Supporting information for

Precision-extension technique for accurate vector-matrix multiplication with CNT transistor crossbar array

Sungho Kim[1], Yongwoo Lee[2], Hee-Dong Kim[1] & 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).

KEYWORDS: carbon nanotube, crossbar array, dot product, matrix multiplication, memristor

1. Electrical properties of the CNT transistor

1-1 Crossbar array integration

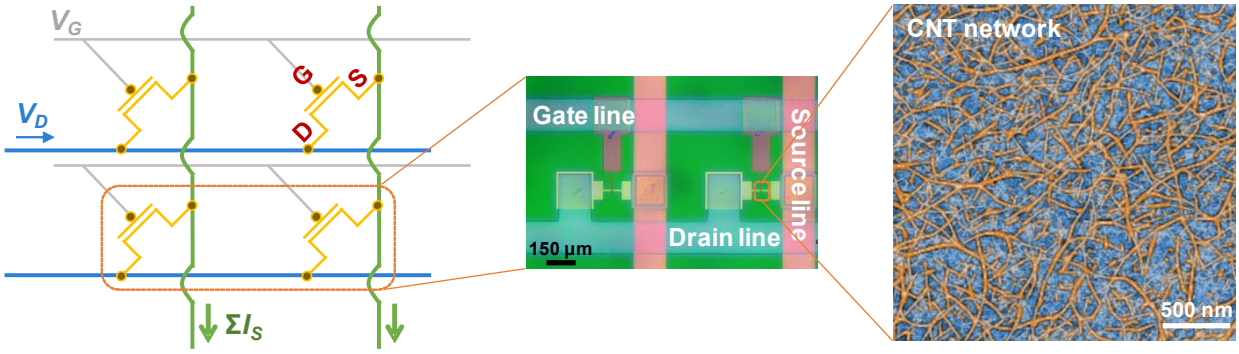


Figure S1. The schematic of the CNT transistor crossbar array, where the input voltage signal is applied to the drain electrode of each transistor ($V_i = V_D$), and the summed source current (ΣI_S) corresponds to the multiplication of V_D by the channel conductance G_i of each transistor. The atomic force microscope image shows the random matrix of single-walled semiconducting CNTs, which serve as the channels of each transistor.

In this study, 10 × 10 CNT transistor crossbar array is demonstrated. As shown in Fig. S1, the gate and drain lines are parallel to each other in a row direction, and the source line is perpendicular to these lines in a column direction. The gate voltage (V_G) is used to update/read the channel conductance of the transistor (G), and the drain voltage (V_D) delivers the input vector information to be calculated in the crossbar array. By exploiting the source current (I_S) which represents the integration of $V_D \cdot G$ (*i.e.*, $\Sigma I_S = V_D \cdot G$), the vector-matrix multiplication (VMM) computation can be performed in the crossbar array without any logical operation.

The channel of CNT transistors is constructed from highly purified 99%-semiconducting CNT solutions processed using a density gradient ultracentrifuge separation method.[S1] The removal of

metallic CNTs via a solution process for the semiconducting CNT network has been shown to dramatically improve the electrical performances of CNT transistors, including a decrease in the leakage path, achieving a high I_{on}/I_{off}. Moreover, a high uniformity and high device yield could be achieved simultaneously.

1-2. The physical mechanism of the hysteresis

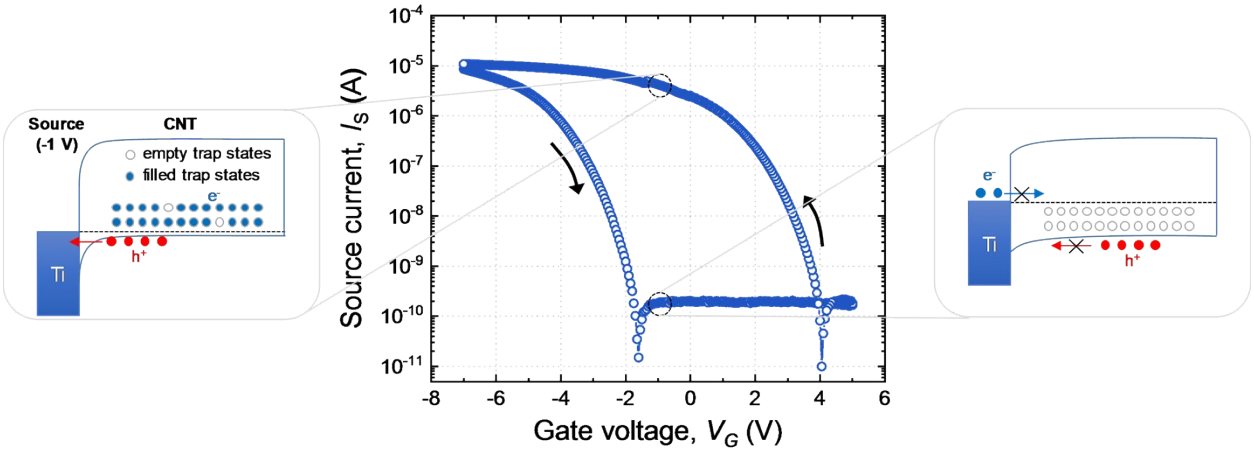


Figure S2. Hysteresis of the source current (I_S) as a function of the gate voltage (V_G) with a constant source voltage (V_S = -1 V).

The channel conductance of the CNT transistor (G) can be modulated gradually, and this behavior is presented by the hysteresis in the source current (I_S) under the gate voltage (V_G) sweep (Fig. S2), where positive/negative V_G increase/decrease G respectively. 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).[S2] Fig. S2 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, 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 source/CNT. Therefore, hole tunneling current can be increased, and consequently, 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 G switching behavior.

1-3. The retention and endurance properties of analog *G* switching.

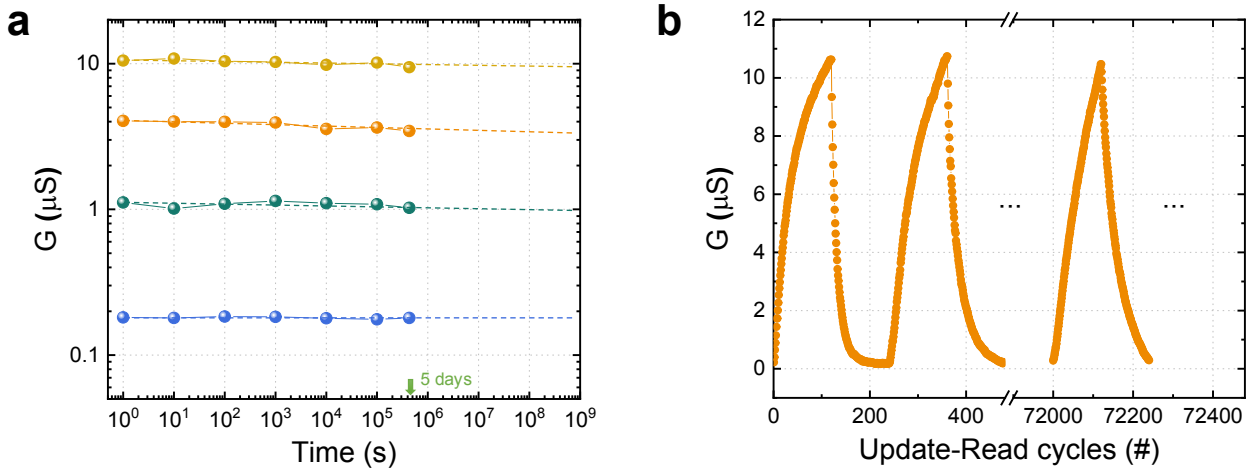


Figure S3. (a) The retention and (b) the endurance properties of the analog G switching.

In this section, we show the other properties of the analog *G* switching not covered in the main text. Fig. S3a shows the retention characteristics for four arbitrarily selected analog *G* states. The stable *G* states can be maintained even for 5 days. In addition, Fig. S3b shows *G* switching behavior up to 7.2×10^4 update-read cycles, where stable and reliable analog *G* switching can be maintained. highly selective and parallel updating is enabled by localized trapping in the CNT synaptic transistor. These results ensure the reliable DPE operation.

2. The update-verify feedback method

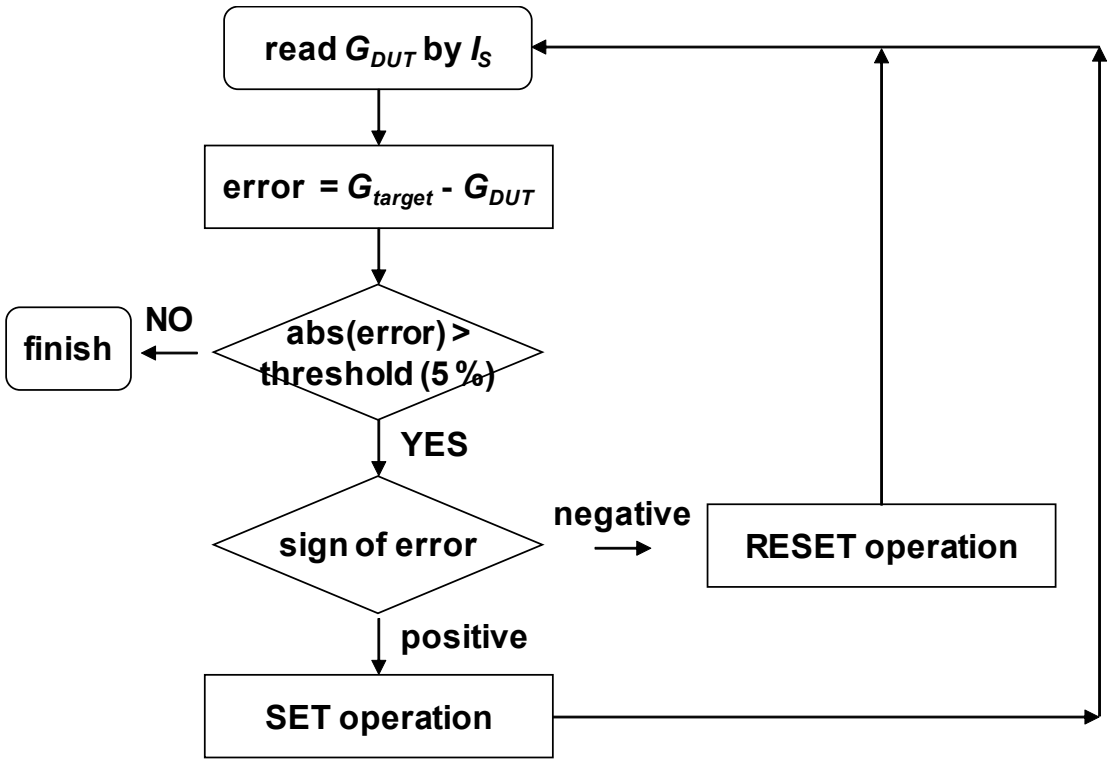


Figure S4. Flow chart of the update-verify scheme.

Fig. S4 shows a flow chart for the update-verify process. To reduce the device variability, we used an update-verify technique to update the channel conductance of each device in the crossbar. Specifically, each cycle is based on a sequence of update–read pulse pairs, each pair including a programming (SET or RESET) 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 SET pulse (V_G = 6V, 100 μs) is applied to increase the device conductance, while for negative errors a RESET pulse (V_G = -6V, 100 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 (for example N_{var} = 5 %). In the experimental implementation, the updating of device conductance typically requires around 20 update-verify cycles.

3. The additional explanation of the DPE operation

3-1. The ADC operation

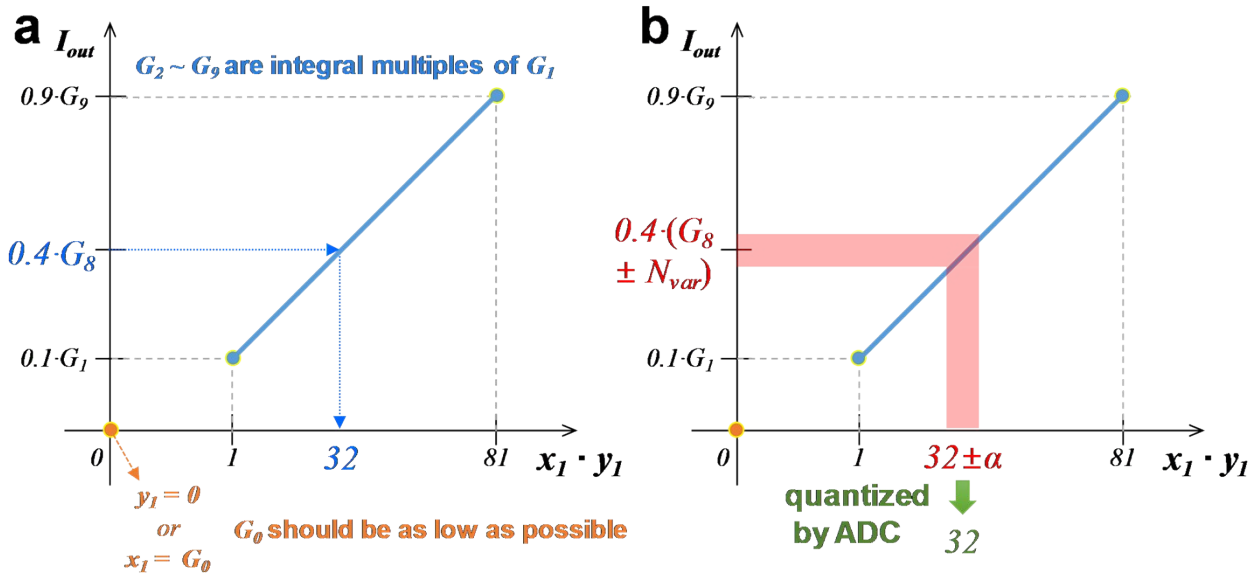


Figure S5. The linear relationship between measured output current (I_{out}) and the inferred multiplication result ($x_1 \cdot y_1$) in (a) an ideal case and (b) a real case.

As mentioned in the main text, the input voltage amplitude (V_i) applied to the row of a crossbar array is determined by an integral multiple of the amplitude when $y_1 = 1$. Similarly, $G_2 \sim G_9$ should be an integral multiple of G_1. Because V_i and G_i are assigned to be an integral multiple of the reference value, the linear relationship can be established between the output current (I_{out}) and the inferred multiplication result ($x_1 \cdot y_1$) as shown in Fig. S5a. Note that the minimum multiplication result (*i.e.*, $x_1 \cdot y_1 = 1 \times 1 = 1$) is mapping to the output current $I_{out} = 0.1 \cdot G_1$; the maximum multiplication result (*i.e.*, $x_1 \cdot y_1 = 9 \times 9 = 81$) is mapping to the output current $I_{out} = 0.9 \cdot G_9$. Based on this linear relationship, the scale ratio R is given as follows.

$$R = \frac{\Delta I_{out}}{\Delta (x_1 \cdot y_1)} = \frac{0.9G_9 - 0.1G_1}{81 - 1}$$

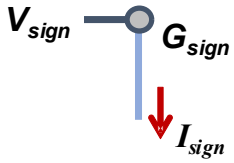
Consequently, the multiplication result ($x_1 \cdot y_1$) can be estimated by exploiting the scale ratio R to the measured I_{out} value.

Here, when either $y_1 = 0$ or $x_1 = 0$, I_{out} should be zero ideally. The situation where $y_1 = 0$ can be easily emulated by adjusting the input voltage amplitude as $V_i = 0$ V. Similarly, the situation where $x_1 = 0$ should be emulated by the device conductance G_0 to be zero, but the device conductance G_0 has a non-zero finite value. Because of the non-zero G_0 value, the multiplication result can be inferred incorrectly. Therefore, the value of G_0 should be set as low as possible to prevent this misunderstanding; $0.9 \cdot G_0$ (*i.e.*, $x_1 \cdot y_1 = 9 \times 0$) should be lower than $0.1 \cdot G_1$ (*i.e.*, $x_1 \cdot y_1 = 1 \times 1$) to clarify that one of x_1 or y_1 is zero.

After re-scaling of I_{out} through R, the multiplication result can be inferred from measured I_{out}. Unfortunately, due to the several sources of error, such as series resistances in wire, sneaky current path, and particularly N_{var}, the inferred multiplication result is inaccurate yet as shown in Fig. S5b; in the example when $x_1 = 8$ and $y_1 = 4$, inferred result will be $32 \pm \alpha$ where α refers to the error. This error should be eliminated by the quantization process of ADC circuitry. The operation principle of the ADC is very simple. In our n-based system (*e.g.*, $n = 0$~9), the ideal output of the multiplication ($x_1 \times y_1$) is the same as the result of the multiplication table, *i.e.*, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, …. , 30 (=5×6 or 6×5), 32(=4×8 or 8×4), 36(=4×9 or 9×4), … 81 (9×9). When $x_1 = 8$ and $y_1 = 4$, the output of ($x_1 \times y_1$) will be $32 \pm \alpha$ where α refers to the error. Here, if the ADC is designed to output only discrete values such as 30, 32, and 36, then the ADC will output 32 closest to the input, which allows the errors to be eliminated. Therefore, if the actual inferred result (*i.e.*, $32 \pm \alpha$) is mapping to the nearest input by the ADC, then the final multiplication result, 32, can be obtained. In the experimental implementation, instead of actual

ADC circuit implementation, the ADC operation is emulated by home-made software (see Supporting Information Note 5).

3-2. The sign determination process



Y \ X	$G_{sign} > G_9$ $(X > 0)$	$G_{sign} < G_0$ $(X < 0)$
$V_{sign} = 1\ \mathrm{V}$ $(Y > 0)$	$I_{sign} > G_9$	$I_{sign} < G_0$
$V_{sign} = -1\ \mathrm{V}$ $(Y < 0)$	$I_{sign} < -G_9$	$I_{sign} < -G_0$

Figure S6. The sign determination process in the precision-extension technique.

Fig. S6 shows the sign determination process in our precision-extension technique. The voltage amplitude can be a negative, but the conductivity of the device cannot be a negative, so an additional device (G_{sign}) is required to represent the sign of the input vector. For example, $G_{sign} > G_9$ indicates a positive matrix element (*i.e.*, $X > 0$), and $G_{sign} < G_0$ represents a negative matrix element ($X < 0$). Similarly, depending on the sign of input vector Y, V_{sign} has a value of -1 V or 1 V. As a result, we can determine the sign of Z (= $Y \times X$) from the magnitude and direction of I_{sign} (=$V_{sign} \cdot G_{sign}$).

4. The scheme for the selective access in the crossbar array

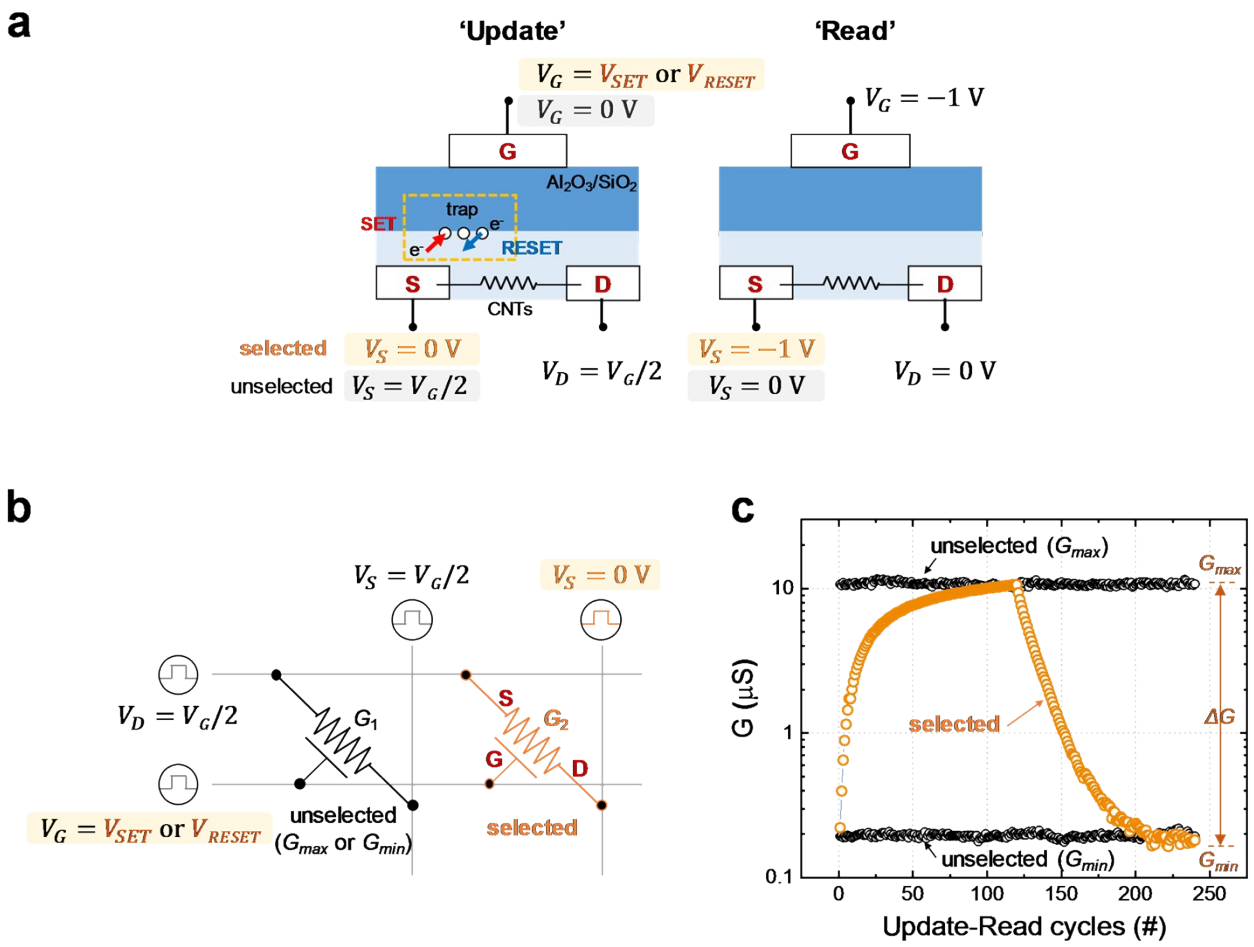


Figure S7. (a) The localized electron trapping/detrapping processes in the CNT synaptic transistor enables selective updating, which is possible because of the three individually adjustable electrodes (*i.e.*, gate, drain and source). (b) Schematic of a 1-by-2 transistor array. Selective updating is carried out by controlling V_G and V_S along the rows and columns, respectively. (c) Selective updating by subjecting G_1 and G_2.

The selective update/read scheme is essential to implement a highly integrated resistive crossbar array. Common two-terminal synaptic devices (such as memristors) cannot effectively

prevent the unwanted current flow through unselected devices without additional selector devices; thus, the resulting voltage drops across the parasitic current path critically reduce the update-read accuracy. For these reasons, existing experimental demonstrations using two-terminal resistive device arrays have been limited to relatively small sizes (<32 × 32).

In contrast, the selective update/read of the CNT transistor can be achieved by individually controlling its three terminals to induce and control localized carrier trapping (Fig. S7a). When a positive voltage is applied to the gate ($V_G = V_{SET}$), the potential difference between the gate and the source ($V_{GS} = V_{SET}$) leads to electron trapping near the drain side. Whereas the drain is always kept at half of the gate voltage ($V_D = V_G/2$); the potential difference between the gate and the drain ($V_{GD} = V_{SET}/2$) is not enough to initiate the electron trapping process. Locally trapped electrons at the drain side can narrow the Schottky barrier width at the source junction (Fig. S2), thereby increasing hole tunneling current and subsequent channel conductance (G). By contrast, the negative gate voltage ($V_G = V_{RESET}$) results in the detrapping of electrons, and leads to a decrease in G. Note that unselected devices can avoid disturbance from the update process of the adjacent selected device by simply applying half of V_G to both the drain and the source; thereby, the G of the unselected devices can be maintained due to the forbidden trapping/detrapping processes. Interestingly, the three-terminal structure of the synaptic transistor can easily eliminate the sneak path current during the read operation. The unwanted current flow through the unselected devices can be prevented by applying the same voltage to both the drain and the source ($V_D = V_S = 0$ V), and this does not require an additional selector device or intrinsic rectifying behavior of the synaptic device.

To experimentally demonstrate the selective access of the crossbar array, we update two CNT transistors with channel conductances, G_1 and G_2, electrically connected as illustrated

schematically in Fig. S7b. The selective access is encoded by voltages applied along the rows (V_G and V_D), and columns (V_S). Initially, $G_{1\ is}$ set to either G_{max} or G_{min}. Selective update can be executed by controlling the source voltage: $V_S = 0$ V for the selected source line, and $V_S = V_G/2$ for the unselected source line to prevent the update. Fig. S7c shows the update result in the crossbar array without disturbing the adjacent devices, where highly selective updating is enabled by localized trapping in the CNT synaptic transistor with individually controllable three terminal structure.

5. The experimental setup for the DPE operation

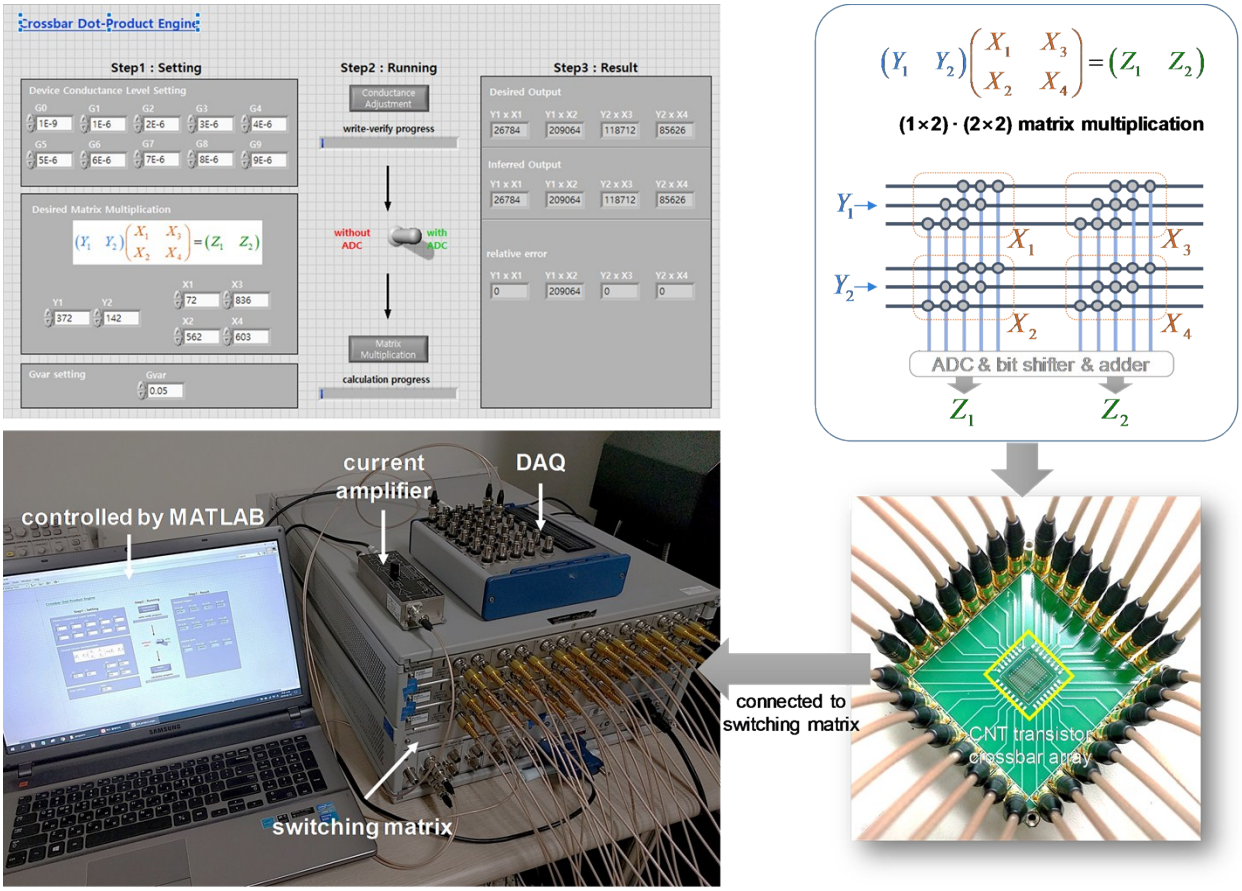


Figure S8. The experimental setup for VMM computation in the CNT transistor crossbar array.

Fig. S8 shows the experimental setup for VMM computation in the CNT transistor crossbar array. Within 10×10 size of the crossbar array, only 6×5 size of array was partially used for VMM computation involving the multiplication of 3-digit number. The crossbar array is connected to the designed PCB test board through a wire-bonding process, and measurement systems are then connected to the test board through the switching matrix (Agilent 5250A). The application of a pulse signal was conducted by DAQ equipment (NI USB-6363), and the amount of output current was measured by DAQ equipment, with transimpedance amplifier (FEMTO, DHPCA-100). Home-made GUI interface allows the user to tune different parameters to control

entire VMM computation process including the matrix element, precision, ADC operation, and conductance levels. The relative error of each multiplication can be automatically obtained by this interface.

The quantization process discussed in the main text is implemented through a MATLAB program rather than an actual circuit implementation. This quantization process is expected to be easily implemented using existing ADC circuitry.

REFERENCES

(S1) Arnold, M.S. *et al.* Sorting Carbon Nanotubes by Electronic Structure using Density Differentiation. *Nat. Nanotech.* **2006**, *1*, 60-65.

(S2) Duong, D. L. *et al.* Origin of Unipolarity in Carbon Nanotube Field Effect Transistors. *J. Mater. Chem.* **2012**, *22* (5), 1994–1997.