Electronic Supplementary Material (ESI) for Nanoscale.

Supporting information for

Parallel Weight Update Protocol for a Carbon Nanotube Synaptic Transistor Array for Accelerating Neuromorphic Computing

Sungho Kim[1], Yongwoo Lee[2], Hee-Dong Kim[1], and Sung-Jin Choi[2,*]

[1]Department of Electrical Engineering, Sejong University, Seoul 05006, Korea

[2]School of Electrical Engineering, Kookmin University, Seoul 02707, Korea

*Correspondence to S. J. C. (sjchoiee@kookmin.ac.kr).

1. The detail of CNT transistor

1-1. Hysteresis and consequent channel conductance modulation in CNT synaptic transistors

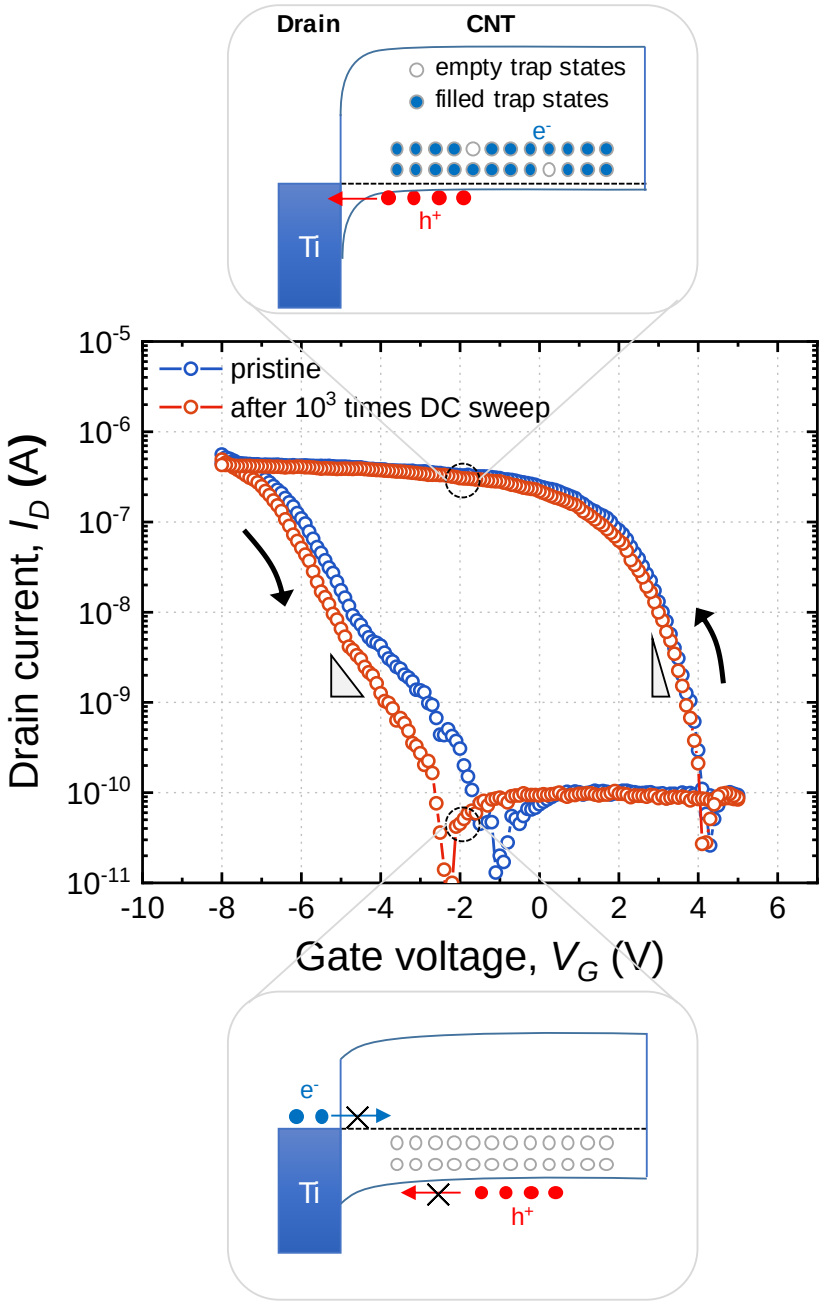

Figure S1. Hysteresis of the drain current (I_D) as a function of the gate voltage (V_G) with a constant drain voltage (V_D = -1 V).

The synaptic weight can be reproduced by an intrinsic analog state of the channel conductance in the CNT synaptic transistor. A hysteresis in the drain current (I_D) under a constant drain voltage (V_D) was observed when the gate voltage (V_G) was swept from -8 V to +5 V and back to -8 V (Fig. S1); a positive V_G increased the channel conductance, which is defined as the 'potentiation' of the synaptic weight, and a negative V_G decreased the conductance, which is defined as the 'depression' of the synaptic weight. As a result, the direction of the hysteresis is counterclockwise.

The physical mechanism of hysteresis in CNT transistor has been studied intensively, and it is generally accepted that the CNT–OH complex located at the CNT/SiO_2 interface acts as an electron acceptor (trap). Fig. S1 explains the hysteresis of a CNT transistor with electron traps when the work function of the metal electrode is lower than that of the CNTs, *e.g.*, Al, Ti, and Cr. Initially, due to the high trap density, the Fermi level of CNT is downshifted to the valence band. When $V_G = 0$, the current cannot flow because due to the Schottky barrier for both hole and electron. However, when $V_G > 0$ is applied for the potentiation, the electrons in the channel can be trapped by the empty trap states. These trapped electrons can bend the energy band upward and consequently narrow the Schottky barrier width at the junction of the drain/CNT. Therefore, hole tunneling current can be increased, and consequently, the channel conductance (G) is increased. By contrast, $V_G < 0$ results in the detrapping process of the electron, and leads to decreasing of G. This trapping/detrapping process of the electrons provides internal dynamics that drive the analog channel conductance switching behavior.

Note that the slope of I_D change when V_G is sweeping from +5V to -8V is steeper than the slope when V_G is sweeping from -8V to +5V. The reason for this difference is as follows. When

V_G is sweeping from + 5V to -8V, the most of traps are filled by electrons, and the Schottky barrier width is already narrowed by the band-bending. Accordingly, the effect of the hole tunneling process on V_G can be enhanced, which leads to a drastic change in I_D. In contrast, when V_G is sweeping from -8V to +5V, the Schottky barrier width is already getting thicker because the electrons are ejected from the trap. Consequently, the hole tunneling process cannot occur and the barrier width control by V_G is not effective either.

1-2. Device variability

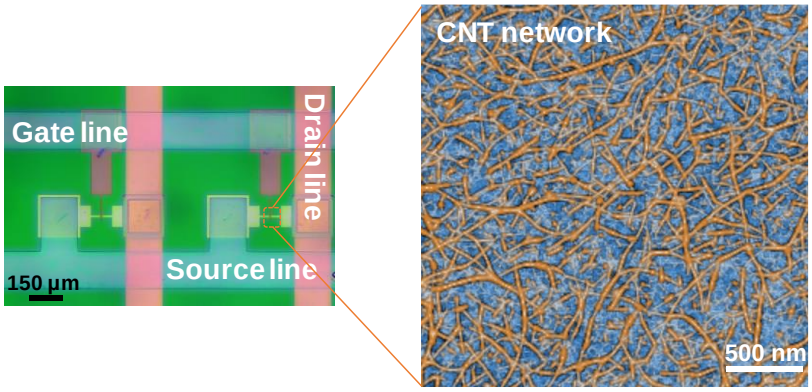

Figure S2. The optical microscope and atomic force microscope images of our CNT transistor (copied from Ref. 20)

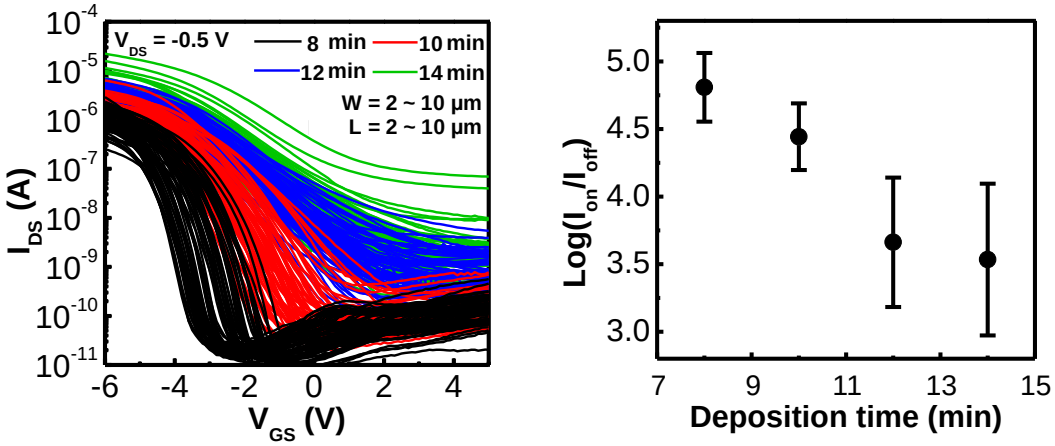

Figure S3. Transfer characteristics (drain current, I_D versus gate voltage, V_G) and summarized on/off ratio (I_{on}/I_{off}) of four sets of devices (total 288 devices) produced with different CNT deposition times of 8 min, 10 min, 12 min, and 14 min (copied from Ref. 18).

Fig. S2 shows the fabricated CNT transistor with a pre-separated, semiconducting enriched single-walled CNT solution (99%) processed by a density gradient ultracentrifuge separation method. The device-to-device variability arises from the variation in the number of connecting paths between the source and drain electrodes. Fig. S3 shows the transfer characteristics of a total of 288 devices in four different sets (8 min, 10 min, 12 min, and 14 min) for the deposition time of solution-processed 99% semiconducting CNTs. It is obvious that as the CNT deposition time was increased from 8 min to 14 min, the average $\log(I_{on}/I_{off})$ decreased from 4.81 to 3.53. We correspondingly select the short 8 min deposition time for a CNT transistor because it has a significantly higher on/off ratio and minimized variability, which can provide reliable neuromorphic computing.

2. The parallel weight update/read process in the crossbar array

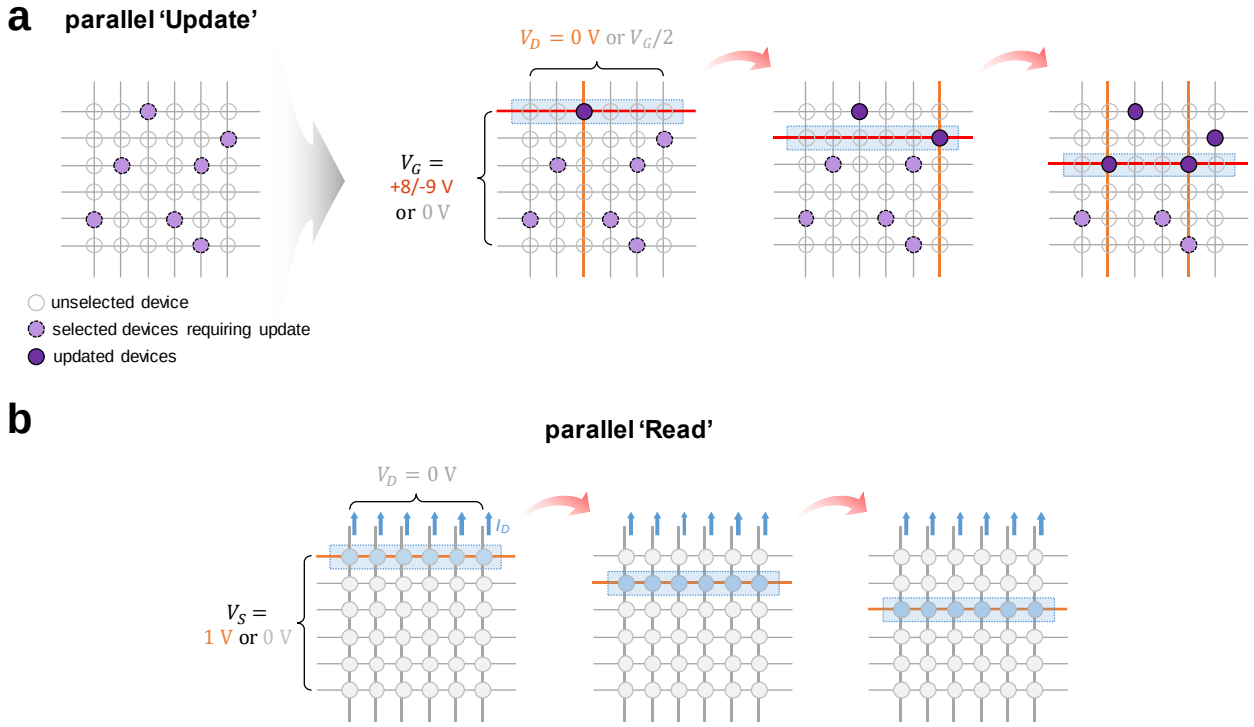

Figure S4. The schematics of the parallel (a) update and (b) read processes respectively.

The most important feature of the proposed weight update protocol in this study is selective and parallel accessing in a row-by-row. For the parallel update process as shown in Fig. S4a, the selected drain and gate lines are set to 0 V and V_W respectively. Whereas, unselected train and gate lines are set to $V_G/2$ and 0 V respectively (all source lines are set to $V_G/2$). Here, multiple drain lines can be selected at the same time, but only one gate line should be selected sequentially; in this way, the weight can be selectively updated without a disturbance from neighboring devices. As a result, in the case of $M \times N$ sized array, the channel conductance of the all synaptic device (G) in the array can be updated at least once during M times update cycles.

A similar approach is applicable to the parallel read process as shown in Fig. S4b. All drain lines are set to 0 V to measure each column current, and the selected source line is sequentially set to 1 V. In addition, all gate lines are set to -1V which should be small enough to not change the channel conductance). Since the source and drain of the unselected device are equally set to 0 V, the unwanted current can not flow through the unselected device, and thus the sneaky current path problem that may occur during the read operation out can be avoided clearly. Consequently, I_D measured in each column allows reading G of all devices in the selected row at a time; all G in the array can be read at least once during M times cycles.

Note that the total number of required update/read cycles in the proposed weight update protocol is only depended on the number of rows (M) in the array. The use of the array with a smaller number of rows can lead to more fast weight accessing, which enables more energy-efficient neuromorphic computing.

3. The performance of analog tuning of *G*

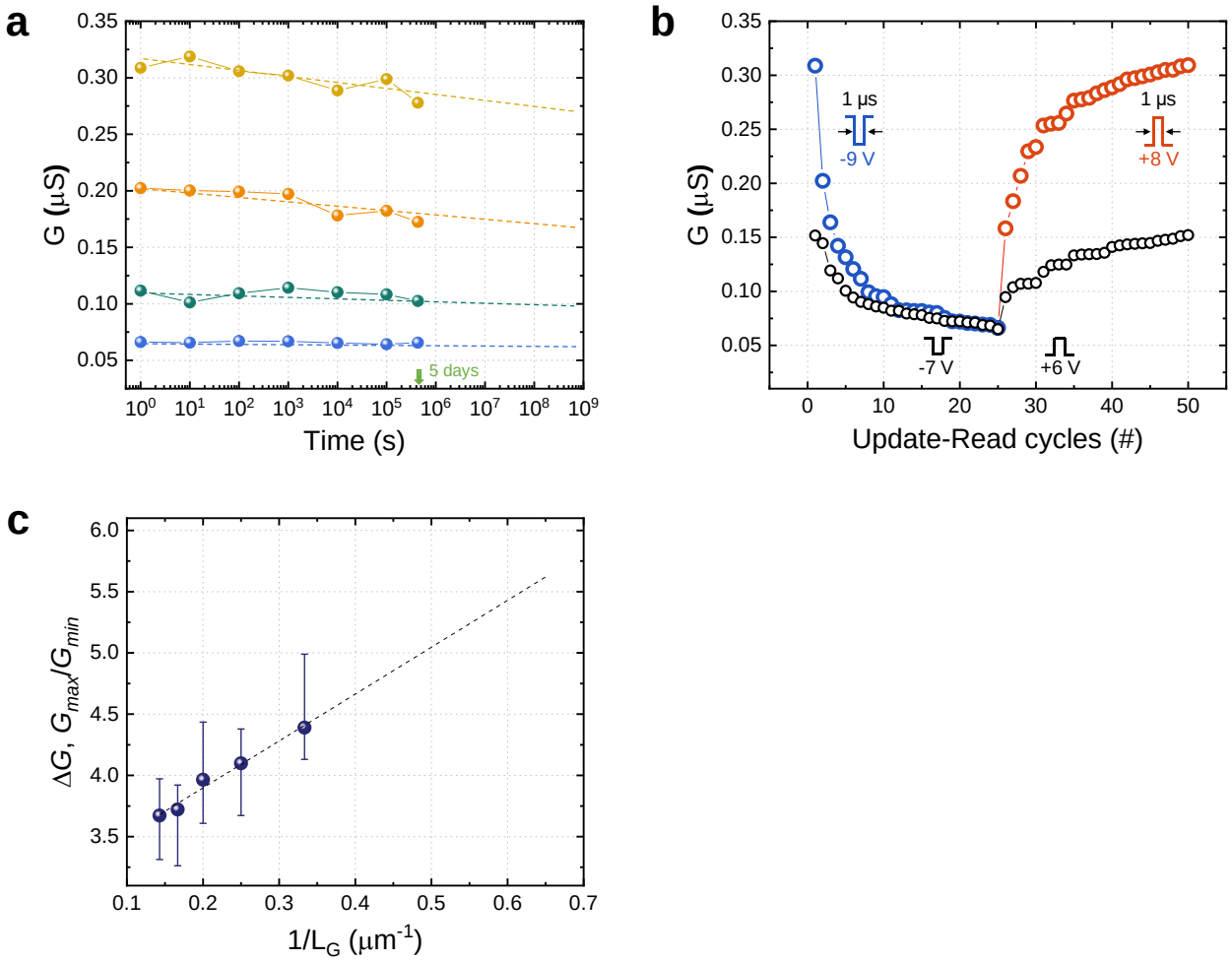

Figure S5. (a) The retention property of the analog *G* states. (b) The analog *G* modulation behavior at two different update voltages. (c) The gate length dependency of Δ*G*.

In this section, we discuss the other properties of the analog tuning of *G* not covered in the main text. Fig. S5a shows the retention characteristics for four arbitrarily selected analog *G* states. The stable weight state can be maintained even for 5 days. In addition, Fig. S5b shows *G* modulation behavior at two different update voltages. An update voltage of V_W = -7 and +6 V results in analog tuning through 25 conductance states, whereas programming with V_W = -9 and

+8 V achieves the same number of states but with higher (2×) modulation. Similarly, hundreds of conductance states can be achieved by adjusting the update pulse duration and/or amplitude. Interestingly, the conductance variation margin ($\Delta G = G_{max}/G_{min}$) can be enhanced by the scaling of the gate length (L_G) as shown in Fig. S5c; ΔG is inversely proportional to the L_G. Unlike the case where it is difficult to modulate ΔG in the conventional memristor, the desired ΔG of the CNT synaptic transistor can be obtained only by the gate length scaling. It is one of the important advantage of 3-terminal synaptic transistor compared to the 2-terminal memristors. Consequently, the CNT synaptic transistor can achieve more analog G states by enlarging ΔG, it will helpful for more accurate neuromorphic computing.

4. The experimental setup for opeationg the synaptic crossbar array

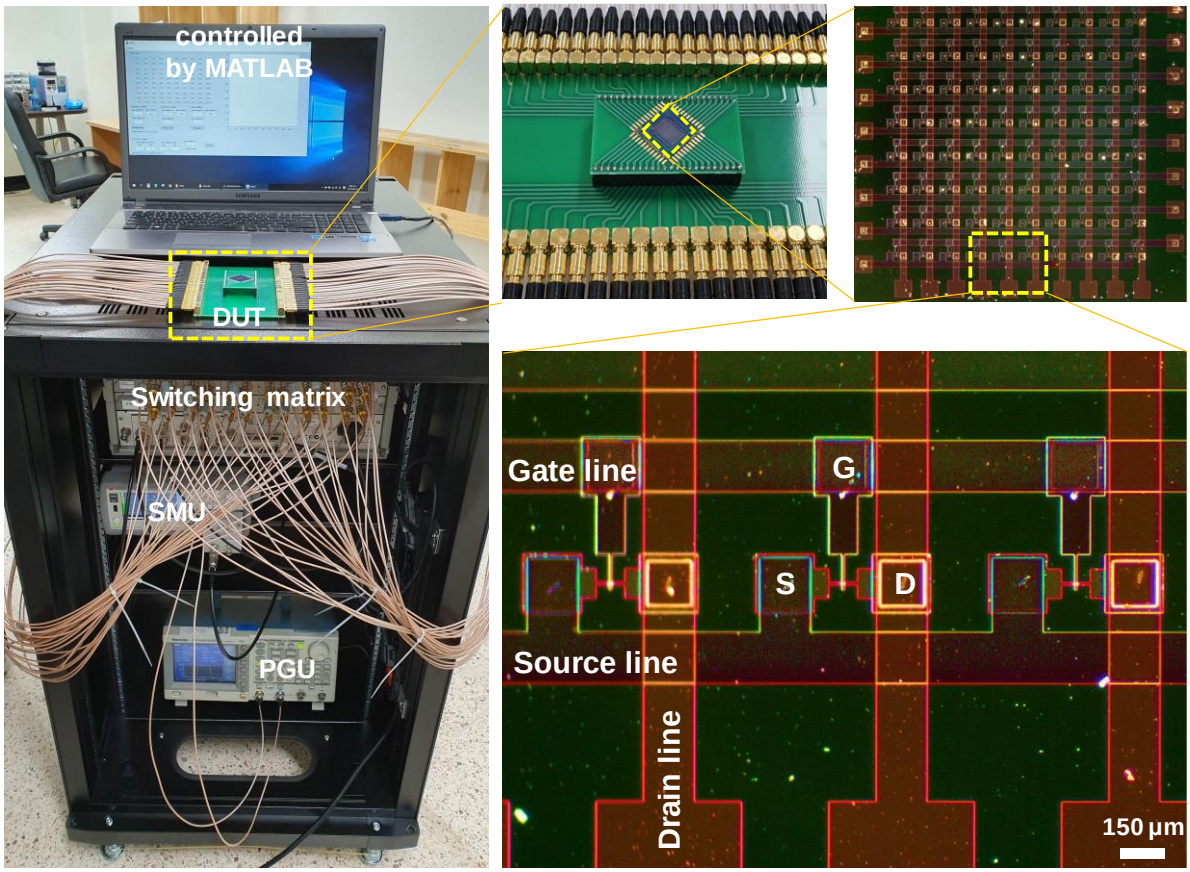

Figure S6. Photos of the measurement system and the fabricated synaptic transistor array.

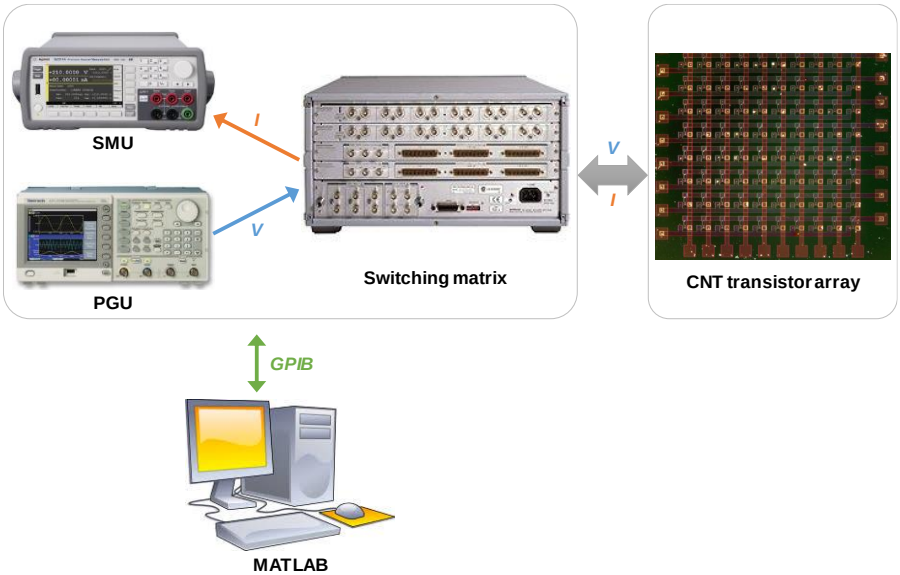

Figure S7. Schematic of measurement system.

Fig. S6 shows the images of experimental setup for operating the synaptic crossbar array. The fabricated synaptic crossbar array is connected to the test board through wire bonding. Actually, 10 × 10 sized array was fabricated, but in this study only 9 × 8 array was partially used for the convolution operation. As shown in Fig. S7, when the application of a pulse signal to the test board is required, the pulse generated by PGU (pulse generation unit, AFG3102) is applied to the crossbar array through the switching matrix (Keysight 5250A). Then, the amount of output current is measured by SMU (source measurement unit, Keysight B2902A). The control of all the measurement equipment is carried out via a home-made MATLAB program. Input image data processing for the convolution operation was also performed by MATLAB program. The data processing procedure for obtaining the data shown in Fig. 4c is discussed in Supporting Information Note 5-2.

5. The data processing procedures

5-1. The update-verify feedback method

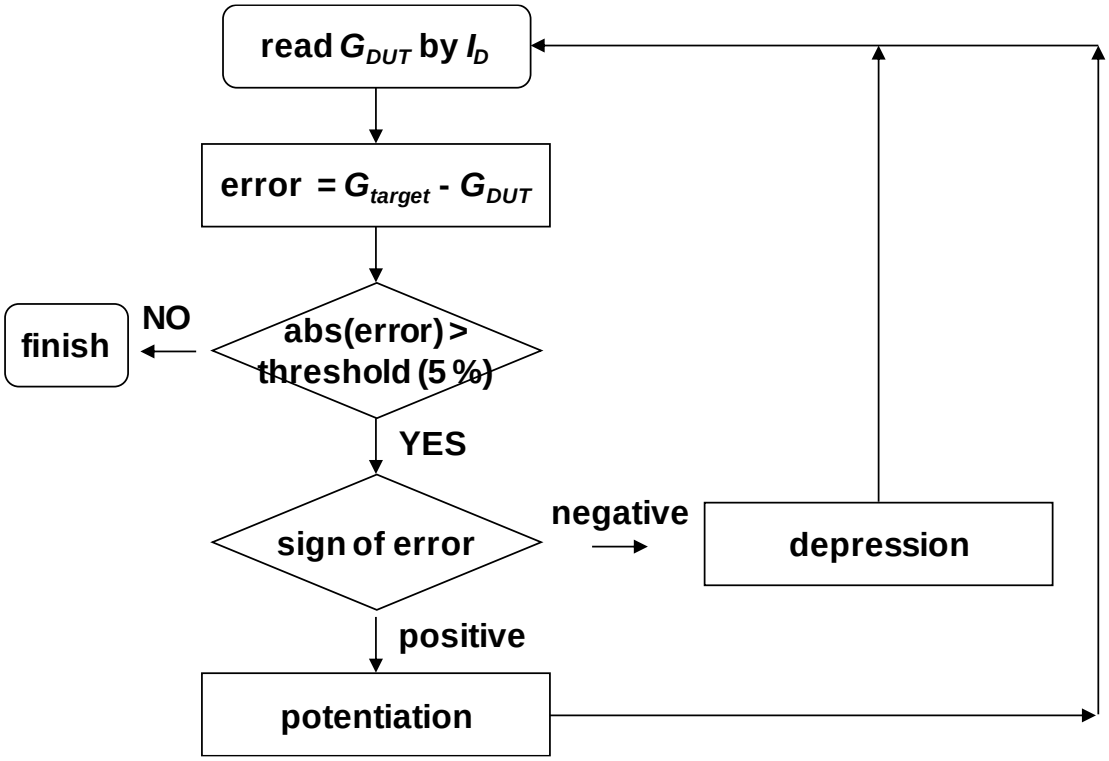

Figure S8. Flow chart of the update-verify scheme.

Fig. S8 shows a flow chart for the update-verify process. We used an update-verify technique to update the channel conductance of devices in the crossbar. Specifically, each cycle is based on a sequence of update–read pulse pairs, each pair including a programming (potentiation or depression) pulse and a subsequent READ pulse (V_G = -1 V, 100 μs) for verification purpose. Current from the READ operation on a target cell is used to compare with a target value and calculate an error. If the error is below a pre-defined threshold, the operation is considered complete and the process stopped, otherwise operations are taken based on the sign of the error. For positive errors, a potentiation pulse (V_G = +8V, 1 μs) is applied to increase the device conductance, while for negative errors a depression pulse (V_G = −9V, 1 μs) is applied to decrease

the device conductance. The procedure is then repeated until the conductance reaches within a pre-determined range of the target value (*e.g.*, $G_{var} = (G_{target}–G_{DUT})/G_{DUT} = 5$ %).

5-2. The procedure of image convolution

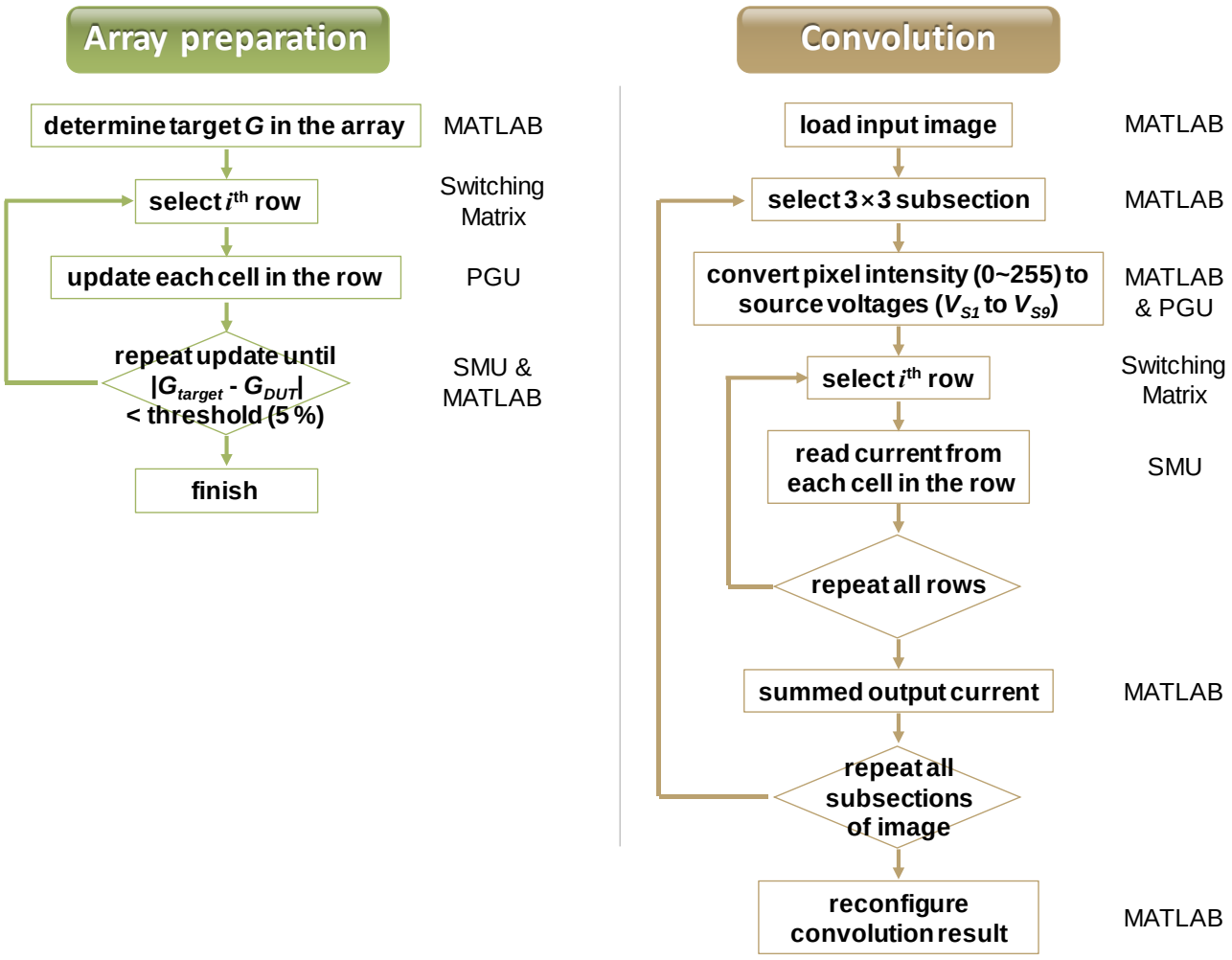

Figure S9. Flow chart of the image convolution

Fig. S9 shows the data processing procedure for image convolution. For the convolution operation, *G* of all cells in the array should be updated first to the desired value. *G* of all cells can be adjusted using the update-verify feedback method discussed in Supporting Information Note

5-1. Next, to obtain the convolution result, the input grayscale image whose pixel intensity is in the range of 0–255 is loaded by MATLAB program. From this image, a 3 × 3 subsection, where the convolution will be performed, is assigned, and the magnitude of the intensity is proportionally converted to the source voltage magnitude (0 V ≤ V_{S1} to V_{S9} ≤ 2 V). These source voltages are sequentially applied to the crossbar array row-by-row with V_D = 0 V set to all drains. Each output column current is measured by SMU, and it is integrated by MATLAB. From the summed column current, the total convolution result can be reconfigured, *i.e.*, the final convolution result shown in Fig. 4c.