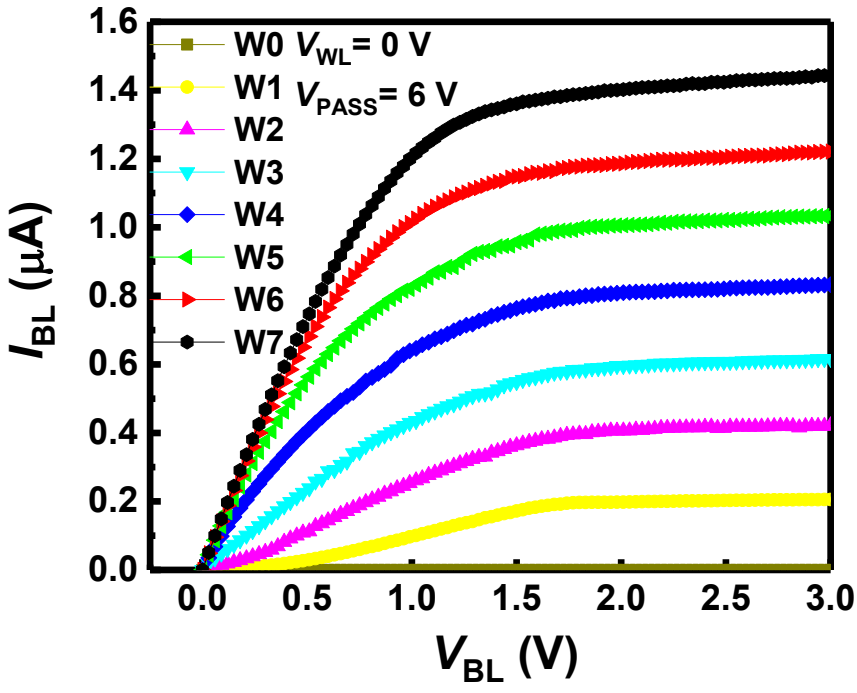

Fig. 3.4. Measured I_{BL}-V_{BL} curves as a parameter of the various weight levels at V_{WL}= 0 V and V_{PASS}= 6 V. The positive and negative weight cells have 8 levels with target currents from 0 to 1.4 μA.

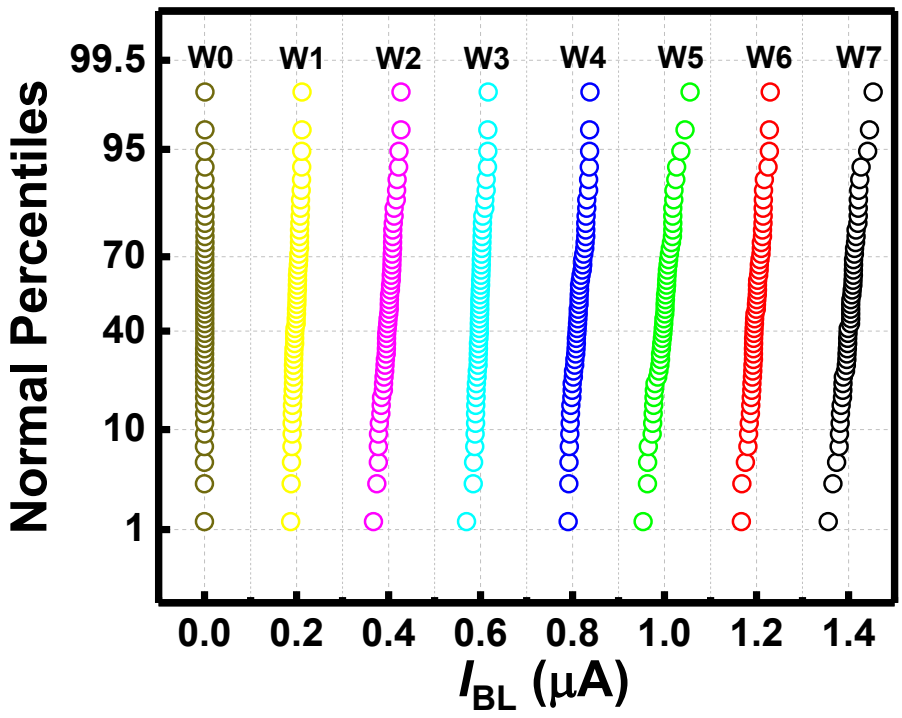

Fig. 3.5. Measured cumulative distribution of I_{BL} as a parameter of the weight level at V_{WL}= 0 V and V_{PASS}= 6 V. Using the read-verify-write scheme, the current can be matched to the target current.

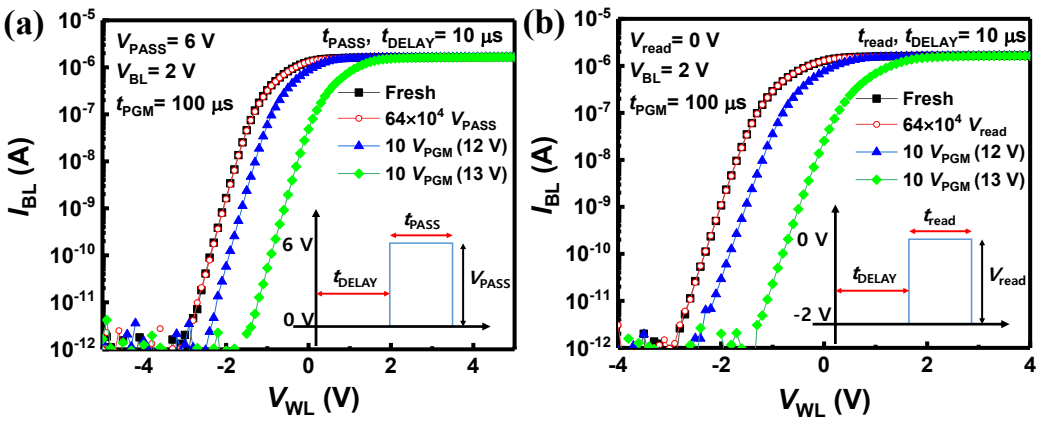

Fig. 3.6. Measured I_{BL}-V_{BL} curves with (a) V_{PASS} and (b) V_{read} disturbance and V_{PGM}.

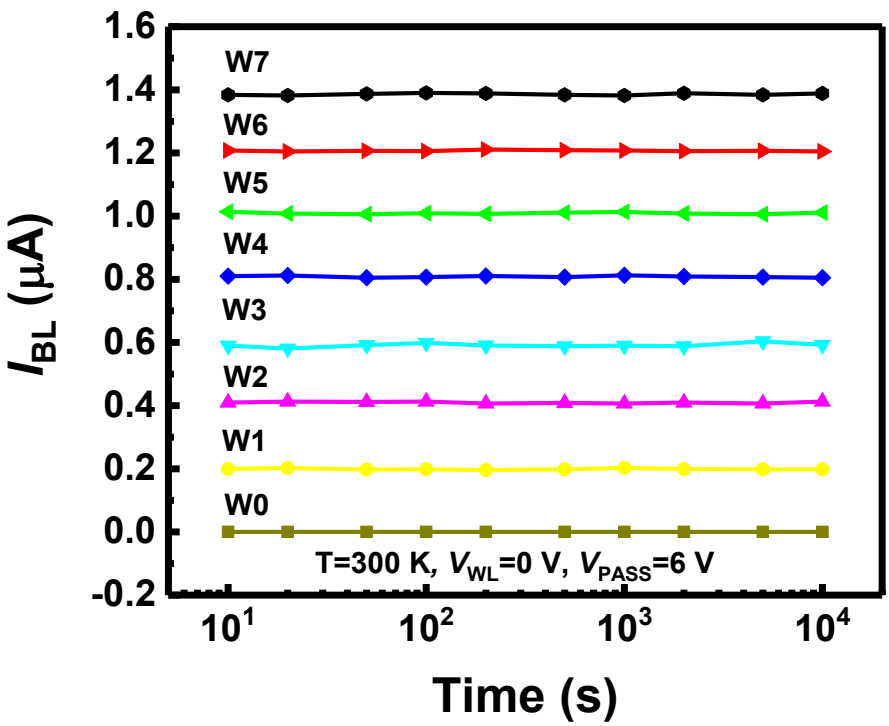

Fig. 3.7. Measured retention characteristics up to 10^4s as a parameter of various weight levels at T=300 K.

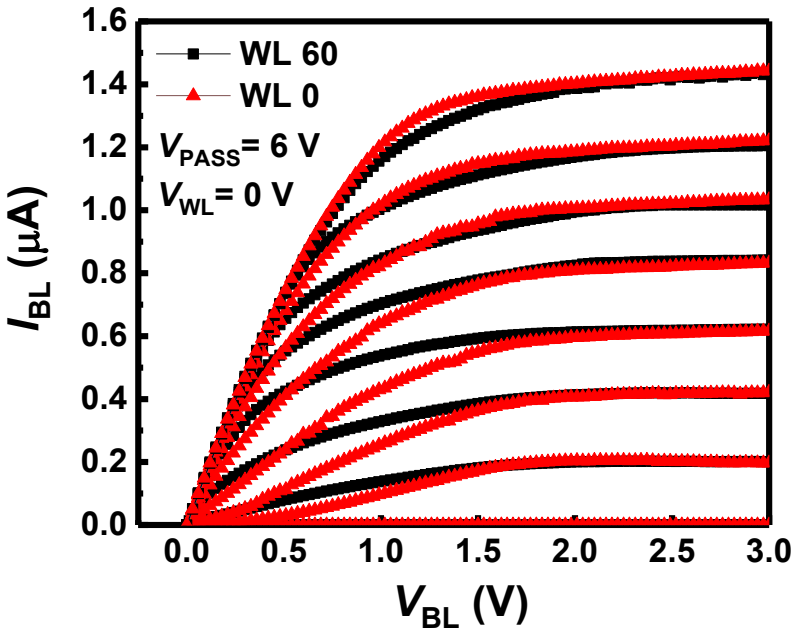

Fig. 3.8. Measured I_{BL}-V_{BL} curves as a parameter of various weight levels at WL 60 (square symbol) and WL 0 (triangle symbol).

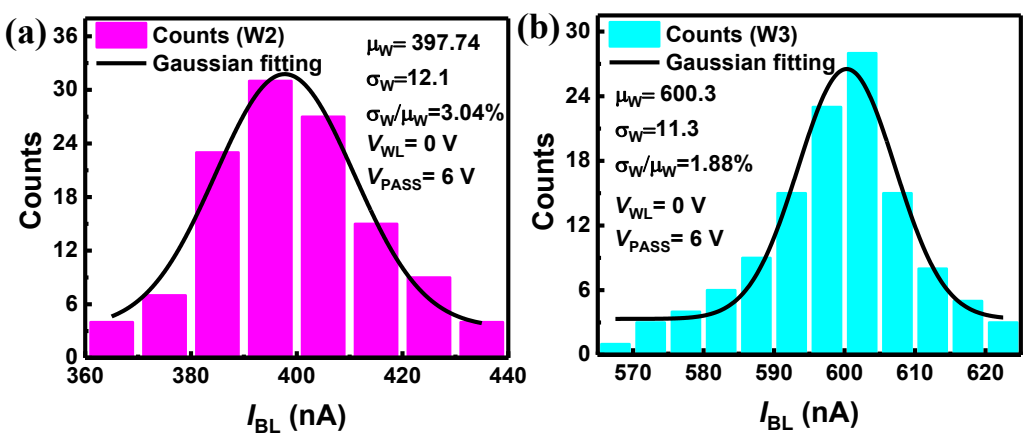

Fig. 3.9. Measured I_{BL} distribution of the NAND string array at the 2nd weight level (W2) and 3rd weight level (W3).

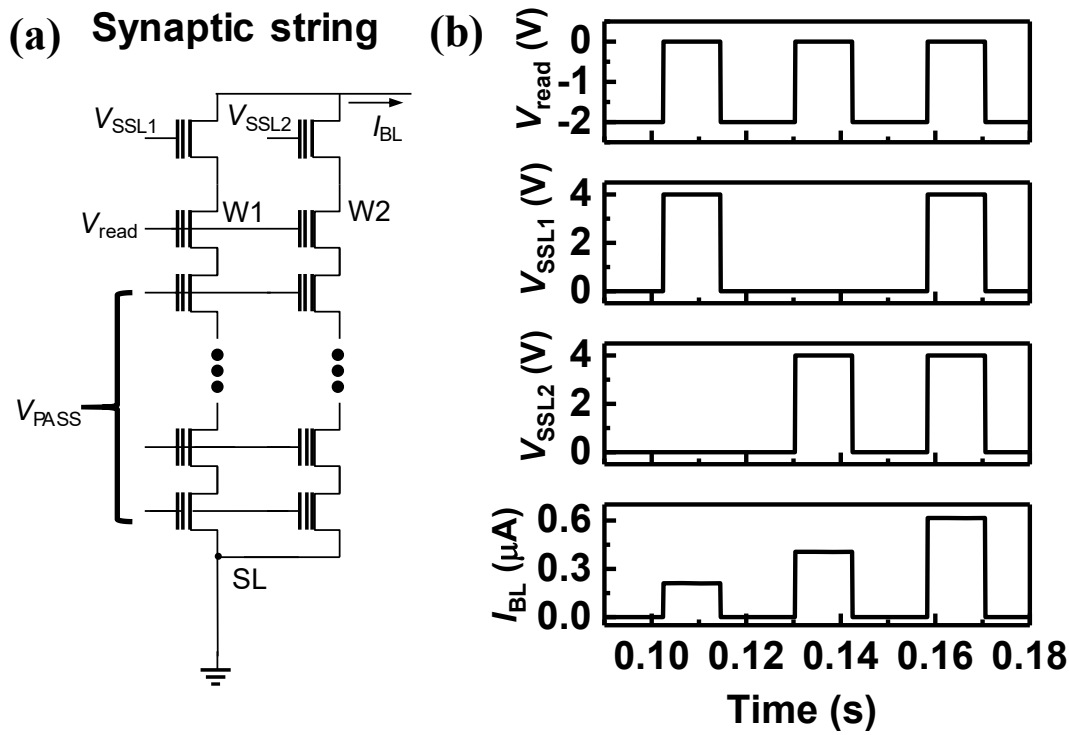

Fig. 3.10. (a) Schematic diagram of a synaptic string for demonstration of the VMM.

(b) Measured transient waveforms of I_{BL} which is the result of the VMM.

3.3 Simulation results

Fig. 3.11 (a) represents a DSA circuit which serves as a binary neuron. The DSA circuit is simulated utilizing a BSIM-CMG model based on a 20-nm FinFET. Fig. 3.11 (b) and (c) show transient waveforms of the DSA circuit when the binary output is 1 ($I_{EVEN}>I_{ODD}$) and 0 ($I_{ODD}>I_{EVEN}$), respectively. In Fig. 3.11 (b), as I_{EVEN} is larger than $I_{ODD,}$ the voltage at node P_B drops to the trip-point voltage faster than node P, raising node P toward VDD. As a result, V_{OUT_B} drops to zero while V_{OUT} remains at VDD. On the other hand, in Fig. 3.11 (c), as I_{ODD} is larger than I_{EVEN}, the voltage at node P drops to the trip-point voltage faster than node P_B, raising node P_B toward VDD. As a result, V_{OUT} drops to zero while V_{OUT_B} remains at VDD.

Because synaptic devices have discrete conductance levels depending on the application of the program pulse, weight quantization needs to be considered. Fig. 3.12 shows a simulated inference accuracy of the neural networks with post-training quantization (PTQ) for the CIFAR 10 and MNIST datasets. Fully connected neural networks consisting of 5 layers (784-1024-1024-1024-10) and convolutional neural networks consisting of 6 convolution layers and 3 fully connected layers are used

for training the MNIST and CIFAR 10 datasets, respectively. Post-training quantization means that floating-point weights are obtained in training without fine-tuning and then they are quantized. The fully connected and convolutional neural networks with the floating-point weight and binary activation are trained without fine-tuning for the MNIST and CIFAR 10 datasets, respectively. Then the trained weights obtained in software with floating-point precision are quantized and transferred to synaptic devices with the linear quantization method. Note that, proposed synaptic architecture in this work utilizes an off-chip learning scheme. As shown in Fig. 3.12, as the bit-width of the weights decreases, the inference accuracy decreases because the quantization error becomes larger. Quantizing the floating-point weights to low-bit values results in the variation of the weighted-sum output, which leads to a decrease of the inference accuracy. As the bit-width of the weights decreases from 9 to 4, the inference accuracy decreases by 1.35 % and 0.32 % with PTQ for the CIFAR 10 and MNIST datasets, respectively.

To reduce the degradation of the inference accuracy, in off-chip training, we adopt a quantization training (QT) method which uses fine-tuning during the

training process [10]. Reported QT methods can cause overhead by adding extra hyper-parameters or modifying the original training procedure. Recent works have tried to reduce the overhead and redundancy in QT [36], [37]. In [9]-[11], QT methods reduce the memory size and accesses, which leads to dramatic improvement in power consumption and computation speed compared to the full-precision model. In addition, they reduce the quantization error compared to PTQ. Algorithm 1 describes the flow of QT [10]. As shown in Algorithm 1, quantized weights and activations are used in the forward and backward propagation, but floating-point gradients of the weights are accumulated in the floating-point variables to ensure that minor gradient updates affect the update of the weight values. Because the weighted-sum error caused by the quantization is reduced during the training process by the QT, the QT reduces the degradation of the inference accuracy by the quantization compared to the PTQ. Fig. 3.13 shows the simulated inference accuracy of the neural networks with the QT method. The quantized fully connected and convolutional neural networks with a 4-bit weight and binary activation are trained for inference of the MNIST and CIFAR 10 datasets,

respectively. As shown in the inset, the QT increases the inference accuracy by 1.24 % and 0.3 % for the CIFAR10 and MNIST datasets, respectively, compared to the PTQ. The final accuracies with the QT method for the CIFAR 10 and MNIST datasets are 88.27 and 98.32%, respectively, which is similar to those obtained in the floating-point neural networks (FNN) with floating-point weights. Therefore, by adopting a quantization training method, QNN can be implemented with a high inference accuracy.

Because the resistance of the metal wire degrades the inference accuracy, we investigate its effect through a simulation. Fig. 3.14 shows the effect of the metal wire resistance as a parameter of the resistance of the synaptic device (R_S) assuming that the resistance of the metal wire between adjacent synaptic devices is 2.5 Ω [17]. As the size of array becomes larger, the effective voltage (V_E) across the synaptic device which is farthest from the voltage source decreases significantly when R_S is 5 kΩ. On the other hand, when R_S is ~24 MΩ which is the output resistance of a NAND cell operating in the saturation region, the V_E across the synaptic device hardly decreases even if the array size increases. Therefore, R_S needs to be large to

reduce the effect of the metal wire and to implement a reliable neuromorphic system. To achieve a ratio of V_E and input voltage (V_{input}) more than 99 %, the R_S needs to be more than 1 MΩ when the number of rows or columns in the array is 64 or less as shown in Fig. 3.14. However, when the resistance in the RRAM devices is greater than 1 MΩ, the device variation becomes large [38]. Fig. 3.15 shows the simulated inference accuracy of the QNN with respect to the device variation (σ_w/μ_w) for the CIFAR 10 and MNIST datasets. In our work, the device variation at the W2 level is the largest, so it is used to evaluate the inference accuracy. The σ_w/μ_ws of the RRAM devices with a R_S larger than 1 MΩ are used for the comparison. The device variation of the NAND cells is much lower than that of the RRAM devices with a variation of 25~48 % [39]-[41], resulting in a higher inference accuracy than the RRAM devices by 2~7 % and 0.04~0.23 % for the CIFAR 10 and MNIST datasets, respectively as shown in Fig. 3.15.

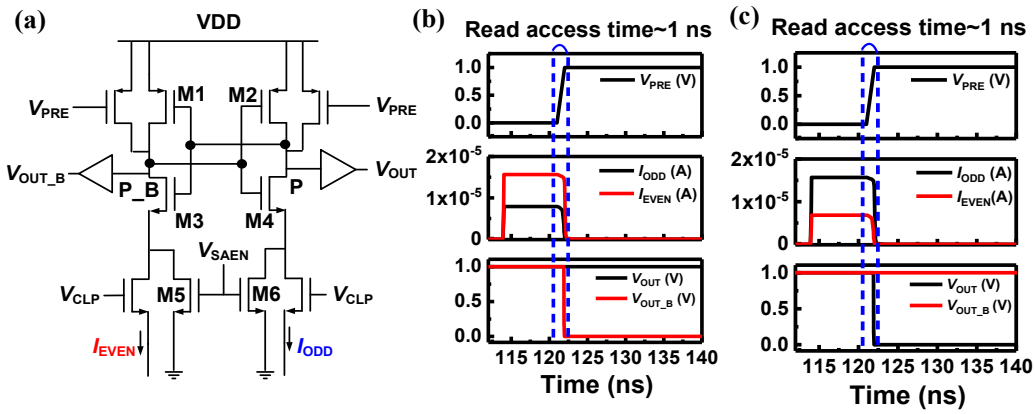

Fig. 3.11. (a) Circuit diagram of the differential current sense amplifier. The simulated transient waveforms of the circuits when the binary output is (b) +1 (I_{EVEN}>I_{ODD}) and (c) 0 (I_{ODD}>I_{EVEN}).

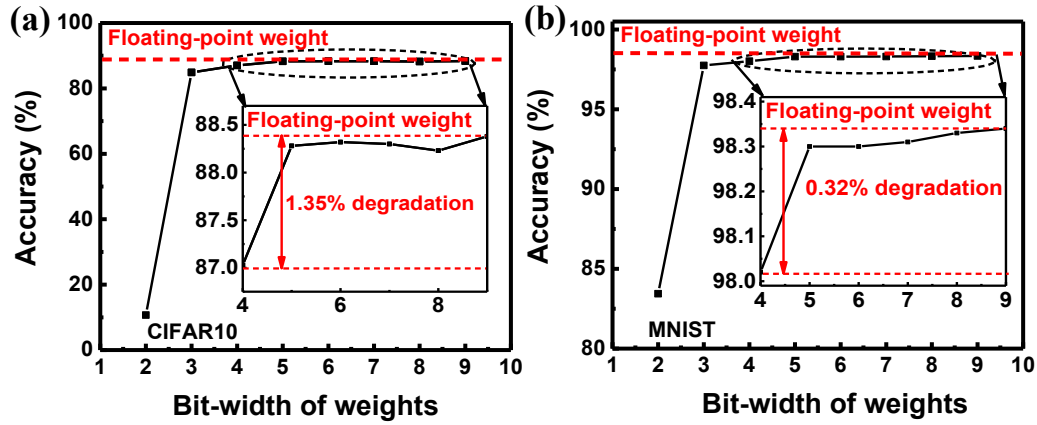

Fig. 3.12. Simulated inference accuracy of the neural networks with PTQ for the CIFAR 10 and MNIST datasets. As the bit-width of the weights decreases to 4, the inference accuracy decreases by 1.35 % and 0.32 % with the PTQ for the CIFAR 10 and MNIST datasets, respectively.

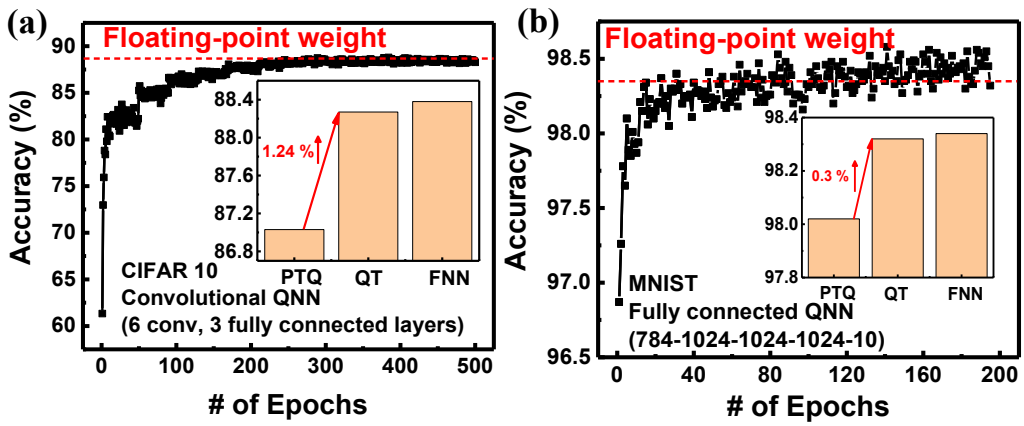

Fig. 3.13. Simulated inference accuracy of the neural networks with the quantized training (QT). The final inference accuracy for the CIFAR 10 and MNIST datasets are 88.27 and 98.32%, respectively. The QT increases the inference accuracy by 1.24 % and 0.3 % for the CIFAR10 and MNIST datasets, respectively, compared to the PTQ.

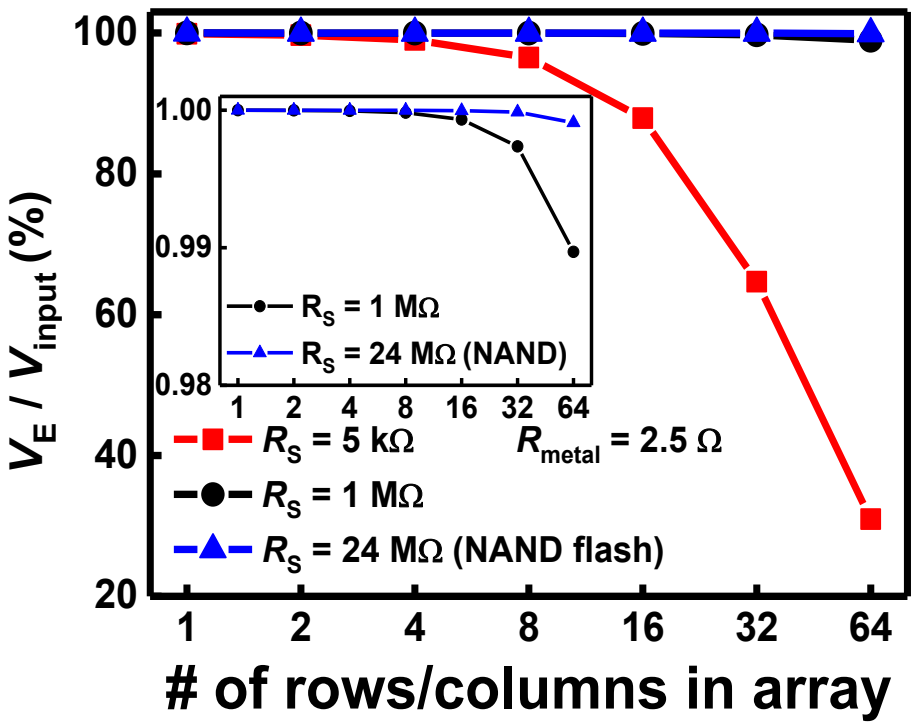

Fig. 3.14. Effective voltage across the synaptic device with the array size.

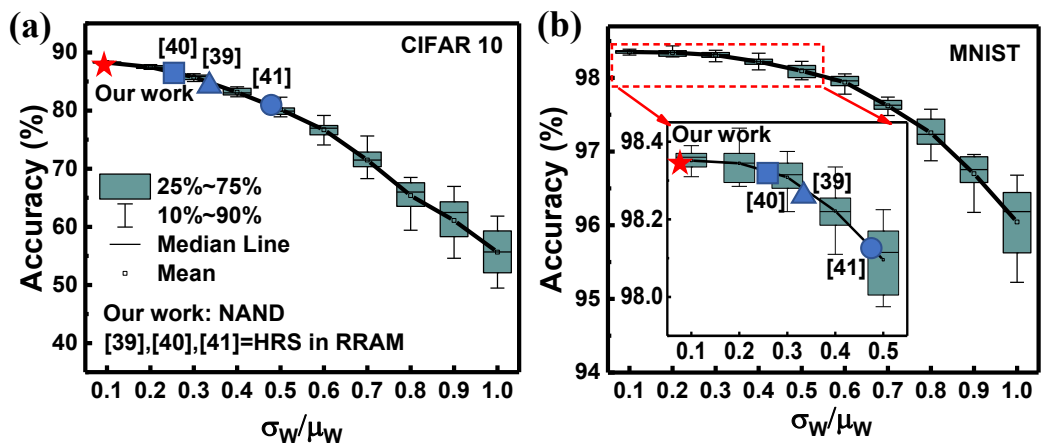

Fig. 3.15. Simulated inference accuracy of the QNN with respect to the device variation (σ_w/μ_w) for the CIFAR 10 and MNIST datasets.

Chapter 4
On-chip learning based on NAND flash memory

4.1 Synaptic architecture for on-chip learning

Fig. 4.1 shows a synaptic device array where two adjacent cells which represent the positive weight (G^+) and negative weight (G^-) of one synapse are adjacently located to each other in a synaptic array. The FP can be correctly performed in the synaptic array as shown in (1). However, as shown in (2), BP cannot be performed in a single time step on the synaptic array. To perform BP in the synaptic array, error input1 (δ_1) is applied to the synaptic array first, then error input2 (δ_2) is applied to the synaptic array.

On the other hand, Fig. 4.2 shows a synaptic array where two cells which represent the G^+ and G^- of one synapse are separated into different synaptic array. In this array, BP can be performed in a single time step as shown in (3) ~ (6). Therefore, G^+ and G^- should be separated to reduce latency in BP.

Figs. 4.3 and 4.4 show the proposed circuit operation of performing FP and BP in the proposed synaptic architecture based on NAND flash memory, respectively, to enable on-chip learning. In the conventional NAND flash memory, source-lines (SL) are connected in a block [21], which impedes the backward propagation. Therefore, in the synaptic architecture proposed to enable both FP and BP, SLs are separated in a direction crossing the bit-line (BL). Two NAND cells which are located in different synaptic weight array (G^+ array and G^- array) are used as one synapse to represent negative weight.

In FP, input biases are applied to the BLs and each weighted sum current is read from a separated SL. In BP, error inputs (δ) are applied to string-select line (SSL) and each weighted sum current (σ) is read from a BL. If error inputs are applied to SLs and weighed sum current is read from BLs, which is the method used in RRAM array [42], the current of the cell at a specific word-line (WL) location can be changed by the resistance of the pass cells depending on the cell's location in the string.

Therefore, error inputs need to be applied to SSLs and amplitude of V_{SL}, V_{SSL}

and V_{BL} should be the same in both the FP and BP to make string currents in the FP and BP equal. In this architecture, the input and error input are provided by the PWM (pulse width modulation) circuit as width-modulated pulses with fixed amplitude. Therefore, by applying the width-modulated pulses to V_{BL} and V_{SSL} in FP and BP, respectively, accurate VMM can be performed in both propagations eliminating the effect of pass cells in the NAND flash memory architecture. Note that the BL currents (I_{BL}) in FP and BP are in the same direction. In this scheme, V_{SSL} is applied to all SSLs to accumulate the currents from NAND strings through BL. On the other hand, in the operation scheme of conventional NAND flash memory, V_{SSL} is applied to a selected SSL to read the information of a NAND cell. Thus, the proposed operation scheme can increase throughput compared to that of the conventional NAND flash memory. Fig. 4.5 shows the schematic diagram of neural network which is composed of n weight layers and pulse diagram with the time. The cells connected to the k^{th} WL of NAND flash memory in Figs. 4.3 and 4.4 represent synapses in the k^{th} weight layer in Fig. 4.5 (a). The read bias (V_{read}) and pass bias are applied to a selected WL and unselected WLs, respectively, as

shown in Fig. 4.5 (b) and (c). Applying V_{read} to the k^{th} WL produces the output of all neurons in the k^{th} neuron layer. In FP, the V_{read} is applied to from the 1^{st} to n^{th} WL to produce weighted-sum output. On the other hand, during BP, the V_{read} is applied to from the n^{th} to 1^{st} WL to produce summed error. By using this scheme, the synaptic weights can be transposed, and FP and BP can be correctly performed in NAND flash memory.

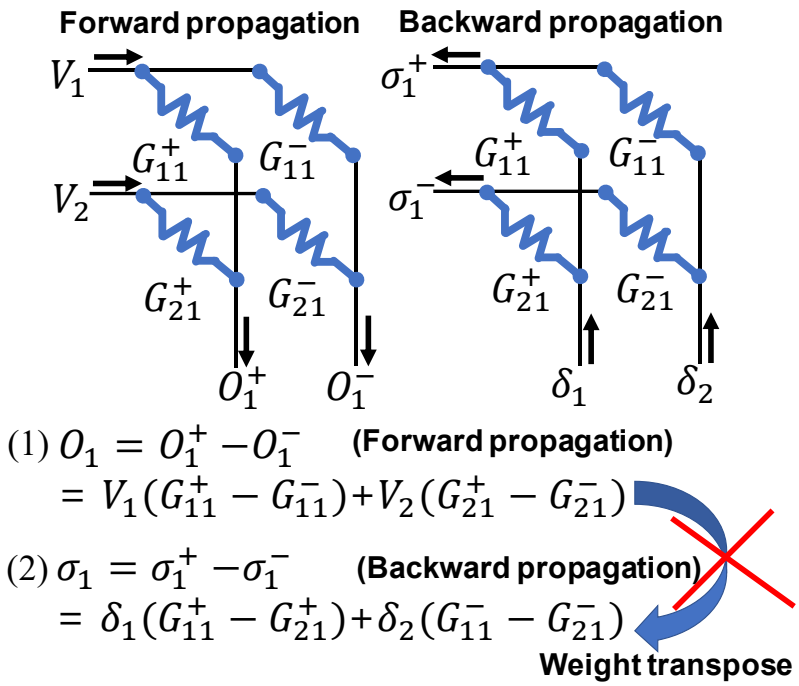

Fig. 4.1. Synaptic array architecture consisting of two adjacent cells representing G^+ and G^-. The weights cannot be transposed.

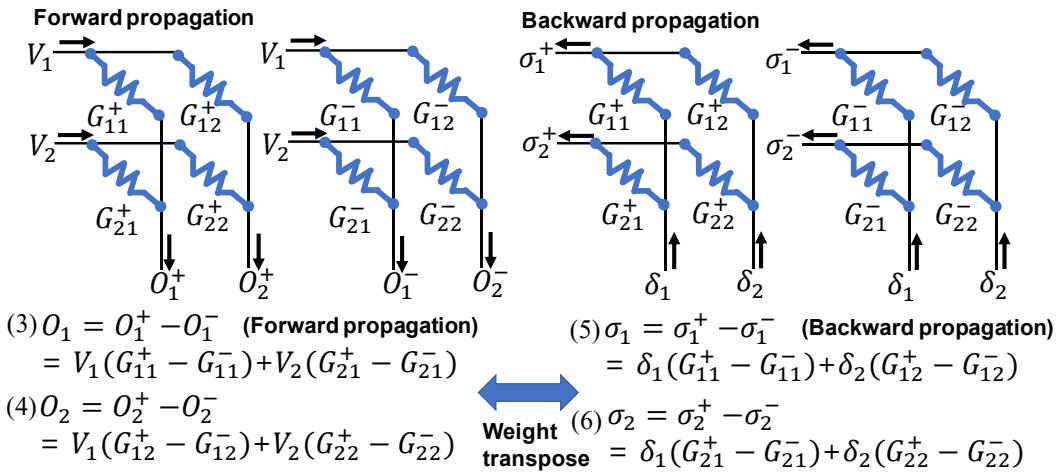

Fig. 4.2. Synaptic array architecture where positive (G^+) and negative (G-) weights are separated in different arrays.

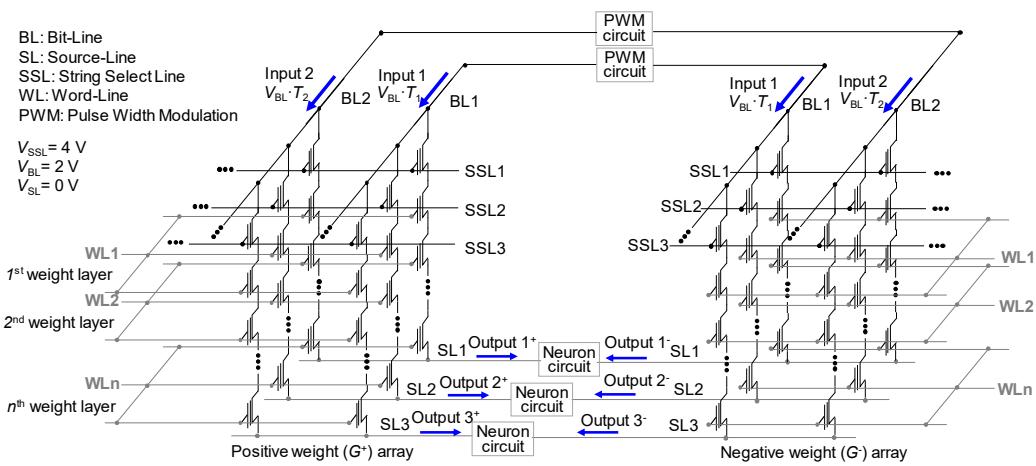

Fig. 4.3. Synaptic array architecture based on NAND flash memory for FP operation of on-chip learning.

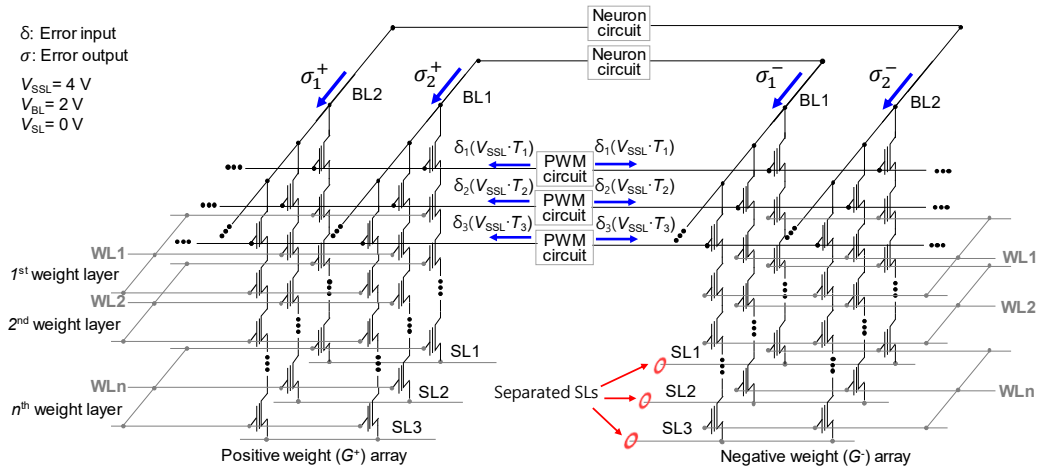

Fig. 4.4. Synaptic array architecture based on NAND flash memory for BP operation of on-chip learning.

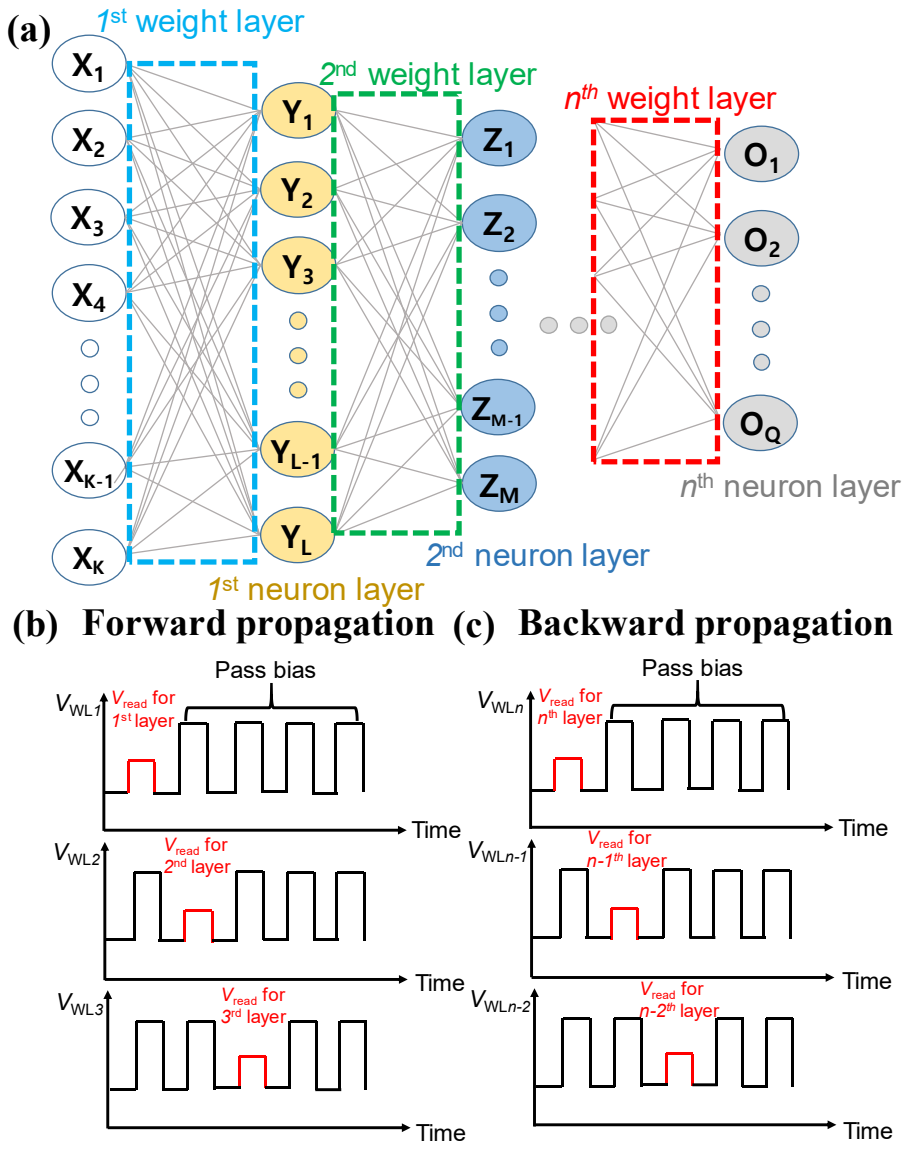

Fig. 4.5. (a) Schematic diagram of neural networks consisting of *n* weight layers. Diagram showing pulses applied to WL over the time in (b) forward propagation and (c) backward propagation.

4.2 Measurement results

The 2-D NAND flash memory fabricated with 26 nm technology is measured in this work to investigate the characteristics of the NAND cells as synaptic devices. One cell string consists of 64 cells including an SSL transistor, a ground select line transistor, and two dummy cells. Linearity of the conductance (G) response of the NAND cell and the G_{MAX} / G_{MIN} of the response is investigated. Fig. 4.6 (a) shows the measured BL current-to-WL voltage (I_{BL}-V_{WL}) curves in a NAND cell with increasing number of program pulses (V_{PGM}) of 14 V. The V_{th} increases from the initial V_{th} as the number of V_{PGM} increases. The off-state voltage (V_{off}) needs to be 0 V to avoid negative bias conditions which puts lots of burden on the peripheral circuit. In addition, the off-current (I_{off}) needs to be kept low by increasing initial V_{th}.

In on-chip learning, linearity of conductance response affects the learning accuracy [43]. To analyze the effect of V_{read} on the linearity of conductance response of the NAND cell, the normalized G responses shown in Fig. 4.6 (b) are obtained from the I_{BL}s measured at V_{WL} of from V_1 to V_5. The G responses of 5-bit represent

the synaptic weight of 6-bit, as weight is represented by difference of G^+ and G^-. The linearity improves as the V_{WL} increases from V_1 to V_5. However, I_{MAX} / I_{MIN} (=G_{MAX} / G_{MIN}) decreases as V_{WL} increases from V_1 to V_5 as shown in Fig. 4.6 (a).

Fig. 4.7 (a) shows the cell's I_{BL}-V_{WL} curve (circle symbols) for V_{PASS} disturbance and the cell's I_{BL}-V_{WL} curves when the cell is programmed with V_{PGM}s of 12 V and 13 V. The I_{BL} exhibits a negligible change with a V_{PASS} disturbance compared to those (triangles) with V_{PGM} as shown in Fig. 4.7 (a). Fig 4.7 (b) shows the cycle-to-cycle variation of a NAND flash cell. In 1 cycle, 31 V_{PGM}s of 14 V and 1 V_{ERS} of -10 V are applied to a NAND cell. The variation of conductance response is negligible up to 1k cycles.

Fig. 4.8 (a) shows the program (P) and erase (E) windows when the P/E cycle is repeated up to 3×10^3. The $\Delta V_{th,PGM}$ (ΔV_{th} by V_{PGM}) increases and $\Delta V_{th,ERS}$ (ΔV_{th} by V_{ERS}) decreases as the number of P/E cycles increases when the V_{PGM} is 16 V. On the other hand, the V_{PGM} of 14 V has a very small effect on the program and erase windows. Fig. 4.8 (b) shows retention characteristic of a NAND cell. Each weight level hardly changes until 10^4 s as shown in Fig. 4.8 (b).

To investigate the effect of pass cells on the NAND string current in the VMM process, we measure I_{BL} using the on-chip learning method in RRAM array. Inputs are applied to rows in forward pass and error inputs are applied to columns in backward pass in previous on-chip learning method commonly used in RRAM array. If it is applied to NAND flash memory array, inputs are imposed on BL in forward pass and error inputs are imposed on SL in backward pass. Fig. 4.9 (a) and (b) show circuits for measuring I_{BL}s when V_{BL}= 2 V and V_{SL}= 0 V and when V_{BL}= 0 V and V_{SL}= 2 V, respectively. As shown in Fig. 9 (c), I_{BL} measured in (a) is different from that measured in (b) due to the channel resistance effect of pass cells. In the NAND cell string, since multiple cells are serially connected between the BL and the SL, the pass cells act as resistors in the read operation. As a result, when bias of V_{BL} and V_{SL} interchanges, the effective voltage across the ground and control gate of the selected cell changes, which causes the error in reading I_{BL}. Therefore, the on-chip learning method used in RRAM array cannot be used in the NAND flash memory. To solve this problem, we propose a novel architecture shown in Figs. 4.3 and 4.4. In the proposed architecture, the input and error input are applied to BL and SSL,

respectively, keeping I_{BL} in FP and BP in the same direction. It allows the weight to be transposed and eliminates the effect of pass cells.

In addition, as the cell current can be changed during the weight update process due to the pattern loading effect, we propose a program scheme to solve this problem. Fig. 4.10 (a) shows circuits for measuring I_{BL}s. All cells in the string initially have a low threshold voltage ($V_{t,low}$), then the 8 cells in the middle of the string are programmed one by one to have a high threshold voltage ($V_{t,high}$). Fig. 4.10 (b) and (c) show the I_{BL}-V_{BL} curves measured in cells closest to BL and SL, respectively. Fig. 4.10 (d) shows the current ratio of I_{BL} and maximum current in the cell (I_{MAX}) with increasing number of programmed cells. I_{MAX} is the current measured when all cells in the string initially have a low threshold voltage ($V_{t,low}$). The I_{BL} of the cell closest to the BL decreases more than that of the cell closest to the SL with increasing number of programmed cells as shown in Fig. 4.10 (d). It is because the effective voltage across the ground and the control gate of the selected cell at a given read WL voltage changes depending on the location of the selected cell in the cell string.

Therefore, I_{BL} of the cell closest to the BL decreases when the number of programmed cells in one string increases because the effective voltage across the ground and control gate of the cell decreases. Thus, a program scheme is proposed in which a program voltage (V_{PGM}) is sequentially imposed on from the cell closest to the SL to the cell closest to the BL in the weight update process. The proposed scheme can reduce the pattern loading effect during weight update.

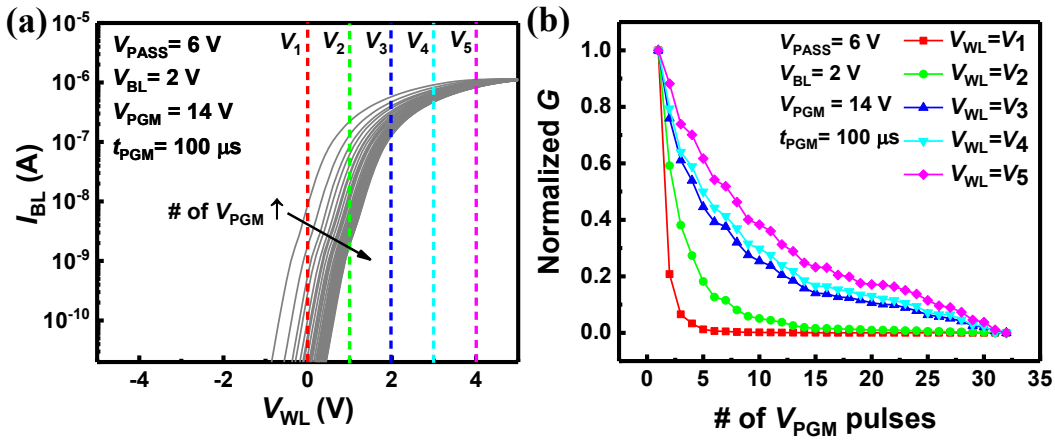

Fig. 4.6. (a) I_{BL}-V_{BL} curves with increasing number of program pulses. (b) Normalized conductance (G) responses measured in (a).

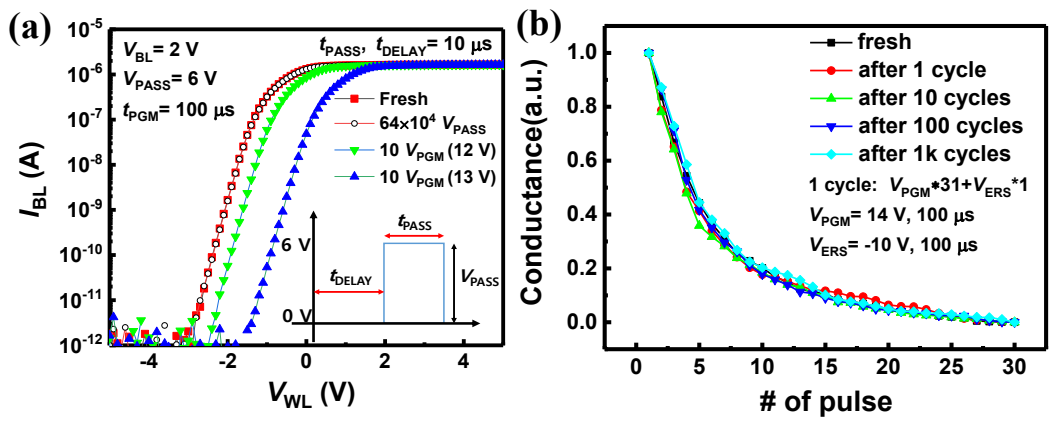

Fig. 4.7. (a) I_{BL}-V_{WL} curves measured in fresh, V_{PASS} disturbed, and programmed cell. (b) Conductance response of fresh and cycled cell.

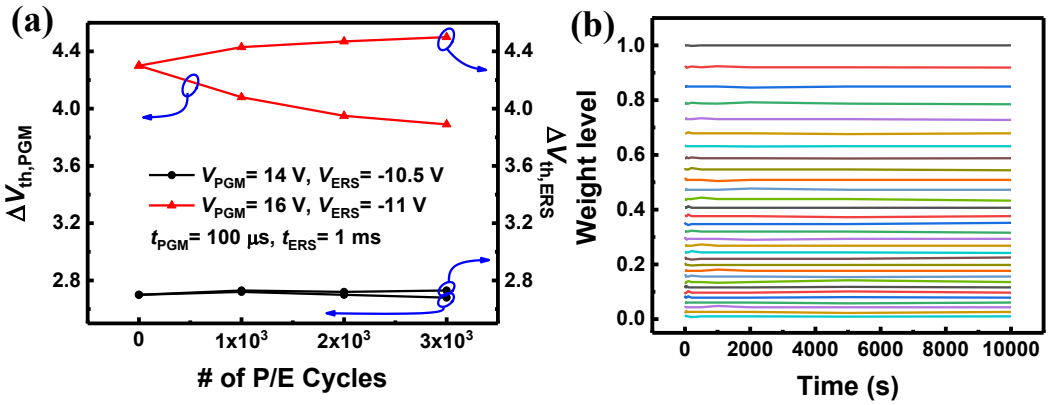

Fig. 4.8. (a) Changes in program and erase windows in the process of applying P/E (program/erase) cycles up to 3×10^3. (b) Retention characteristics of a NAND cell.

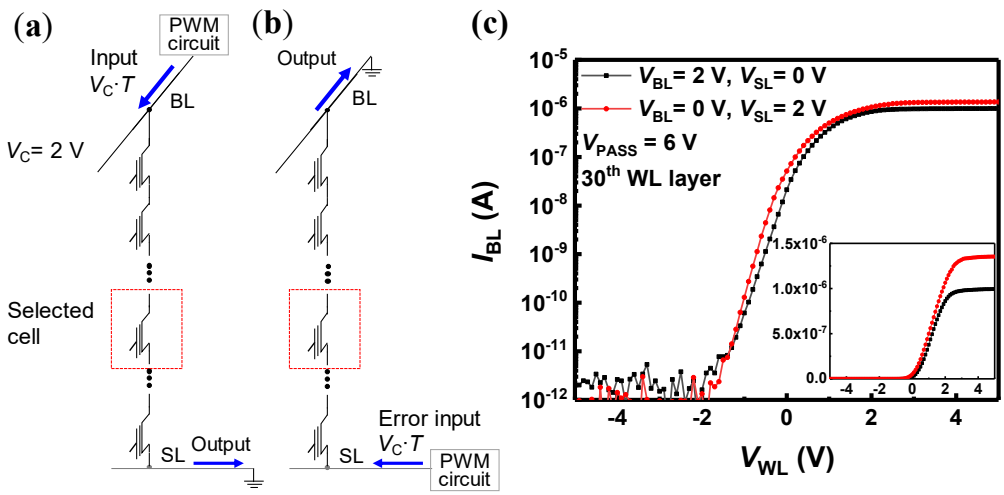

Fig. 4.9. Circuits that measure I_{BL} in a selected cell when (a) V_{BL}=2 V and V_{SL}=0 V and when (b) V_{BL}=0 V and V_{SL}=2 V. (c) I_{BL}-V_{WL} curves measured under the conditions of (a) and (b).

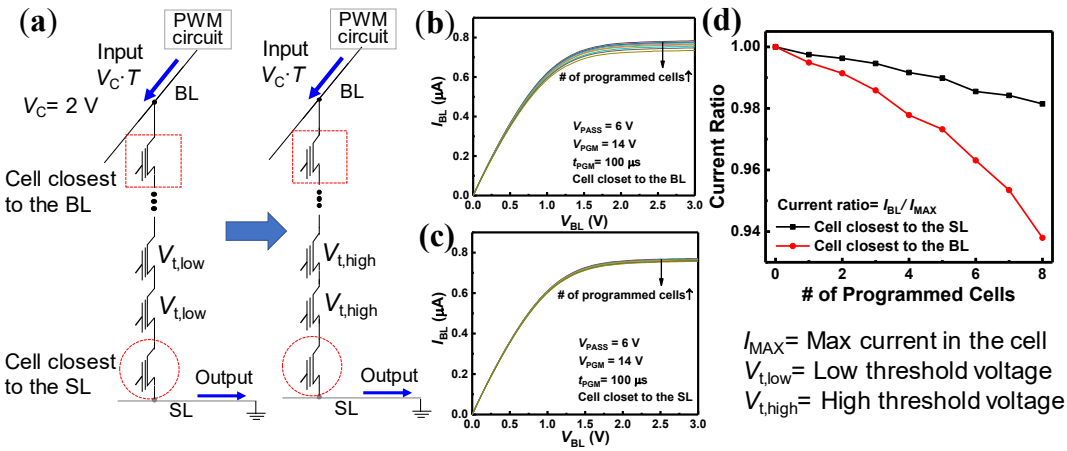

Fig. 4.10. (a) Circuits for measuring I_{BL}s at a V_{BL} of 2 V and V_{SL} of 0 V. The I_{BL}-V_{BL} curves are measured in the cells closest to the (b) BL and (c) SL, with increasing number of programmed cells in one string. (d) Current ratio of I_{BL} and maximum current in the cell (I_{MAX}) with increasing number of programmed cells.

4.3 Neuron circuits

Fig. 4.11 (a) represents the neuron circuit consisting of two current mirrors and one capacitor. The relationship between voltage of the capacitor (V_C) and the difference of current from G^+ and G^- array represents a hard-sigmoid function, which is the activation function. Fig. 4.11 (b) and (c) represent PWM circuit and simulated results of it, respectively. The PWM circuit converts the voltage of capacitor (V_C) to the width-modulated pulse (V_P). The width of output pulse (V_P) has linear relationship with the amplitude of the input pulse (V_C). The PWM circuit replaces analog to digital converter (ADC), which greatly curtail the burden of neuron circuit and power consumption. Note that, in technologically mature and commercial NAND flash memory, read time is 45 μs. This value is optimized read time including precharge/discharge time and considering non-ideality of waveform. In commercial NAND flash memory, a peripheral circuit controls the enormous size of array with 16Kb page size. However, partitioning NAND flash blocks into smaller array can significantly decrease read time compared to the commercial NAND flash memory. It is because WL/SL loading is greatly decreased, which

reduces RC delay time. On the other hand, on-current measured in NAND cells in this work is ~1 μA. However, we can adopt an extremely thin body of 3 nm to provide low on-current of ~2 nA. In addition, current mirrors in neuron circuit can reduce the current which flows into the capacitor of the neuron circuit. Therefore, capacitance of the capacitor can be significantly reduced.

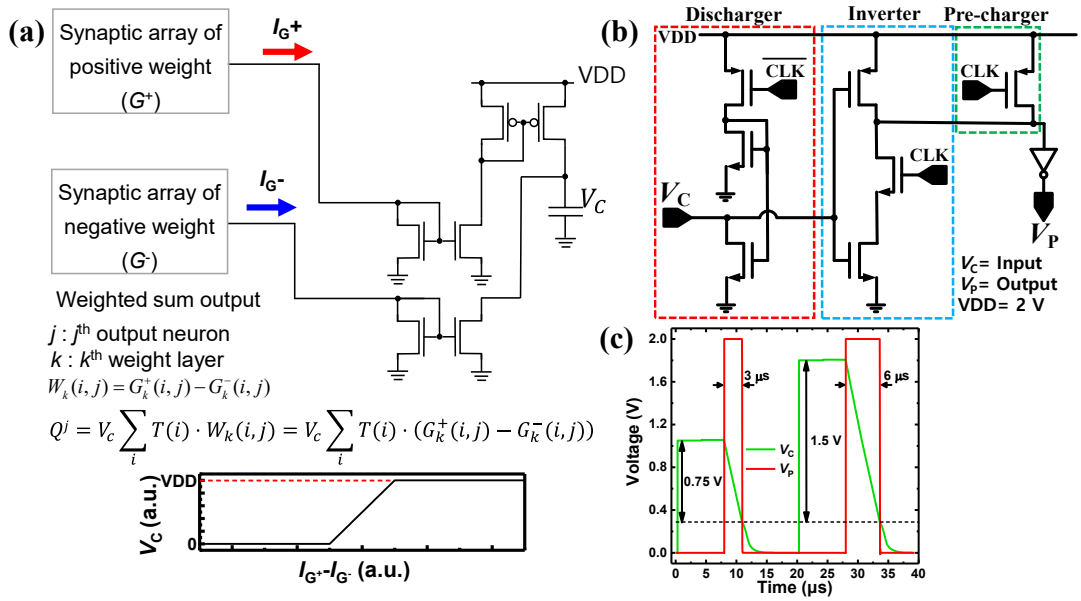

Fig. 4.11. (a) I_{BL}-V_{BL} curves with increasing number of program pulses. (b) Normalized conductance (G) responses measured in (a).

4.4 Simulation results

Fig. 4.12 (a) shows the on-chip learning scheme in this work. Hard-sigmoid function is used in the on-chip learning scheme, as it can be implemented by the simple circuit shown in Fig. 4.11 (a). In software-based algorithm, the amount of weight change (ΔW) is the product of error (δ), activated neuron value, and learning rate. However, if multiple V_{PGM}s are required to update weight of synaptic devices, we need to check the current value of conductance and calculate the number of pulses to reach the target conductance, which imposes a big burden on the peripheral circuitry. Furthermore, on-chip learning system requires a lot of weight updates during training, so read-verify-write (RVW) method which consumes lots of time and energy cannot be used for each weight update. Therefore, in the on-chip learning scheme, the weights are updated based on the sign of ΔW (sgn(ΔW)). The sgn(ΔW) is determined by the sign of error (δ) and sign of activated neuron value. When sgn(ΔW)>0, the G^- is decreased to increase weight. When sgn(ΔW)<0, the G^+ is decreased to decreases weight. In addition, we use the same single program bias (V_{PGM}) in each weight update, greatly reducing the burden on the peripheral

circuit. Fig. 4.12 (b) represents the weight update method. To increase weight, G^- is decreased by a V_{PGM}. However, when G^- is saturated to the minimum conductance (G_{min}), the weight cannot be increased. In this case, the G^+ is initialized to the maximum conductance (G_{max}) and decreased to a target value by applying a series of program pulses sequentially as shown in Fig. 4.12 (b). In initialization process, individual erase scheme is needed and we introduce an erase scheme to enable individual cell erase as an illustration. A high bias is applied to a selected BL and a selected SSL is turned off to provide gate-induced drain leakage (GIDL) to only one selected cell string while a high bias is applied to other SSLs not to generate GIDL in unselected cell strings. In this scheme, 0 V and high bias are applied to a selected WL and unselected WLs, respectively, to make one cell erased, and ground-select line (GSL) is turned off.

Fig. 4.13 (a) represents the inference accuracy of the proposed on-chip learning system for the MNIST dataset on fully connected neural networks which consists of 3 neuron layers (784-200-10) with 6-bit weight precision. The conductance response in Fig. 4.6 (b) at V_{WL} of from V_1 to V_5 is used. The inset shows

the final inference accuracy after 100 epochs. Linearity of conductance response improves as the V_{WL} increases from V_1 to V_5, which increases inference accuracy. The conductance response at a V_{WL} of V_3 is used to obtain a G_{MAX} / G_{MIN} of ~5.1. The final accuracy is 95.58 % at a V_{WL} of V_3, which is comparable to 95.81 % obtained with ideal linear conductance response.

Fig. 4.13 (b) shows the inference accuracy with respect to the number of hidden layers. The inference accuracies are 95.58 %, 95.65 % and 95.56 % when the number of hidden layers is 1, 3 and 5, respectively. Therefore, the proposed scheme can be applied to deeper neural networks. The accuracy increases as the number of hidden layers increases up to 3. On the other hand, the accuracy does not increase, when the number of hidden layers increases from 3 to 5, because regularization method does not used in this work. It needs proper regularization method to improve accuracy in deep networks. However, it is beyond scope of this work, and there are recent works which propose regularization method in neuromorphic system. If the suitable regularization methods are applied, the proposed scheme can achieve higher accuracy with deep networks.

Fig. 4.14 (a) shows the effect of synaptic weight variation (σ_G/μ_G : standard deviation / mean) on the inference accuracy in off-chip learning and on-chip learning. As σ_G/μ_G increases from 0 to 1, the inference accuracy decreases by 0.7 % in on-chip learning while 21.2 % in off-chip learning. Fig. 4.14 (b) shows the effect of cycle-to-cycle variation on the inference accuracy in on-chip learning. The variation is applied to the change of device conductance whenever an V_{PGM} is applied. The accuracy of the on-chip training system hardly changes even if σ/μ increases to 1.

To investigate whether on-chip learning is possible in the proposed architecture, circuit simulations were performed using SPICE as shown in Fig. 4.15. The reduced MNIST (8×8) data is trained in fully connected neural networks (64-64-10). The NAND flash memory architecture with 2 WL layers is constructed to have 16,384 synapses. NAND cells are modeled to have a conductance response of Fig. 4.6 (b). The NAND flash memory architecture is trained using 1700 training images. The designed circuit that performs the activation function and PWM can be considered as a neuron circuit. The inset shows final inference accuracy after 8

epochs. As shown in Fig. 4.15, the inference accuracy of the on-chip learning system obtained from circuit simulation is 94.99 %, which is comparable to 95.604 % obtained from system-level simulation. Furthermore, during on-chip learning process, the neural networks correct the deviation caused by pass cells through backward propagation. In Fig. 4.15, the difference between accuracy of system-level simulation and SPICE simulation decreases as the number of epochs increases.

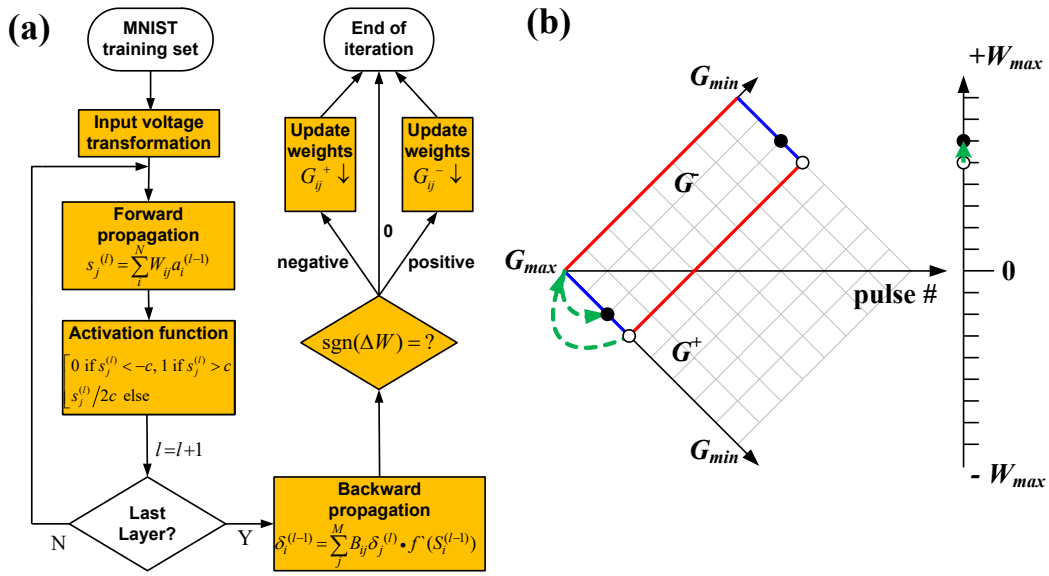

Fig. 4.12. (a) On-chip learning scheme. (b) Weight update method.

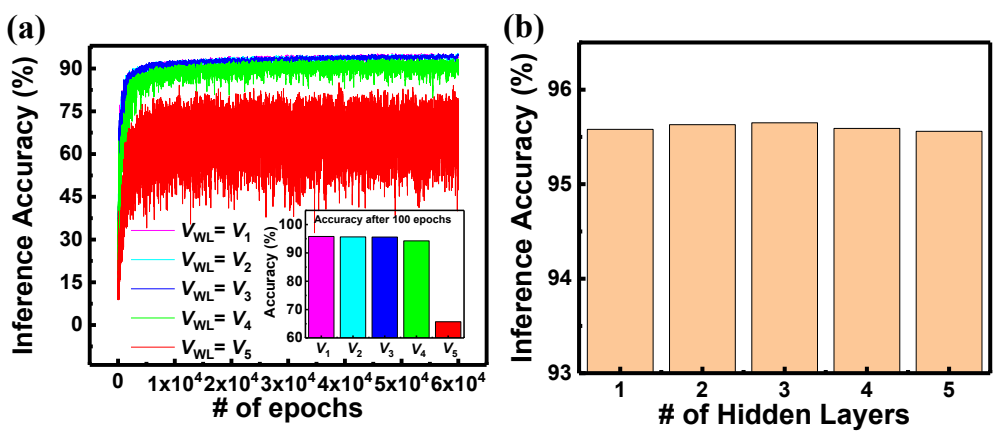

Fig. 4.13. (a) Inference accuracy of the proposed on-chip learning system for MNIST dataset using the *G* response in Fig. 4.6 (b). (b) Inference accuracy with the number of hidden layers.

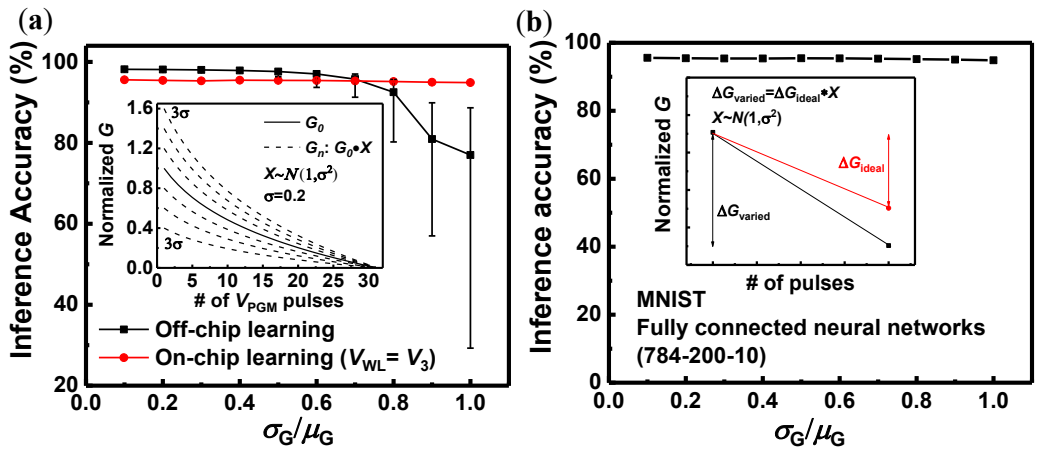

Fig. 4.14. (a) Inference accuracy with synaptic weight variation (σ_G/μ_G). (b) Inference accuracy with cycle-to-cycle variation.

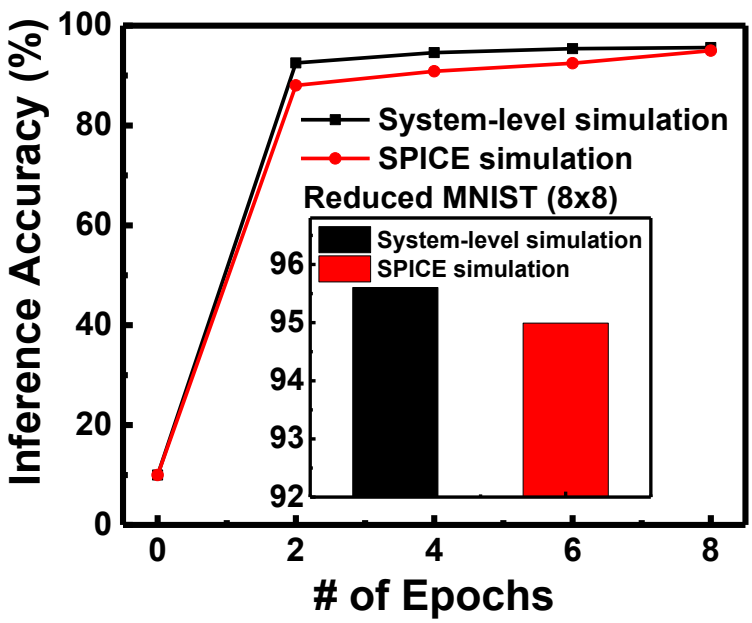

Fig. 4.15. Inference accuracy of the proposed on-chip learning system for reduced MNIST dataset in fully connected neural networks (64-64-10).

Chapter 5
Conclusion

In this work, we have proposed a novel synaptic architecture based on NAND cell strings and an operation scheme for high-density and highly-reliable BNN, for the first time. Reliability has been verified from retention, endurance, and V_{pass} disturbance measurement results. The bit-error rate (4.2×10^{-8} %) of the proposed synapses was 4 orders of magnitude lower than that of RRAMs. When V_{PGM} is below 16V, single pulse is enough to write a weight without ISPP method. In 128-stack NAND flash memory, the estimated synapse density is ~100 times higher than that of RRAMs. Thus, the proposed architecture is very promising for high-density and highly-reliable BNNs.

A novel synaptic string architecture based on a NAND flash memory for highly robust and high-density quantized neural networks (QNN) with binary neuron activation was proposed, for the first time. The differential sensing scheme and neuron activation of (1, 0) instead of (1, -1) are appropriately compatible with

the conventional NAND flash memory architecture consisting of cell strings. Moreover, a binary neuron enables adopting the 1-bit sense amplifier instead of the multi-level sense amplifier or analog-to-digital converter (ADC), which serves as the area-saving and energy-efficient peripheral circuits enabling bitwise communication between the layers of the neural networks. Operating NAND cells in the saturation region achieved a high synaptic resistance (24 MΩ) eliminating the effect of the IR drop in the metal wire and the serial resistance of the pass cells in the NAND cell string. By using a read-verify-write scheme, a low-variance conductance distribution was demonstrated for 8 levels, where the maximum device variation (σ_w/μ_w) of the 8 levels is about 3.04 %. Vector-matrix multiplication (VMM) of 4-bit weight and binary activation could be accomplished by only one input pulse eliminating the need of multiplier and additional logic operations. High on/off current ratio (>10^5) and small I_{off} (<10pA) of the NAND cells can implement a high bandwidth to sum the currents in parallel from many cells in a synaptic string array. In addition, quantization training reduced the degradation of the inference accuracy by 1.24 % and 0.3 % for the CIFAR 10 and MNIST datasets, respectively,

compared to the PTQ. Finally, the low-variance conductance distribution (~3.04 %) of the NAND cells achieves a higher inference accuracy compared to RRAM devices.

We have proposed a novel synaptic architecture based on NAND flash memory for on-chip learning. By separating SL and applying input and error input to BL and SSL in NAND cell array, accurate vector-matrix multiplication is successfully performed in both FP and BP eliminating the effect of pass cells. In addition, weight update method in which a V_{PGM} is sequentially applied from the cell closest to the SL to the cell closest to the BL was proposed to eliminate the effect of cell position on the string current. At an optimized read bias of 2 V, a satisfying accuracy of 95.58 % is achieved which is comparable to that of 95.81 % obtained with perfect linear device. It was shown that the inference accuracy decreases by 0.7 % in on-chip learning while 21.2 % in off-chip learning when weight variation (σ_G/μ_G) increases from 0 to 1. In addition, the superiority of the proposed on-chip learning architecture was verified in circuit simulation. The proposed synaptic architecture based on a technologically mature NAND flash

memory in this work has great advantages when implementing a high-density and highly reliable on-chip learning system.

Bibliography

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015. doi: 10.1038/nature14539

[2] J. Fan et al., "Human tracking using convolutional neural networks." *IEEE Transactions on Neural Networks*, vol. 21, no. 10, pp. 1610–1623, 2010. doi**:** 10.1109/TNN.2010.2066286

[3] A. Karpathy and L. Fei-Fei., "Deep visual-semantic alignments for generating image descriptions," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128-3137, 2015. doi: 10.1109/CVPR.2015.72 98932

[4] K. He, X. Zhang, S. Ren, and J. Su, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. doi: 10.1109/CVPR.2016.90

[5] J. Gu, M. Zhu, Z. Zhou, F. Zhang, Z. Lin, Q. Zhang, and M. Breternitz, "Implementation and evaluation of deep neural networks (DNN) on mainstream heterogeneous system," *Proceedings of 5th Asia-Pacific Workshop on Systems*, 2014. doi: 10.1145/2637166.2637229

[6] M. Hu, *et al*., "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," In *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1-6, 2016.

[7] R. Andri, L. Cavigelli, D. Rossi, and L. Benini, "YodaNN: An Architecture for Ultralow Power Binary-Weight CNN Acceleration," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 1, pp. 48-60, 2018. doi: 10.1109/TCAD.2017.2682138

[8] S. Yu, “Neuro-inspired computing with emerging nonvolatile memorys,” *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260-285, 2018. doi: 10.1109/JPROC.2018.2790840

[9] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized neural networks: Training neural networks with low precision weights and activations,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869-6898, 2017.

[10] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, “Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients,” *arXiv preprint arXiv:1606.06160*, 2016.

[11] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1,” arXiv:1602.02830 [cs], Mar. 2016.

[12] S. Yu, Z. Li, P. Y. Chen, H. Wu, B. Gao, D. Wang, W. Wu, H. Qian, “Binary neural network with 16 Mb RRAM macro chip for classification and online training,” *in IEEE Int. Electron Devices Meeting (IEDM)*, 2016. doi: 10.1109/IEDM.2016.7838429

[13] M. Bocquet, T. Hirztlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal and D. Querlioz, “In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks” *in IEEE Int. Electron Devices Meeting (IEDM)*, 2018. doi: 10.1109/iedm.2018. 8614639

[14] L. Ni, Y. Wang, H. Yu, W. Yang, C. Weng, and J. Zhao, “An energy-efficient matrix multiplication accelerator by distributed in-memory computing on binary RRAM crossbar,” in *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 280-285), 2016.

[15] M. Hu, *et al*., "Memristor-based analog computation and neural network classification with a dot product engine," *Advanced Materials*, vol. 30, no. 9, 2018.

[16] J. Woo, X. Peng, and S. Yu, "Design considerations of selector device in cross-point RRAM array for neuromorphic computing," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2018. doi: 10.1109/ISCAS.2018.8351735

[17] B. Liu, *et al*., "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 63-70, 2014.

[18] C.-P. Lo, et al., "Embedded 2Mb ReRAM macro with 2.6ns read access time using dynamic-trip-point-mismatch sampling current-mode sense amplifier for IoE applications", in *Symp. VLSI Circuits*, 2017.

[19] X. Sun, S. Yin, X. Peng, R. Liu, J.-S. Seo, and S. Yu, "XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks," in *Proc. IEEE/ ACM Design Autom. Test Eur. (DATE)*, 2018.

[20] S.-H. Lee "Technology Scaling Challenges and Opportunities of Memory Devices," *in IEEE Int. Electron Devices Meeting (IEDM), Dec. 2016*. doi: 10.1109/IEDM.2016.7838026

[21] D. Kang, W. Jeong, C. Kim, D. Kim, Y. Cho, K. Kang, J. Ryu, K. Kang, S. Lee, W. Kim, H. Lee, J. Yu, N. Choi, D. Jang, J. Ihm, D. Kim, Y. Min, M. Kim, A. Park, J. Son, I. Kim, P. Kwak, B. Jung, D. Lee, H. Kim, H. Yang, D. Byeon, K. Park, K. Kyung, J. Choi, "256 Gb 3b/cell V-NAND flash memory with 48 stacked WL layers," *in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 130–131, Jan./Feb. 2016. doi: 10.1109/JSSC.2016.2604297.

[22] C. Kim, J. Cho, W. Jeong, I. Park, H. Park, D. Kim, D. Kang, S. Lee, J. Lee, W. Kim, J. Park, Y. Ahn, J. Lee, J. Lee, S. Kim, H. Yoon, J. Yu, N. Choi, Y. Kwon, N. Kim, H. Jang, J. Park, S. Song, Y. Park, J. Bang, S. Hong, B. Jeong, H. Kim, C.

Lee, Y. Min, I. Lee, I. Kim, S. Kim, D. Yoon, K. Kim, Y. Choi, M. Kim, H. Kim, P. Kwak, J. Ihm, D. Byeon, J. Lee, K. Park, K. Kyung, "A 512 Gb 3b/cell 64-stacked WL 3D V-NAND flash memory," *in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, pp. 202–203, Feb. 2017. doi: 10.1109/JSSC.2017.2731813.

[23] LO, Chieh-Pu, *et al*., "Embedded 2Mb ReRAM macro with 2.6 ns read access time using dynamic-trip-point-mismatch sampling current-mode sense amplifier for IoE applications," In: *2017 Symposium on VLSI Circuits*. IEEE, 2017. p. C164-C165.

[24] J. P. Duarte, S. Khandelwal, A. Medury, C. Hu, P. Kushwaha, H. Agarwal, A. Dasgupta, and Y. S. Chauhan, "BSIM-CMG: Standard FinFET compact model for advanced circuit design," *in IEEE 2015-41st European Solid-State Circuits Conference (ESSCIRC)*, pp. 196-201, Sep. 2015. doi: 10.1109/ESSCIRC.2015.7313862

[25] R. Chen, Z. Qin, Y. Wang, D. Liu, Z. Shao, & Y. Guan, (2014). On-demand block-level address mapping in large-scale NAND flash storage systems. *IEEE Transactions on Computers*, *64*(6), 1729-1741. doi: 10.1109/TC.2014.2329680

[26] C. Ma, Y. Wang, Z. Shen, R. Chen, Z. Wang, & Z. Shao, (2020). MNFTL: An Efficient Flash Translation Layer for MLC NAND Flash Memory. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, *25*(6), 1-19. doi: 10.1145/3398037

[27] R. Chen, C. Zhang, Y. Wang, Z. Shen, D. Liu, Z. Shao, & Y. Guan, (2018). DCR: Deterministic crash recovery for NAND flash storage systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *38*(12), 2201-2214. doi: 10.1109/TCAD.2018.2878179

[28] B. Li, *et al*., "Merging the interface: Power, area and accuracy co-optimization for rram crossbar-based mixed-signal computing system," in *Proceedings of the 52nd Annual Design Automation Conference*, pp. 1-6, 2015.

[29] Lee, S. T., & Lee, J. H, "Neuromorphic Computing Using NAND Flash Memory Architecture With Pulse Width Modulation Scheme," *Frontiers in Neuroscience*, *14*, 2020.

[30] S. T. Lee, S. Lim, N. Choi, J. H. Bae, C. H. Kim, S. Lee, D. Lee, T. Lee, S. Chung, B. G. Park, and J. H. Lee, "Neuromorphic Technology Based on Charge Storage Memory Devices," *in 2018 IEEE Symposium on VLSI Technology*, pp. 169-170, Jun. 2018. doi: 10.1109/VLSIT.2018.8510667.

[31] S. T. Lee, *et al*., "High-Density and Highly-Reliable Binary Neural Networks Using NAND Flash Memory Cells as Synaptic Devices," in *2019 IEEE International Electron Devices Meeting (IEDM)*, pp. 38-4, 2019.

[32] Y. Cai, E. F. Haratsch, O. Mutlu, and K. Mai, "Threshold voltage distribution in MLC NAND flash memory: Characterization, analysis, and modeling," in *2013 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1285-1290, 2013.

[33] S. T. Lee, *et al*., "Operation Scheme of Multi-Layer Neural Networks Using NAND Flash Memory as High-Density Synaptic Devices," *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 1085-1093, 2019.

[34] Lee, S. T., Kwon, D., Kim, H., Yoo, H., & Lee, J. H, "NAND Flash Based Novel Synaptic Architecture for Highly Robust and High-Density Quantized Neural Networks With Binary Neuron Activation of (1, 0)," *IEEE Access*, *8*, 114330-114339, 2020.

[35] LEE, S.-T., *et al*., “Novel Method Enabling Forward and Backward Propagations in nand Flash Memory for On-Chip Learning,” *IEEE Transactions on Electron Devices*, 2021.

[36] Y. Li, *et al*., "Build a compact binary neural network through bit-level sensitivity and data pruning." *Neurocomputing*, 2020.

[37] P. Kapoor, *et al*., "Computation-Efficient Quantization Method for Deep Neural Networks." 2018.

[38] J. H. Lee, *et al*., “Review of candidate devices for neuromorphic applications,” in *ESSDERC 2019-49th European Solid-State Device Research Conference (ESSDERC)*, pp. 22-27, 2019.

[39] H. Y. Chen, ‘‘HfOx based vertical resistive random access memory for cost-effective 3D cross-point architecture without cell selector,’’ in IEDM Tech. Dig., Jun. 2012, pp. 7–20.

[40] H. W. Pan, K. P. Huang, S. Y. Chen, P. C. Peng, Z. S. Yang, C.-H. Kuo, Y.-D. Chih, Y.-C. King, and C. J. Lin, ‘‘1Kbit FinFET dielectric (FIND) RRAM in pure 16nm FinFET CMOS logic process,’’ in IEDM Tech. Dig., Dec. 2015, pp. 5–10.

[41] G. Molas, G. Piccolboni, M. Barci, B. Traore, J. Guy, G. Palma, E. Vianello, P. Blaise, J. M. Portal, M. Bocquet, A. Levisse, B. Giraud, J. P. Noel, M. Harrand, M. Bernard, A. Roule, B. De Salvo, and L. Perniola, ‘‘Functionality and reliability of resistive RAM (RRAM) for non-volatile memory applications,’’ in Proc. Int. Symp. VLSI Technol., Syst. Appl. (VLSI-TSA), Apr. 2016, pp. 1–2.

[42] D. Soudry, D. Di Castro, A. Gal, A. Kolodny, and S. Kvatinsky, “Memristor-based multilayer neural networks with online gradient descent training,” *IEEE transactions on neural networks and learning systems*, vol. 26, pp. 2408-2421, 2015,

doi: 10.1109/tnnls.2014.2383395.

[43] S. Lim *et al*., "Adaptive learning rule for hardware-based deep neural networks using electronic synapse devices," *Neural Computing and Applications*, vol. 31, pp. 8101-8116, 2019, doi: 10.1007/s00521-018-3659-y.

초 록

DNN에서 중요한 작업인 벡터-매트릭스 곱셈 (VMM)을 효율적으로 처리하기 위해 시냅스 소자를 사용하는 뉴로모픽 컴퓨팅이 활발히 연구되고 있다. 지금까지 RRAM (Resistive RAM)이 주로 뉴로모픽 컴퓨팅의 시냅스 소자로 사용되었다. 그러나 RRAM은 소자의 산포가 크고 신뢰성이 좋지 않으며 CMOS 주변 회로와 통합이 어려운 문제로 인해 대규모 시냅스 소자 어레이를 구현하는 데는 여전히 많은 제한이 있다. 이러한 문제로 인해 성숙한 실리콘 메모리인 SRAM 셀이 시냅스 소자로 제안되고 있다. 그러나 SRAM은 셀 당 면적 (~150 F^2 per bitcell)이 크고 또한 온칩 SRAM 용량 (~a few MB) 은 많은 파라미터를 수용하기에 충분하지 않다.

본 논문에서는 오프 칩 학습과 온 칩 학습을 위해 NAND 플래시 셀 스트링을 기반으로 하는 시냅스 아키텍처를 제안한다. NAND 셀 스트링 기반의 새로운 시냅스 아키텍처는 오프 칩 학습에서 이진 신경망 (BNN)을 위한 XNOR 연산이 가능한 고밀도 시냅스로 사용된다. 상호 보완적인 방식으로 NAND 플래시 셀의 임계 전압과 입력 전압을 변경함으로써 XNOR 연산을 성공적으로 수행한다. NAND 플래시 셀의 큰 온/오프 전류 비율(~ $7x10^5$)은 ECC 없이 고밀도 및 고신뢰성의 BNN을 구현할 수 있다. 우리는 4비트 가중치를 갖는 매우 견고하며 고집적의 양자화된 신경망(QNN)을 위한 NAND 플래시 메모리를 기반의 새로운 시냅틱 아키텍처를 제안한다. 양자화 학습은 훈련 후 양자화에 비해 추론 정확

도의 저하를 최소화할 수 있다. 제안하는 동작 방식은 BNN에 비해 더 높은 추론 정확도를 가지는 QNN을 구현할 수 있다.

온 칩 학습은 훈련 중 시간과 에너지 소비를 크게 줄이고 시냅스 소자의 산포를 보상하며 변화하는 환경에 실시간으로 적응할 수 있다. NAND 플래시 메모리 구조의 높은 집적도를 사용한 온 칩 학습은 매우 유용하다. 그러나 기존의 RRAM 어레이에 사용되는 온 칩 학습 방법은 NAND 플래시 메모리의 셀 스트링 구조로 인해 NAND 플래시 셀을 시냅스 소자로 사용하는 경우 활용할 수 없다. 이 연구에서는 온 칩 학습을 위해 NAND 플래시 메모리에서 순방향 전파 (FP) 및 역방향 전파 (BP)를 가능하게 하는 새로운 시냅스 어레이 아키텍처를 제안한다. 제안된 시냅스 아키텍처에서는 가중치가 올바르게 전치될 수 있도록 양의 시냅스 가중치와 음의 시냅스 가중치가 서로 다른 어레이로 분리된다. 또한 기존 NAND 플래시 메모리와 달리 소스 라인 (SL)을 분리하여 NAND 플래시 메모리에서 순방향 전파와 역방향 전파를 모두 연산할 수 있다. NAND 셀 어레이의 비트 라인 (BL) 및 스트링 선택 라인 (SSL)에 각각 입력 및 오류 입력을 인가함으로써 PASS 셀의 효과를 제거하여 순방향 전파 및 역박향 전파 모두에서 정확한 벡터 행렬 곱셈이 성공적으로 수행되도록 한다. 제안된 온 칩 학습 시스템은 오프 칩 학습 시스템에 비해 소자의 산포에 대해 훨씬 영향이 적다. 마지막으로, 제안된 온 칩 학습 아키텍처의 우수성을 신경망의 회로 시뮬레이션을 통해 검증하였다.

주요어 : 하드웨어 기반 신경망, NAND 플래시 메모리, 신경 모방 시스

템, 메모리내 연산, 이진 신경망, 온칩 학습.

학번 : 2016-20951