AFRL-AFOSR-VA-TR-2024-0231

Intelligent Neuromorphic Network Based on Carbon Nanotube/Polymer Composite

Yong Chen
UNIVERSITY OF CALIFORNIA LOS ANGELES
11000 KINROSS AVE STE 102
LOS ANGELES, CA,
US

05/27/2024
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20240527	Final	START DATE 20150115	END DATE 20200114

4. TITLE AND SUBTITLE
Intelligent Neuromorphic Network Based on Carbon Nanotube/Polymer Composite

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-15-1-0056	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Yong Chen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF CALIFORNIA LOS ANGELES
11000 KINROSS AVE STE 102
LOS ANGELES, CA
US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Office of Scientific Research 875 N. Randolph St. Room 3112 Arlington, VA 22203	AFRL/AFOSR RTB2	AFRL-AFOSR-VA-TR-2024-0231

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The goal of the project is to develop an intelligent neuromorphic network based on carbon nanotube (CNT) composites to emulate biological neuron networks and create intelligent systems in complex erratic environments. In the last year, we have made progresses in the following categories: (1) Design of the device and circuit architecture of the neuromorphic
network; (2) the development and test of novel composites, devices, and fabrication processes of the neuromorphic network; (3) modeling and analysis of the self-programming and optimization mechanism of the neuromorphic network; (4) the demonstration of the intelligent behaviors of the neuromorphic network.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	c. THIS PAGE		
U	U	U	UU	27

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
BYUNG LEE	426-8483

Standard Form 298 (Rev.5/2020)
Prescribed by ANSI Std. Z39.18

Progress summary

The goal of the project is to develop an intelligent neuromorphic network based on carbon nanotube (CNT) composites to emulate biological neuron networks and create intelligent systems in complex erratic environments. In the last year, we have made progresses in the following categories: (1) Design of the device and circuit architecture of the neuromorphic network; (2) the development and test of novel composites, devices, and fabrication processes of the neuromorphic network; (3) modeling and analysis of the self-programming and optimization mechanism of the neuromorphic network; (4) the demonstration of the intelligent behaviors of the neuromorphic network.

Technical

1. Design of the device and circuit architecture of the neuromorphic network

The neuromorphic network have been designed by integrating 120 of synapstors and 12 Si soma circuits. In a biological neuron network, synapses, the junctions between neurons, are essential for signal processing, memory, and learning functions. We have designed a device, synapstor, based on CNT composites to emulate the functions of the synapses. The synapstor conductance is modified based on the correlation between the output of the circuit and the system performance. We have also designed Si soma circuits based on Schmitt switches to process the output currents from the synapstors based on the integrate-and-fire mechanism. In the neuromorphic network, spikes represent the sensing signals from an external system are input to the synapstors, the spikes trigger postsynaptic currents via synapstors, and the accumulative currents trigger output spikes from the Si soma circuit based on the integrate-and-fire mechanism. The synaptic weights in the synapstors are also modified based on the correlation between the input and output spikes. The circuit has the functions of high-speed parallel signal processing, memory, and self-programming functions with low power consumption.

2. Development and test of novel composites, devices, and fabrication processes for the neuromorphic network

We have explored different CNT-based composites, including metals, liquid ionic polymers, oxides, C60 molecules, Au nanoparticles to optimize the synapstor and neuromorphic network properties. Various CNT densities, polymer thicknesses, and processing conditions have been explored to optimize the device and circuit functions. When a signal is input to a synapstor, an electronic charge is injected in the composite matrix surrounding the CNTs, which in turn modifies the electronic concentration and generates dynamic post-synaptic current in the CNT in the synapstor. The input signal can also induce the electronic charge stored in the matrix permanently, which generates long-term memory and learning capabilities in the synapstor. The new materials have significantly improved the device reliability, stability, memory, and learning functions of the synapstors. We have also developed new device structures and fabrication processes for the neuromorphic network. The new device structures have dramatically improve the device learning functions, and the new processes to prepare the CNT composites and fabricate devices and circuits have significantly improved the yield, throughput, and reliability of the circuits. We have also developed the process to fabricate the devices on the polymer soft substrates, which will lead to the large-scale 3D neuromorphic network and the integration of sensing network and neuromorphic circuits in the future.

3. Modeling and analysis of the neuromorphic network

A key issue for developing the neuromorphic network is to understand its self-programming and optimization mechanism. We have simulated the neuromorphic network and investigated its interaction with simulated systems, such as an abstract nonlinear systems, speech samples, etc. We have compared the behaviors of the neuromorphic network in nonlinear, stochastic, and critical environments to identify key features and limitations of the neuromorphic network. We have developed a stochastic model to simulate neuromorphic network and its behavior in response to noisy, erratic environments, which allows us to quantify and optimize self-programming process for the real applications. We are using a stochastic process approach to investigate the circuit self-programming rule based on correlation between the inputs and outputs of the circuits.

4. Demonstration of the intelligent behaviors of the neuromorphic network

We have used the neuromorphic network to drive an unmanned aerial vehicle (UAV). During the dynamic interaction, the signals from a sensing network on the UAV are sent to the network in a parallel mode, and trigger output signals to control the UAV. The signal-processing algorithm of the network is also self-programmed via a learning process in parallel, and the intelligent behavior of the network evolves spontaneously to improve the UAV performance functions such as stability and obstacle recognition in erratically changing environments. Unlike the signal processing and control algorithms of the Si circuits, which need to be programmed and defined precisely by human, but become invalid when the environment erratically changes beyond the designed ranges and conditions, the algorithms in the neuromorphic network are established via a self-organized learning process, which can improve the UAV performance functions even when the UAV encounters erratic changes and exogenous variations in a complex environment.

We have also developed a perceptive neuromorphic network by integrating a sensing network with the neuromorphic network based on carbon nanotube composites. The sensing network was fabricated using carbon nanotube composites on a mechanically flexible, optically transparent substrates. The neuromorphic network enables perceptive sensing, low-power consumption, high-speed parallel signal processing without external computing and human inference. We have also used the neuromorphic network to learn and recognize speech signals.

When Turing established his universal computing model, his overarching ambition was to emulate the human brain.[1] Following Moore's law, the miniaturization of the transistors has exponentially improved the performance and energy efficiencies of computers[2] (**Figure 1**a), leading to an information revolution and artificial intelligent systems that can simulate learning functions of the human brain.[3] Based on the Turing model, digital computers execute algorithms in serial mode by physically separated logic and memory transistors (Figure 1b), and the computing energy is predominantly consumed by data memory and signal transmissions between memory and logic units,[4-7] referred to as the "von Neumann bottleneck."[8] Transistor-based circuits with parallel computing architectures and distributed memories, such as graphics processing units (GPUs) from Nvidia,[9] tensor processing units (TPUs) from Google,[3, 10] field-programmable gate arrays (FPGAs) from Intel,[11] and the TrueNorth neuromorphic circuit from IBM[12] have been developed to improve their energy efficiencies (Figure 1a) to the range of

$10^{10} - 10^{11}\ FLOPS/W$ (floating point operations per second per watt) by increasing parallelism and reducing global data transmission. However, their energy efficiencies are fundamentally limited by the energy consumptions on memory ($\sim 10^{-15}\ J/bit$) and signal transitions ($\sim 10^{-11}\ J/bit$) in digital computing circuits.[5, 6] When transistors approach the limitations of their minimal sizes near the end of Moore's law, the energy efficiencies of transistor-based computing circuits are asymptotically saturated[4-6, 13, 14] (Figure 1a). Meanwhile, the information industry generates "big data" with exponentially increasing volumes, and leads to exponentially increasing power requirements for computations.[4, 6, 14] This trajectory is unsustainable as it would exceed the entire global power production in one or two decades[15] (Figure 1a). It is imperative to develop a new platform to facilitate inference and learning from "big data" in emerging intelligent systems with significantly higher energy efficiency than that of the transistor-based Turing computing platform.

The human brain performs inference and learning from "big data" with an estimated speed ($\sim 10^{16}\ FLOPS$)[16] comparable to the speed ($\sim 10^{17}\ FLOPS$) of the fastest supercomputer, Summit,[17] but consumes much less power ($\sim 20\ W$) than the supercomputer ($\sim 10^{7}\ W$), and is much more energy-efficient ($\sim 10^{15}\ FLOPS/W$) than the supercomputer ($\sim 10^{10}\ FLOPS/W$, Figure 1a). By contrast, the human brain concurrently performs spatiotemporal inference and learning in analog parallel mode[16, 18, 19] (Figure 1c) via a network of neurons connected by $\sim 10^{14}$ synapses (Figure 1d). For inference, a wave of voltage pulses, $V_i^m(t)$, in the mth presynaptic neuron is processed by a synapse connected with the mth presynaptic and nth postsynaptic neurons, and induces a current in the nth postsynaptic neuron[20], $I^{nm} = \kappa^{nm} \circledast (w^{nm} V_i^m)$, where w^{nm} denotes the synaptic weight (conductance), $\kappa^{nm}(t)$ denotes a temporal kernel function, and $\kappa^{nm} \circledast (w^{nm} V_i^m)$ represents the temporal convolution between κ^{nm} and $w^{nm} V_i^m$. For spatiotemporal parallel inference, a wave of voltage pulses in presynaptic neurons induces a collective current via synapses in the nth postsynaptic neuron (Figure 1d), which can be expressed as,[20]

$$I^n(t) = \sum_m \kappa^{nm} \circledast (w^{nm} V_i^m) \qquad (1)$$

and the current induces voltage pulses, $V_o^n(t)$, in the nth postsynaptic neuron. When the voltage pulse is fired in the postsynaptic neuron ($V_o^n \neq 0$), the postsynaptic current $I^n = 0$. The w^{nm}

matrix is also modified concurrently by the spatiotemporal waves of voltage pulses in the presynaptic and postsynaptic neurons for learning,[19-21]

$$\dot{w}^{nm} = \alpha\, V_i^m V_o^n \qquad (2)$$

where α denotes the conductance modification coefficient, and V_o^n and V_i^m voltage pulses have the same amplitudes and durations. w^{nm} is modified when $V_i^m = V_o^n$, with the learning coefficient $\alpha > 0$ in Hebbian learning, and $\alpha < 0$ in anti-Hebbian learning. α is a function of the timing difference between V_i and V_o pulses in the learning based on synaptic spike-timing-dependent plasticity (STDP). Based on Equation 2, general correlative learning algorithms in machine learning[19] can also be implemented (Supporting Information). Following Equation 2, when $V_i^m \cdot V_o^n = 0$ (e.g. $V_i^m \neq 0$ and $V_o^n = 0$ during inference), $\dot{w}^{nm} = 0$, i.e. w remains nonvolatile for memory. By integrating the analog convolutional processing (Equation 1), correlative learning (Equation 2), and nonvolatile memory functions in a single synapse, the brain circumvents the fundamental limitations such as physically separated memory units, data transmission between memory and logic units in computers, and concurrently executes the inference (Equation 1) and learning (Equation 2) algorithms in a neural network in analog parallel mode with an energy efficiency more than five orders of magnitudes higher than that of the Summit supercomputer.

Analog memory devices such as memory transistors,[22, 23] memory resistors (memristors)[24, 25, 26, 27], and phase change memory (PCM)[28, 29] have been developed to emulate synapses. For inference, the neuromorphic circuits based on these devices processed input voltage pulses, V_i^m, and generated an output current by following Ohm's law, $I^n = \sum_m w^{nm} V_i^m$, in parallel analog mode with energy efficiencies of $\sim 10^{10} - 10^{14}\ FLOPS/W$,[25, 26, 27-29] which significantly exceeded the energy efficiencies of digital circuits (Figure 1a). However, these devices lacked the function to trigger currents that lasted over time after the V_i^m pulses ended, as described by Equation 1, which prevented the devices from convolutional signal processing of dynamic signals for spatiotemporal inference.[20] For learning, the conductance of a memory transistor or memristor were modified by simultaneously applying writing voltage signals on their input and output electrodes, and correlative learning algorithms such as STDP were executed on individual devices by applying tailored voltage signals,[23, 27]

but a writing voltage signal applied on its input or output electrode alone would also change its conductance, therefore the devices did not follow Eq. 2 (i.e. $\dot{w}^{nm} = \alpha\, V_i^m V_o^n \neq 0$ when $V_i^m \neq 0$ and $V_o^n = 0$ or $V_o^n \neq 0$ and $V_i^m = 0$). To avoid change of conductance during signal inference, the voltage signals for inference were decreased to significantly smaller magnitudes than the voltage signals for learning, thus when the inference algorithm was executed in the circuits, the learning algorithm was interrupted, and vice versa.[22, 23, 24, 25, 26, 27-29] Moreover, in order to modify device conductances accurately in a circuit, learning algorithms were executed in external digital circuits to obtain targeted conductance values, then devices were modified to the targeted conductance values by applying different writing voltages on different devices sequentially in iterative writing and reading processes.[25, 26, 27-29] The energy efficiencies for the writing processes were $\sim 10^5 - 10^{13}\ FLOPS/W$[25, 26, 27-29] (Figure 1a), but the energy and time for the external digital computing circuits to execute correlative learning algorithms from "big data" with M-dimensional variables increase versus M exponentially,[3, 7] referred to as the "curse of dimensionality".[30] Although circuits based on existing analog memory devices execute inference algorithms with high speeds and energy efficiencies, the differences between voltage signals for inference and learning prevent the circuits from executing inference (Equation 1) and learning (Equation 2) algorithms concurrently. They also required separated memory and logic circuits, and signal transmissions between the circuits to execute learning algorithms, which limits their speeds and energy efficiencies for learning ($\lesssim 10^{11}\ FLOPS/W$).[14]

Here we report a synaptic resistor, abbreviated as synstor hereinafter, to emulate a synapse by integrating analog convolutional processing (Equation 1), correlative learning (Equation 2), and nonvolatile memory functions in a single device (Figure 1c). By transmitting waves of voltage pulses with the same amplitude and width to the input and output electrodes of an $M \times N$ crossbar synstor circuit, the spatiotemporal convolutional inference (Equation 1) and correlative learning algorithms (Equation 2) can be executed concurrently in the circuit in parallel analog mode (Figure 1d, Supporting Information, Figure S1). We have demonstrated a 4×2 crossbar synstor circuit executing speech inference and learning concurrently with an energy efficiency of $\sim 1.6 \times 10^{17}\ FLOPS/W$ (Figure 1a), which is about seven orders of magnitudes higher than that of the Summit supercomputer.

The structure of a carbon nanotube (CNT) synstor is shown in **Figure 2**a and 2b, and the device fabrication process is described in the Supporting Information and Figure S2. The

synstor has an input electrode, an output electrode, as well as a reference electrode as a common electric ground like a synapse. The synstor is composed of a 20 μm-wide p-type semiconducting CNT network which forms Schottky contacts with the Al input and output electrodes. The CNT networks are fabricated on a 6.5 nm-thick HfO_2 dielectric layer on a 2.5 nm-thick and 10 μm-wide TiO_2 charge storage layer on a 22 nm-thick HfO_2 dielectric layer on a 50 nm-thick and 10 μm-wide Al reference electrode (Figure S3). There is a 5 μm lateral space between the TiO_2 charge storage layer/Al reference electrode and Al input/output electrodes. The synstor has a transistor-like structure, but its reference electrode is always grounded, and it does not need a programmed gate voltage to control the current between the input and output electrodes like a transistor.

A synstor was tested with a continuous voltage sweep, V_i, on its input electrode and a grounded output electrode. The current flowing through a synstor, I, was measured as a function of V_i, and displayed in Figure 2c. The nonlinear rectifying $I - V_i$ curves indicate that Schottky barriers were formed between the Al input/output electrodes and the p-type semiconducting CNTs, as reported previously.[31] The DC conductance w of the synstor is a nonlinear function of V_i. The device was modified by applying 50 pairs of 5 ms-wide V_i and V_o voltage pulses on their input and output electrodes simultaneously with the same amplitude ($V_i = V_o$). As shown in Figure 2c, after the device experienced 50 pairs of V_i and V_o pulses with $V_i = V_o = 1.75\ V$, w was decreased; after the device experienced 50 pairs of V_i and V_o pulses with $V_i = V_o = -1.75\ V$, w was increased. The changes of w, $\Delta w = w - w_0$, were also measured before and after the synstors experienced 50 pairs of 5 ms-wide V_i and V_o pulses with various amplitudes ranged between $-2\ V \leq V_i = V_o \leq 2\ V$, and the percentage changes of w, $\Delta w/w_0$, are plotted versus the pulse amplitudes in Figure 2d. The w is modified by following Equation 2, $\dot{w} = \alpha\ V_i \cdot V_o$. When $V_i = V_o \gtrsim 1.0\ V$, w was decreased ($\alpha < 0$); when $V_i = V_o \lesssim -0.8\ V$, w was increased ($\alpha > 0$); when $-0.8\ V \lesssim V_i = V_o \lesssim 1.0\ V$, $\dot{w} \approx 0$ ($\alpha \approx 0$). Synstors were modified to its high and low conductance values, w_H and w_L, by applying 50 pairs of 5 ms-wide V_i and V_o pulses with $V_i = V_o = -1.75\ V$ and $V_i = V_o = 1.75\ V$ alternatively in 2930 modification cycles, and no deterioration of device conductance modification was observed (Figure S4). 108 synstors on a chip were modified to w_H and w_L respectively, and the distributions of w_H and w_L values are shown in Figure S5. The average w_H value, $\overline{w}_H = 1.9$ nS, and the standard deviation of w_H, $\sigma_H = 0.44$ nS. $w_L < 0.2$ nS, which is the limit of the measurement module.

After synstors were modified to their analog conductances, the synstors were tested under $V_i \cdot V_o \leq 0$ by applying 50 pairs of various 5 ms-wide V_i and V_o pulses under the conditions: (1) $-2\ \mathrm{V} < V_i < 2\ V$ and $V_o = 0$; (2) $-2\ \mathrm{V} < V_o < 2\ V$ and $V_i = 0$; and (3) $-2\ \mathrm{V} < V_i = -V_o < 2\ V$. It was observed that w remained unchanged under these conditions (Figure 2d), which indicates that the synstor has a nonvolatile memory of w (i.e. $\dot{w} \approx 0$ (Equation 2)) under $V_i \cdot V_o \leq 0$. After synstors were modified to different analog conductances, w_0, the nonvolatile memory of the synstors was examined by measuring their conductances versus time over $1.75 \times 10^5\ s$ at room temperature. Within the test period, the average percentage changes of the conductances $\overline{|\Delta w/w_0|} \approx 3\ \%$, and the extrapolations of the experimental data indicate their long-term (~10 years) nonvolatile analog memory (Figure S6).

A series of 10 ns-wide V_i and V_o pulses with the same amplitude ($V_i = V_o$) were applied on a synstor simultaneously (without the timing difference between V_i and V_o pulses), and its DC conductance, w, was measured before and after the pulses were applied. As shown in Figure 2e, when a series of 10 ns-wide paired V_i and V_o pulses with $V_i = V_o = 1.75\ V$ (or $V_i = V_o = -1.75\ V$) were applied on the synstor, w was gradually decreased (or increased) versus the number of pulse pairs, n. When a series of 10 ns-wide V_i and V_o pulses with $V_i = 0$ or $V_o = 0$ were applied on the synstor, the average percentage changes of the conductances, $\overline{|\Delta w/w_0|} \lesssim 3.3\ \%$. A series of periodic V_i pulses with an amplitude of $-1.75\ V$, a period of 30 ns, and different durations ($8, 10, 15, 20\ ns$) were applied periodically on a synstor under $V_o = 0$, and the currents flowing through the synstor, $|I(t)|$, increased versus t during the pulse, decayed versus time after the pulse (Figure 2f and Figure S7). $|I|$ also increased with increasing pulse number and duration.

A cross-section of the synstor with an Al/CNT/Al structure of a resistor and a CNT/HfO_2/TiO_2/HfO_2/Al structure of a capacitor is shown in **Figure 3**a. When a voltage pulse is applied on the input electrode of a synstor, it drives a current through the CNT network toward the grounded output electrode of the synstor, and simultaneously charges the capacitor between the CNT network and Al reference electrode. After the pulse ends, the capacitor is discharged, leading to a current through the output electrode. As shown in Figure 2f and Figure S7, the output current triggered by a series of $-1.75\ V$ periodic pulses with a period of 30 ns and a duration $t_d = 8, 10, 15, 20\ ns$ from a synstor changed versus time by following Equation 1, $I(t) = (w\ \kappa) \circledast V_i$, with $V_i(t) = \sum_n V_a \delta(t - t_n)$, V_a as the amplitude of the pulse, and t_n as

the moment to trigger the n^{th} pulse. $\kappa(t) = 1 - e^{-\beta_p t}$ during the pulse ($t \leq t_d$), and $\kappa(t) = (1 - e^{-\beta_p t_d})e^{-\beta_d t}$ after the pulse ($t > t_d$), with t_d as the duration of the pulse, β_p and β_d as the parameters related to the resistance and capacitance of the CNT network (Supporting Information, Equation S1). $I(t)$ was fitted by $I(t) = w\,\kappa \circledast V_i$ with $\beta_d = 41.5\,MHz$, and $\beta_p = 0.73,\ 0.67, 0.47, 0.36\ MHz$ when $t_d = 8, 10, 15, 20\,ns$, respectively.

As shown in the simulated electronic band structures in Figure 3b, a pair of negative (or positive) V_i and V_o voltages with $V_i = V_o = -1.75\,V$ (or $+1.75\,V$) increases (or decreases) the Fermi energy of the CNT network, and induces a difference of $+1.66\,eV$ (or $-1.59\,eV$) between the CNT and TiO_2 Fermi energies (**Figure 4**a), which injects electrons into (or depletes electrons from) the TiO_2 charge-storage layer by electronic hopping through the HfO_2 dielectric barrier layer[32] (Figure S8). The changes of the charge density in the TiO_2 layer, $\Delta\rho_s$, induced by the paired V_i and V_o voltages with $V_i = V_o$ were measured by capacitance-voltage tests on a synstor (Supporting Information and Figure S9), and plotted as a function of V_i and V_o in Figure 4b. When the synstor experienced V_i and V_o voltages with $V_i = V_o \gtrsim V_t^+$ and $V_i = V_o \lesssim V_t^-$, $\Delta\rho_s$ increased versus V_i, V_o. The $\Delta\rho_s - V_i, V_o$ data were fitted by $\Delta\rho_s = k_\rho^+[V_a - V_t^+]$ under $V_i = V_o \gtrsim V_t^+$ or $\Delta\rho_s = k_\rho^-[V_a - V_t^-]$ under $V_i = V_o \lesssim V_t^-$ with $k_\rho^+ = -145\,nF/cm^2$, $k_\rho^- = -106\,nF/cm^2$, $V_t^+ = 0.92\,V$, and $V_t^- = -0.85\,V$ as the positive and negative threshold voltages to modify the charges in their storage layer (Supporting Information, Equation S2). When the synstor experienced V_i and V_o voltages with $V_t^+ \gtrsim V_i = V_o \gtrsim V_t^-$, the external voltage could not drive a significant amount of electrons to overcome the potential barrier in the HfO_2 layer, therefore, no significant charge modification in the charge storage layer was observed ($\Delta\rho_s \approx 0$) (Figure 4b). When the synstor experienced V_i and V_o voltages with $V_i = V_o \gtrsim V_t^+$ and $V_i = V_o \lesssim V_t^-$, the negative (or positive) charges in the storage layer attract (repel) the holes in the p-type semiconducting CNT network (Figure 3c), and increase (or decrease) the device conductance w exponentially versus the magnitudes of V_i and V_o voltages (Supporting Information, Equation S3). The experimental data, $\Delta w/w_0$, (Figure 2d) were fitted by $\Delta w/w_0 = e^{\beta_v^+(V_i - V_t^+)} - 1$ with $\beta_v^+ = 4.06/V$ and $V_t^+ = 1.05\,V$ under $V_i = V_o > V_t^+$; and $\Delta w/w_0 = e^{-\beta_v^-(V_i - V_t^-)} - 1$ with $\beta_v^- = 3.69/V$ and $V_t^- = -0.81\,V$ under $V_i = V_o < V_t^-$; $\Delta w \approx 0$ under $V_t^+ > V_i = V_o > V_t^-$. When a series of paired V_i and V_o pulses with $V_i = V_o$ were applied, ρ_s was also modified by the external potential as a function of the number of the applied voltage pulses, n. ρ_s also gradually builds up an internal potential against the external potential, resulting in $\Delta w(n)/w_0$ to change as a logarithm function of n (Supporting Information,

Equation S4). The experimental data, $\Delta w(n)/w_0$, were fitted by $\Delta w(n)/w_0 = k_n^+ Ln\left(\frac{n}{n_0^+}+1\right)$ with $k_n^+ = 7.5$ and $n_0^+ = 1.7\times10^3$ for $V_i = V_o = 1.75\ V$; and by $\Delta w(n)/w_0 = k_n^- Ln\left(\frac{n}{n_0^-}+1\right)$ with $k_n^- = 15.3$ and $n_0^- = 1.76\times10^5$ for $V_i = V_o = -1.75\ V$ (Figure 2e). A single-wall CNT with an average diameter of ~$1\ nm$ locally forms a capacitor with the TiO_2 layer with an extremely low capacitance (~$10^{-19}\ F/nm$)[33]. The voltages applied on the CNT network with respect to the Al reference predominantly drop across the CNT/HfO_2/TiO_2 capacitor locally, driving electrons to hop through the HfO_2 dielectric layer. The low capacitance also allows the charge stored in the TiO_2 layer to influence the hole concentration and conductance of the CNT network significantly. The Fermi energies of the p-type CNT, Al, and TiO_2 materials are approximately equal (with differences less than $0.2\ eV$),[31, 34] resulting in the symmetric Fermi energy differences, and similar charge and conductance modification rates by the positive and negative voltages with the equal magnitude following the learning algorithm $\dot{w} = \alpha\ V_i \cdot V_o$ (Equation 2).

When V_i and V_o voltages with $V_i \cdot V_o = 0$ were applied on a synstor, its electronic band structures were also simulated under the conditions of (1) $V_i = 1.75\ V$ and $V_o = 0$ (Figure 3c), (2) $V_i = -1.75\ V$ and $V_o = 0$ (Figure 3c), and (3) $V_i = 0$ and $V_o = 0$ (Figure S8b). Under these asymmetric V_i and V_o voltages, the positive V_i or V_o voltage mainly drops across the reverse-biased Schottky contact between the Al electrode and p-type CNTs, and the negative V_i or V_o voltage mainly drops across the hole-depletion region on the lateral space beyond the TiO_2 charge storage layer/Al reference electrode, which leads to small differences ($\lesssim 0.23\ eV$) between the Fermi energies of the CNT network and the recessed TiO_2 layer (Figure 4a). The changes of the charge density in the TiO_2 layer, $\Delta\rho_s$, induced by the V_i and V_o voltages under $-2\ V \le V_i \le 2\ V$ and $V_o = 0$ were measured by capacitance-voltage tests on a synstor (Supporting Information and Figure S9), and plotted as a function of V_i and V_o in Figure 4b. When the synstor experienced V_i and V_o voltages with $V_i \neq 0$ and $V_o = 0$, the observed charge density changes, $|\Delta\rho_s| < 25\ nF/cm^2$, which are less than 15% of the charge density changes induced by the paired pulses with the same magnitude (Figure 2d and 2e). A control device was fabricated with a structure similar to the synstor, but in which the Al input and output electrodes were replaced by Au ones. An Ohmic contact was formed between the Au electrode and the p-type CNTs, and a V_i or V_o voltage alone significantly modified the conductance of the control

device (Figure S10), which confirms that the large Al/CNT Schottky barriers are necessary to keep w nonvolatile under $V_i \cdot V_o = 0$.

The concurrent signal inference and learning were demonstrated in a 4×2 crossbar circuit (**Figure 5**a) composed of 72 synstors and 2 integrate-and-fire "neuron" circuits with the functions according to the Hodgkin–Huxley neuron model[35] (Supporting Information and Figure S11). Nine synstors were connected in parallel with an input, an output, and a reference electrode with serpentine structures at each cross point of the electrodes to reduce noises from the devices (Figure 5b). Original speech signals consisted of unlabeled "yes" and "no" utterances were pre-processed to generate the wave of voltage pulses, $V_i^m(t)$, input to the crossbar synstor circuit (Figure 5a and Supporting Information). The input pulses had an amplitude of 1.75 V or -1.75 V, a duration of 10 ns, and firing rates, r_i^m, proportional to mel frequency cepstral coefficients (MFCCs)[36] of the speech signals at the different frequency ranges (Figure 5c and Supporting Information). The synstor circuit processed the $-1.75\ V$ $V_i^m(t)$ pulses and triggered currents by following $I^n(t) = \sum_m \kappa^{nm} \circledast (w^{nm} V_i^m)$ (Equation 1) under $V_o^n = 0$, which flowed into the integrate-and-fire "neuron" circuits to trigger 10 ns-wide $\pm 1.75\ V$ back-propagating pulses, $V_o^n(t)$, on the output electrodes (Figure 5d), and forwarding-propagating $1.0\ V$ output pulses, $V_f^n(t)$ (Figure 5e). The firing rates of the V_f^n pulses, r_f^n, increased monotonically as a nonlinear sigmoid function of I^n (Figure S11b). When a V_f^n pulse was triggered from the n^{th} "neuron" as a "winner" (i.e. $r_f^n > r_f^{n'}$, $n \neq n'$), a series of 10 ns-wide $-1.75\ V$ V_o pulses was triggered on the n^{th} output electrode of the "winner", and then a series of 10 ns-wide $1.75\ V$ V_o pulses was triggered on the n'^{th} output electrode of the "loser" (Figure 5d and 5e). To reduce the currents in the circuit, no voltage pulses with opposite polarity were applied on the input and output electrodes simultaneously. In the parallel unsupervised learning process, when the waves of V_i^m and V_o^n pulses encountered each other in the synstors (Figure 5f and 5g), the conductance matrix $[w^{nm}]_{NM}$ was modified by following Equation 2 and a "winner-take-all" learning algorithm[37], $\dot{w}^{nm} = \alpha\ V_i^m V_o^n \propto (r_f^n - r_f^{n'}) r_i^m$ (Supporting Information, Equation S6), resulting in the change of r_f^n, $\delta r_f^n \propto (\bar{r}_{f,Y}^n - \bar{r}_{f,Y}^{n'})$ for "yes" or $\delta r_f^n \propto (\bar{r}_{f,N}^n - \bar{r}_{f,N}^{n'})$ for "no", with $\bar{r}_{f,Y}^n$ and $\bar{r}_{f,N}^n$ as the average firing rates of the V_f^n pulses triggered by "yes" word and "no" words, respectively. (Supporting Information, Equation S7). Before learning started (i.e. $V_o = 0$ when $t < 16\ s$), the synstors had unspecified random conductances, the "yes" and "no" words triggered V_f pulses from the two "neurons", which

were not distinguished or orthogonal. After the learning started ($t > 16\ s$), the "yes" words triggered asymmetric V_f pulses from the two "neurons" with $\bar{r}^2_{f,Y} > \bar{r}^1_{f,Y}$, thus "neuron" 2 was the "winner" with $\delta r^2_{f,Y} \propto (\bar{r}^2_{f,Y} - \bar{r}^1_{f,Y}) > 0$, and $r^2_{f,Y}$ increased and stabilized at $\hat{r}^2_{f,Y} \approx 16\ Hz$; $\delta r^1_{f,Y} \propto (\bar{r}^1_{f,Y} - \bar{r}^2_{f,Y}) < 0$, and $r^1_{f,Y}$ decreased to $\hat{r}^1_{f,Y} = 0$. The "no" words triggered asymmetric V_f pulses from the two "neurons" with $\bar{r}^1_{f,N} > \bar{r}^2_{f,N}$, thus the "neuron" 1 was the "winner" with $\delta r^1_{f,N} \propto (\bar{r}^1_{f,N} - \bar{r}^2_{f,N}) > 0$, and $r^1_{f,N}$ increased and stabilized at $\hat{r}^1_{f,N} \approx 14\ Hz$; $\delta r^2_{f,N} \propto (\bar{r}^2_{f,N} - \bar{r}^1_{f,N}) < 0$, and $r^2_{f,N}$ decreased to $\hat{r}^2_{f,N} = 0$ (Figure 5e). After the circuit processed and learnt from 2-3 unlabeled speech signals, the "yes" and "no" speech signals were stably mapped to two distinguishable orthogonal waves of output pulses with average firing rates $\vec{\hat{r}}_{f,Y} \approx \begin{bmatrix} 0 \\ 16 \end{bmatrix} Hz$ and $\vec{\hat{r}}_{f,N} \approx \begin{bmatrix} 14 \\ 0 \end{bmatrix} Hz$. In comparison with computers, the 4×2 synstor circuit concurrently executed the signal inference (Equation 1) and learning (Equation 2) algorithms with an equivalent computational speed of ~$2.4 \times 10^9\ FLOPS$ (Supporting Information, Equation S8), a power consumption of ~15 nW (Supporting Information, Equation S9), and an equivalent computational energy efficiency of ~$1.6 \times 10^{17}\ FLOPS/W$ (Supporting Information, Equation S10).

We have demonstrated a synaptic resistor (synstor) with analog convolutional signal processing, correlative learning, and nonvolatile memory functions. The device is composed of a p-type semiconducting CNT network which formed Schottky contacts with input and output Al electrodes as a resistor, and a recessed TiO_2 charge storage layer embedded in a HfO_2 dielectric layer sandwiched between an Al reference electrode and the CNT network as a capacitor. For inference, a synstor processes a series of voltage pulses, $V_i(t)$, on its input electrode by charging the capacitor during the pulses, and discharging the capacitor after the pulses, and triggering a current via the CNT resistor, $I(t) = \kappa \circledast (w\, V_i)$ (Equation 1 and Supporting Information, Equation S1) on its grounded output electrode ($V_o = 0$) as a convolution of $V_i(t)$ and the product of its DC conductance, w, and a kernel function $\kappa(t)$. When a series of paired V_i and V_o voltage pulses with the same amplitude and duration (i.e. $V_i = V_o$) are applied on the synstor simultaneously, w is modified by following the Hebbian learning rule, $\dot{w}^{nm} = \alpha\, V_i^m V_o^n$ (Equation 2), where α is a nonlinear function of the amplitudes and numbers of V_i and V_o pulses (Supporting Information, Equation S3 and S4). The paired negative (positive) pulses generate a potential difference between the CNT network and TiO_2 layer to

increase (decrease) the electronic charge stored in the TiO_2 layer, which in turn attracts (repels) the holes in the p-type semiconducting CNT network to increase (decrease) its conductance with $\alpha > 0$ ($\alpha < 0$). Otherwise, when a synstor experiences V_i and V_o pulses under the condition $V_i \cdot V_o = 0$, the V_i or V_o potential mainly drops beyond the TiO_2 charge storage layer/Al reference electrode, and the magnitudes of the potential differences between the CNT network and the recessed TiO_2 layer are below the threshold values to modify the charge stored in the TiO_2 layer, thus $\dot{w} = 0$ (Equation 2) for nonvolatile memory. A 4×2 crossbar synstor circuit connected with integrate-and-fire "neuron" circuits was demonstrated for concurrent inference and learning from "yes" and "no" speech signals. The speech signals were converted to a wave of input voltage pulses, $\vec{V}_i$, processed by the synstor circuit in parallel to generate output currents (Equation 1), which in turn triggered waves of forward-propagating output voltage pulses, $\vec{V}_f$, from the integrate-and-fire "neuron" circuits for inference, and back-propagating voltage pulses, $\vec{V}_o$, on the output electrodes of the synstors. During the inference, the conductance matrix $[w^{nm}]_{NM}$ was concurrently modified by following the correlative learning algorithm (Equation 2) in a parallel learning process, leading to the orthogonal waves of output $\vec{V}_f$ pulses to distinguish "yes" and "no" speech signals for inference.

A synstor circuit can execute spatiotemporal inference (Equation 1) and correlative learning (Equation 2) algorithms concurrently with high energy efficiency by circumventing the fundamental computing limitations in existing electronic circuits such as physically separated logic and memory units, data transmission between memory and logic, the execution of the inference and learning algorithms in serial mode in different circuits, and the signal transmissions between the circuits. The equivalent computing energy efficiency of the 4×2 synstor circuit is $1.6 \times 10^{17}\ FLOPS/W$ (Figure 1a), which exceeds the energy efficiencies of digital transistor circuits ($\sim 10^9 - 10^{11}\ FLOPS/W$),[9-11, 17] and of the analog neuromorphic circuits of memristors and PCM ($\sim 10^5 - 10^{14}\ FLOPS/W$, excluding learning algorithm computations).[25, 29] In digital serial mode, transistors operate at high conductance ($\sim 10^5\ nS$) in order to enhance computing speed ($\sim 10^9\ Hz$) and accuracy; in analog parallel mode, synstors (synapses) operate at low conductance ($\lesssim 2\ nS$), and the computing speed and accuracy of an $M \times N$ synstor (synapse) circuit increase with increasing M and N, the numbers of parallel input/output electrodes (Supporting Information, Equation S8, S9, and S10, Figure S12 and S13). The energy consumption of a synstor circuit decreases with decreasing synstor

conductance, therefore the energy efficiency of the circuit can be further improved by decreasing synstor conductance (Supporting Information, Equation S10, Figure S14). The phase shifts of the pulses and the sneak currents in an $M \times N$ crossbar synstor circuit increase with increasing circuit scale, M and N, which limits $M, N \lesssim 10^4$ for concurrent inference and learning in parallel, and can be improved by decreasing the conductance of synstors, and the resistance and capacitance of input/output electrodes (Supporting Information). The speed of an $M \times N$ crossbar synstor circuit increases with increasing circuit scale, M and N (Supporting Information, Equation S8). With the energy efficiency of the 4×2 circuit reported in this work, a $2k \times 2k$ circuit with a power consumption of $\sim 10\, mW$ could have a speed of $\sim 10^{15}\ FLOPS$ (Figure S15), exceeding the speeds of digital transistor circuits such as TPU, GPU, and FPGA ($\sim 10^{13}\ FLOPS$),[9-11] and the analog memory devices ($\sim 10^{12} - 10^{14}\ FLOPS$).[25, 29] The microscale synstor circuit has a performance density of $\sim 1.3 \times 10^{11}\ FLOPS/mm^2$, which is superior to that of nanoscale transistor circuits ($\sim 10^{9} - 10^{11}\ FLOPS/mm^2$),[9-11, 17] and inferior to that of nanoscale memristor and PCM circuits ($\sim 10^{9} - 10^{12}\ FLOPS/mm^2$).[25, 29] Based on simulation of nanoscale devices (Supporting Information), synstors could potentially be miniaturized to nanoscale ($\sim 40\, nm$) with a performance density of $\sim 10^{17}\ FLOPS/mm^2$. There is "plenty of room at the bottom" to miniaturize synstor size, scale up synstor circuits, optimize their materials and fabrication processes, improve their energy efficiency, speed, power consumption, and uniformity for concurrent inference and learning from "big data" in intelligent systems.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

C.D.D. and C.M.S. contributed equally to this work. The authors acknowledge the support of this work by the Air Force Office of Scientific Research (AFOSR) under the programs, "Intelligent Neuromorphic Network (FA9550-15-1-0056)" and "Avian-Inspired Multifunctional Morphing Vehicle (FA9550-16-1-0087)".

Received: ((will be filled in by the editorial staff))
Revised: ((will be filled in by the editorial staff))
Published online: ((will be filled in by the editorial staff))

References

[1] A. M. Turing, Mind 1950, 49, 28.

[2] M. M. Waldrop, Nature 2016, 530, 144; J. Koomey, S. Berard, M. Sanchez, H. Wong, IEEE Annals of the History of Computing 2011, 33, 46.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Nature 2016, 529, 484.

[4] R. S. Williams, Comput Sci Eng 2017, 19, 7.

[5] V. Zhirnov, R. Cavin, L. Gammaitoni, in *Minimum energy of computing, fundamental considerations, ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology*, InTech, 2014.

[6] Semiconductor Industry Association https://www.semiconductors.org/clientuploads/Resources/RITR%20WEB%20version%20FINAL.pdf, 2015.

[7] A. J. G. Hey, R. P. Feynman, *Feynman and computation : exploring the limits of computers*, Westview Press/Perseus Books, Cambridge, Mass. 2002.

[8] J. Backus, Commun Acm 1978, 21, 613.

[9] Nvida, http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pd, 2017.

[10] N. P. e. a. Jouppi, 44th Annual International Symposium on Computer Architecture (Isca 2017) 2017, 1.

[11] E. Nurvitadhi, G. Venkatesh, J. Sim, D. Marr, R. Huang, J. O. G. Hock, Y. T. Liew, K. Srivatsan, D. Moss, S. Subhaschandra, G. Boudoukh, in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ACM, Monterey, California, USA 2017, 5.

[12] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, D. S. Modha, Science 2014, 345, 668.

[13] J. Hasler, H. B. Marr, Front Neurosci-Switz 2013, 7, 118.

[14] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, Y. Shi, Nature Electronics 2018, 1, 216.

[15] J. Conti, P. Holtberg, J. Diefenderfer, A. LaRose, J. T. Turnure, L. Westfall, USDOE Energy Information Administration (EIA), Washington, DC (United States). Office of Energy Analysis, 2016.

[16] R. Kurzweil, *The singularity is near : when humans transcend biology*, Viking, New York 2005.

[17] Oak Ridge National Laboratory, America's newest and smartest supercomputer, https://www.olcf.ornl.gov/summit/, 2018.

[18] S. Zeki, Philos T R Soc B 2015, 370, 103; K. Friston, Nature Reviews Neuroscience 2010, 11, 127.

[19] Z. Chen, S. Haykin, J. J. Eggermont, S. Becker, Correlative Learning: A Basis for Brain and Adaptive Systems 2007, 1.

[20] W. Gerstner, W. M. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*, Cambridge university press, 2002.

[21] D. O. Hebb, *The organization of behavior; a neuropsychological theory*, Wiley, New York, 1949.

[22] a) C. Diorio, P. Hasler, A. Minch, C. A. Mead, Ieee T Electron Dev 1996, 43, 1972; b) X. Gu, S. S. Iyer, Ieee Electr Device L 2017, 38, 1204.

[23] a) Q. X. Lai, L. Zhang, Z. Y. Li, W. F. Stickle, R. S. Williams, Y. Chen, Adv Mater 2010, 22, 2448; b) K. Kim, C. L. Chen, Q. Truong, A. M. Shen, Y. Chen, Adv Mater 2013, 25, 1693.

[24] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. Adam, K. K. Likharev, D. B. Strukov, Nature 2015, 521, 61.

[25] a) M. Hu, C. E. Graves, C. Li, Y. N. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. S. Yang, Q. F. Xia, J. P. Strachan, Adv Mater 2018, 30; b) C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Nature Electronics 2018, 1, 52.

[26] a) A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, V. Srikumar, ACM SIGARCH Computer Architecture News 2016, 44, 14; b) P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory", presented at *ACM SIGARCH Computer Architecture News*, 2016; c) P. Yao, H. Q. Wu, B. Gao, S. B. Eryilmaz, X. Y. Huang, W. Q. Zhang, Q. T. Zhang, N. Deng, L. P. Shi, H. S. P. Wong, H. Qian, Nat Commun 2017, 8.

[27] Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Nat Mater 2017, 16, 101.

[28] S. B. Eryilmaz, D. Kuzum, R. Jeyasingh, S. Kim, M. BrightSky, C. Lam, H. S. P. Wong, Front Neurosci-Switz 2014, 8.

[29] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha, Nature 2018, 558, 60.

[30] R. Bellman, Rand Corporation., *Dynamic programming*, Princeton University Press, Princeton, 1957.

[31] Z. H. Chen, J. Appenzeller, J. Knoch, Y. M. Lin, P. Avouris, Nano Lett 2005, 5, 1497.

[32] M. Rinkio, A. Johansson, M. Y. Zavodchikova, J. J. Toppari, A. G. Nasibulin, E. I. Kauppinen, P. Torma, New J Phys 2008, 10; S. M. Sze, K. K. Ng, *Physics of semiconductor devices*, John wiley & sons, 2006.

[33] S. J. Tans, A. R. Verschueren, C. Dekker, Nature 1998, 393, 49.

[34] Z. Xu, J. H. Wu, T. Y. Wu, Q. L. Bao, X. He, Z. Lan, J. M. Lin, M. L. Huang, Y. F. Huang, L. Q. Fan, Energy Technol-Ger 2017, 5, 1820.

[35] A. L. Hodgkin, A. F. Huxley, J Physiol-London 1952, 117, 500.

[36] M. Sahidullah, G. Saha, Speech Commun 2012, 54, 543.

[37] W. Maass, Neural Comput 2000, 12, 2519.

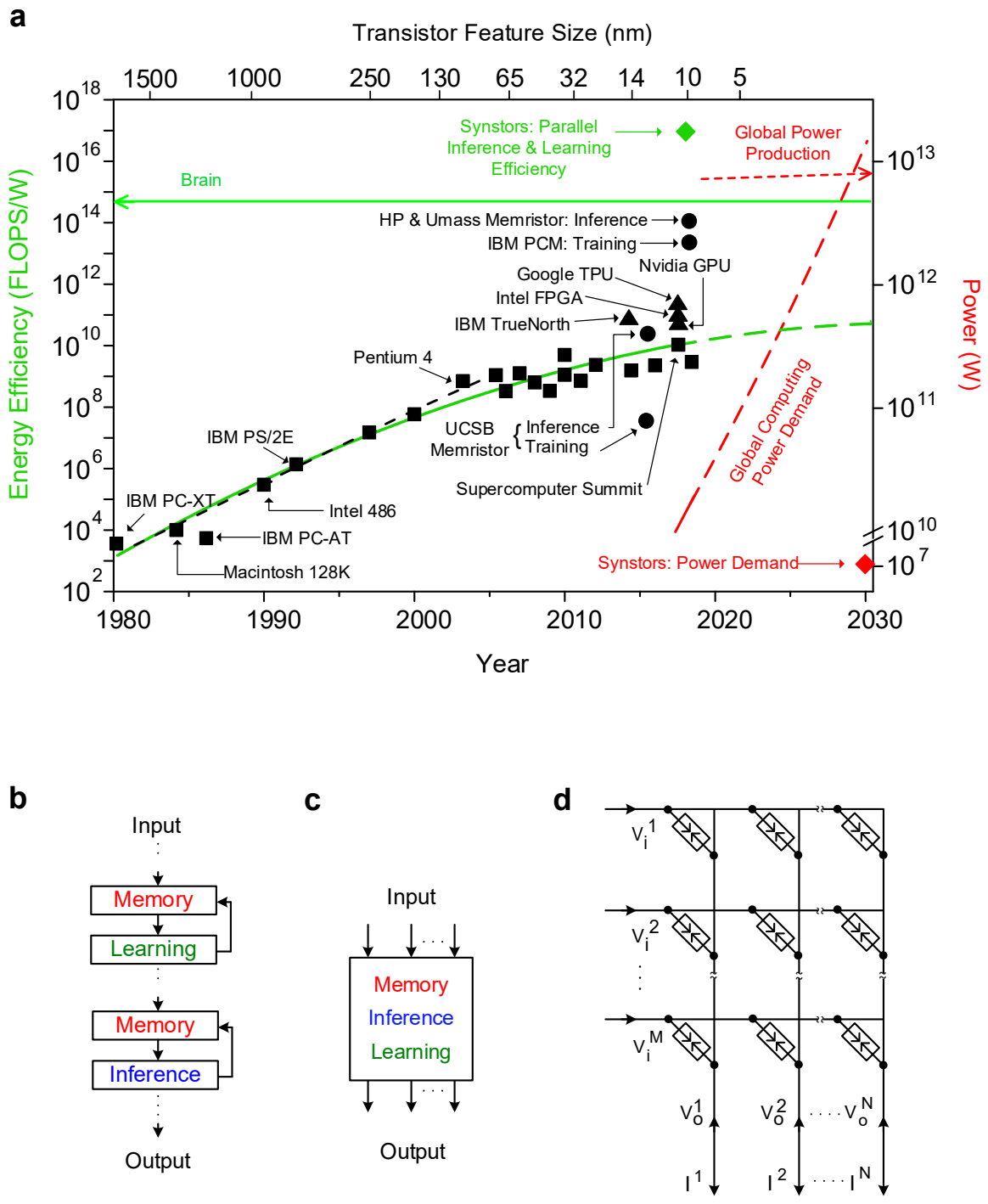

Figure 1. a) The energy efficiencies of the human brain (green line), Summit supercomputer, personal computers (squares), Volta V100 graphics processing units (GPUs) from Nvidia, tensor processing units (TPUs) from Google, Stratrix 10 field-programmable gate array (FPGA) from Intel, TrueNorth neuromorphic circuit from IBM (triangles), memristor circuits from

UCSB and UMass/HP, phase change memory (PCM) circuit from IBM (circles), and a synstor circuit reported in this work (green diamond) are shown (left y-axis) in the unit of floating point operations per second per watt ($FLOPS/W$) versus their introduction years. The different energy efficiencies of the memristor and PCM circuits for signal inference and training are displayed separately. The trend lines of the energy efficiencies of digital computers based on Si-transistors from 1975 to 2009 (black dashed line) and from 1980 to 2018 (green line) are displayed. The projected global power production (dot-dashed red line) and global computing power demands (red dashed line) based on exponentially increasing data volume and the energy efficiencies of digital computers are also displayed. The global power demand in 2030 based on the energy efficiency of synstor circuits reported in this work is shown as a red diamond. b) Based on the Turing model, a computer executes inference and learning algorithms on separated logic and memory units in serial mode with data transitions between them. c) By integrating analog convolutional processing, correlative learning, and nonvolatile analog memory functions on each synapse (or synstor), a circuit of synapses (or synstors) perform inference and learning on multi-dimensional signals concurrently in parallel mode. D) An M×N crossbar circuit to illustrate a network of synapses (synstors) connected with M presynaptic neurons (input electrodes) and N post-synaptic neurons (output electrodes). V_i^m denotes an input voltage pulse on the mth presynaptic neuron (input electrode), V_o^n denotes a back-propagating voltage pulse on the nth post-synaptic neuron (output electrode), and I^n denotes a current flowing into the nth post-synaptic neuron (output electrode).

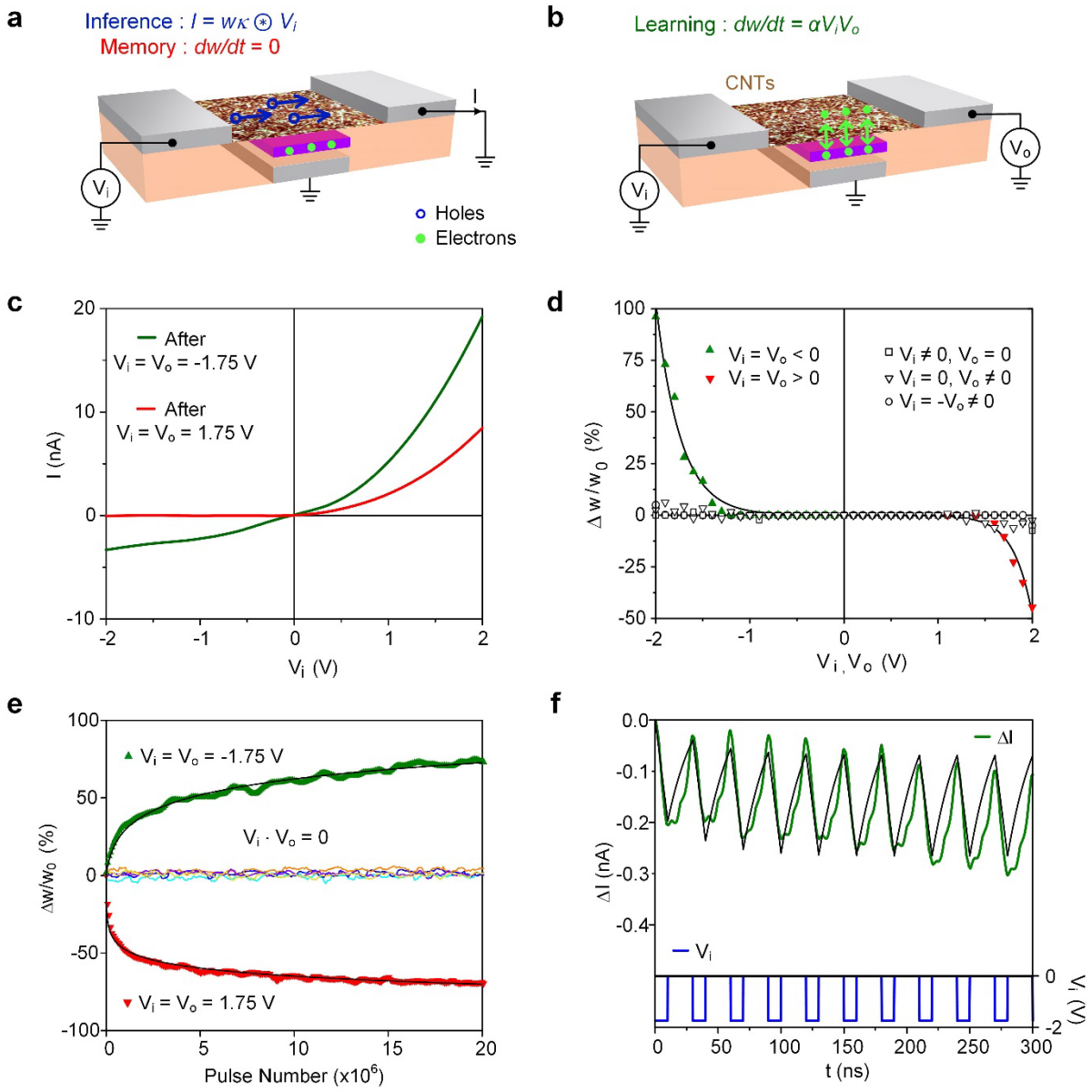

Figure 2. a), b) Schematic illustration of the structure of a carbon nanotube (CNT) synstor. An atomic force microscopy image shows the synstor comprised of a CNT network (orange) connected by Al input and output electrodes (grey). A TiO_2 charge-storage layer (purple) is embedded in a HfO_2 dielectric layer (light orange) on top of a grounded Al reference electrode (grey). a) An input voltage pulse $V_i \neq 0$ induces an output current via the CNT network under an output voltage $V_o = 0$ for inference. When $V_i \cdot V_o = 0, \dot{w} = 0, w$ remains nonvolatile for memory. b) When $V_i = V_o \neq 0$, V_i and V_o modify the electrons stored in the TiO_2 layer, resulting in the change of the hole concentration and the conductance of the p-type CNT network for learning. c)

Synstor currents, I, are plotted versus V_i, after the synstor was modified by 50 pairs of $5\ ms$-wide V_i and V_o pulses with $V_i = V_o = -1.75\ V$ (green line), and $V_i = V_o = 1.75\ V$ (red line) on the synstor. d) The percentage changes of the synstor conductance, $\Delta w/w_0$, induced by 50 pairs of various $5\ ms$-wide V_i and V_o pulses are plotted versus the pulse amplitudes. The $\Delta w/w_0$ data are fitted (solid lines) by $\Delta w/w_0 = e^{\beta_v^+(V_i - V_t^+)} - 1$ when $V_i = V_o > V_t^+$; and $\Delta w/w_0 = e^{-\beta_v^-(V_i - V_t^-)} - 1$ when $V_i = V_o < V_t^-$; $\Delta w = 0$ when $V_t^+ > V_i = V_o > V_t^-$. e) The percentage changes of the synstor conductance, $\Delta w/w_0$, induced by various $10\ ns$-wide V_i and V_o pulses with $V_i = V_o = -1.75\ V$ (green triangles), $V_i = V_o = 1.75\ V$ (red triangles), $V_i = V_o = 0$ (purple), $V_i = 0$ and $V_o = -1.75\ V$ (orange), $V_i = 0$ and $V_o = 1.75\ V$ (cyan), $V_i = 1.75\ V$ and $V_o = 0$ (yellow), and $V_i = -1.75\ V$ and $V_o = 0$ (blue) are plotted versus the applied pulse numbers, n. The $\Delta w(n)/w_0$ data are fitted (black lines) by $\Delta w(n)/w_0 = k_n^+ Ln\left(\frac{n}{n_0^+} + 1\right)$ when $V_i = V_o = 1.75\ V$; and by $\Delta w(n)/w_0 = k_n^- Ln\left(\frac{n}{n_0^-} + 1\right)$ when $V_i = V_o = -1.75\ V$. f) The change of the current, $\Delta I(t)$, (green line) triggered by a series of $-1.75\ V$ $10\ ns$-wide V_i pulses (blue line) from a synstor under $V_o = 0$ are plotted versus time and fitted (black) by $\Delta I(t) = w\ \kappa \circledast V_i$ (Equation 1) with w as the DC conductance of the device, and $\kappa(t)$ as a kernel function.

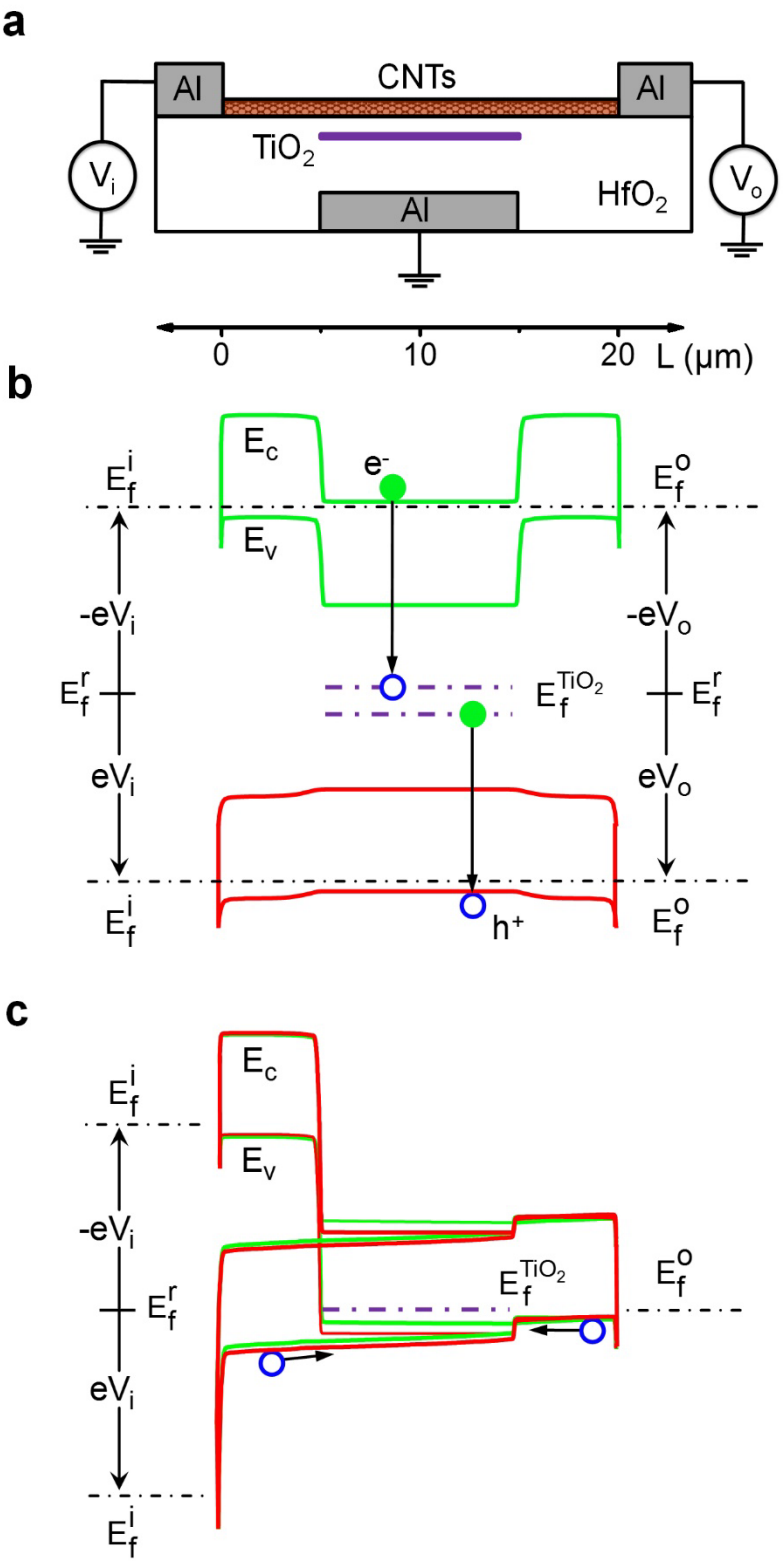

Figure 3. a) A scheme shows the cross-sectional structure of a synstor with a scale to mark the lateral distance, L. Simulated electronic band diagrams are plotted along the Al input electrode, the CNT network (orange), and the Al output electrode in the synstor under, b) $V_i = V_o = -1.75\ V$ (green), $V_i = V_o = 1.75\ V$ (red), c) $V_o = 0$, $V_i = -1.75\ V$ (top), and $V_i = 1.75\ V$ (bottom). CNT energy-band diagrams with negative charge (green lines), and positive charge (red lines) stored in the TiO_2 layer are also shown in c). E_f^i , E_f^o, E_f^r, and $E_f^{TiO_2}$ denote the Fermi energies of the Al input, output, reference electrodes, and TiO_2 charge storage layer, respectively. The electronic charge is represented by "e". E_c and E_v denote the edges of the CNT conduction and valence bands, respectively. Electrons injected into or depleted from the TiO_2 layer are illustrated as the filled green circles, and holes in CNTs, TiO_2, or transported laterally along CNTs are illustrated as the open blue circles. The purple dot-dashed lines represent the Fermi energy of the TiO_2 charge storage layer. The black dot-dashed lines represent the Fermi energies of the CNT network and the Al input and output electrodes.

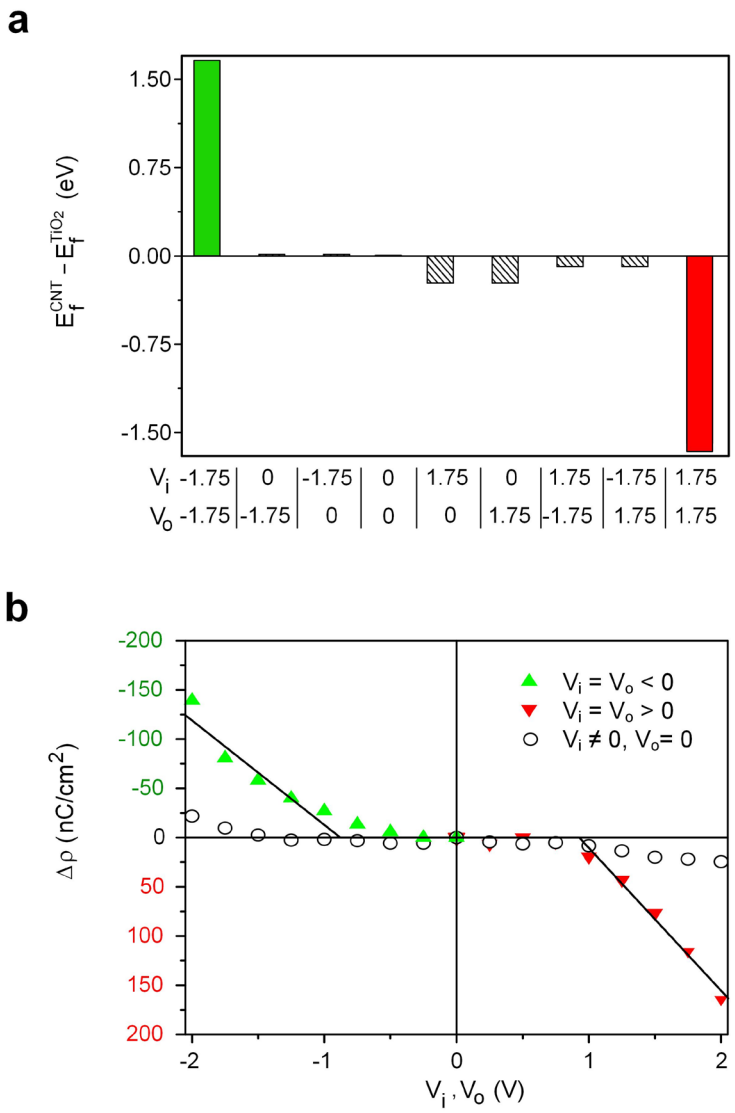

Figure 4. a) The simulated differences between the average Fermi energies of the CNT network, E_f^{CNT}, and the TiO_2 charge storage layer, $E_f^{TiO_2}$, in a synstor under various combinations of V_i voltages on its input electrode and V_o voltages on its output electrode. b) The change of the charge density in the TiO_2 layer of a synstor, $\Delta\rho_s$, induced by various V_i and V_o voltages are measured by capacitance-voltage test and plotted versus the amplitudes of the V_i and V_o voltages. $\Delta\rho_s$ data are fitted by $\Delta\rho_s = k_\rho^+[V_a - V_t^+]$ (solid lines) under $V_i = V_o > V_t^+ > 0$ and $\Delta\rho_s = k_\rho^-[V_a - V_t^-]$ under $V_i = V_o < V_t^- < 0$ with $k_\rho^+ = -145\, nF/cm^2$, $k_\rho^- = -106\, nF/cm^2$, $V_t^+ = 0.92$ V, and $V_t^- = -0.85$ V.

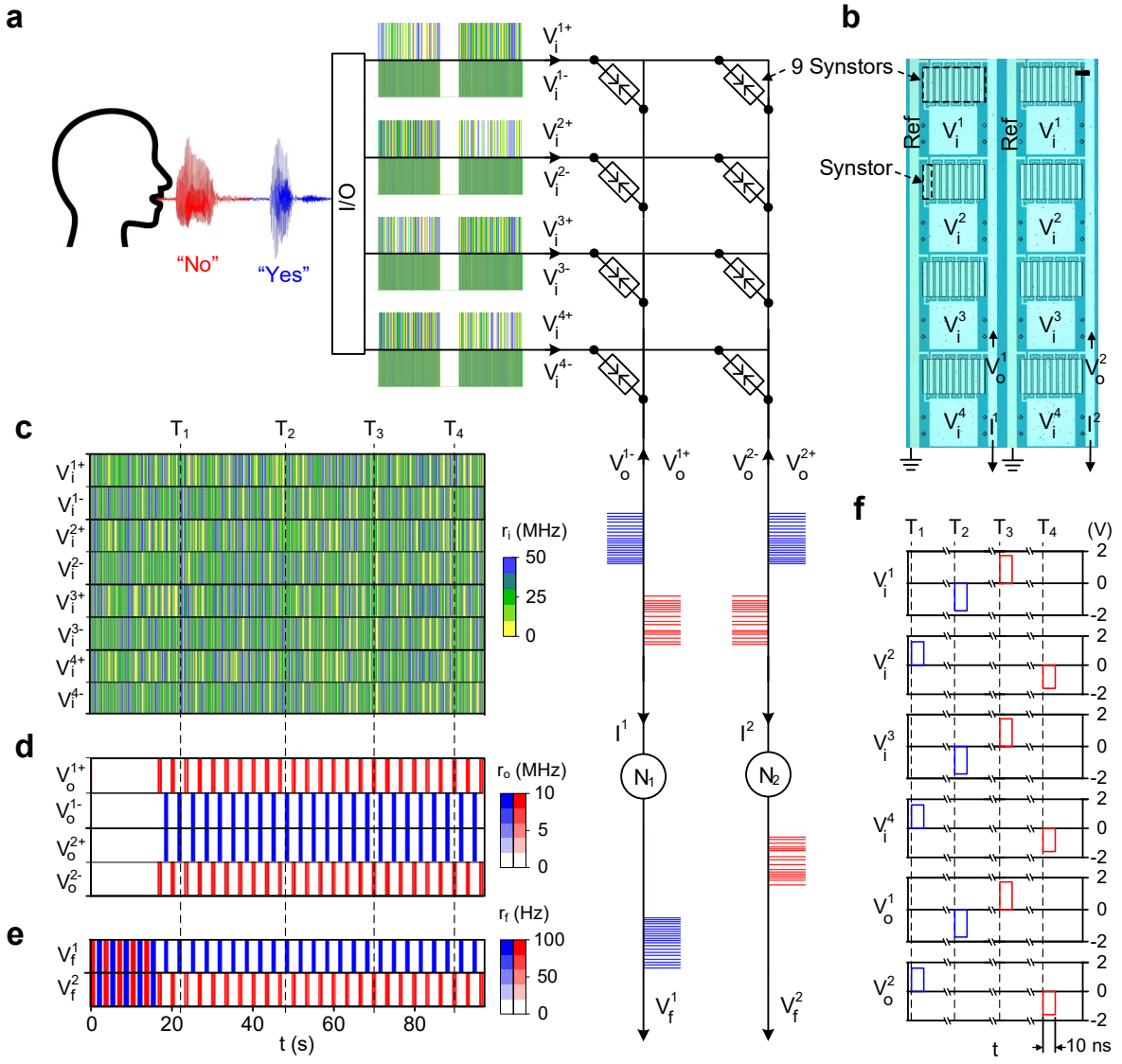

Figure 5. a) A 4 × 2 crossbar circuit composed of 72 synstors connected with two horizontal input electrodes, two vertical output electrodes. Each output electrode is connected to an integrate-and-fire "neuron" circuit, N_1 or N_2. The speech signals of "yes" and "no" words are converted to waves of voltage pulses, V_i^1, V_i^2, V_i^3, and V_i^4, input to the 1st, 2nd, 3rd, and 4th electrodes to trigger currents, I^1 and I^2, on the 1st and 2nd output electrodes via the synstors, which in turn trigger back-propagating voltage pulses, V_o^1 and V_o^2, on the 1st and 2nd output electrodes and forward-propagating voltage pulses, V_f^1 and V_f^2, from the 1st and 2nd "neurons". b) An optical image of a

synstor chip with input electrodes connected with their contact pads (labeled by V_i^1, V_i^2, V_i^3, and V_i^4), output electrodes (labeled by V_o^1 and V_o^2, and I^1 and I^2), reference electrodes (labeled by "Ref"), and CNT networks connected by the input and output electrodes. A single synstor and nine synstors are marked separately. At each crosspoint between an input and an output electrodes, the input and output electrodes are interdigitated to connect with 9 synstors, and a shared serpentine reference electrode runs between them. The scale bar is 200 μm. c) The firing rates of $\pm 1.75\ V$ 10 ns-wide pulses (V_i^{1+}, V_i^{1-}, V_i^{2+}, V_i^{2-}, V_i^{3+}, V_i^{3-}, V_i^{4+}, and V_i^{4-}) on the four input electrodes. d) The firing rates of $\pm 1.75\ V$ 10 ns-wide back-propagating pulses (V_o^{1+}, V_o^{1-}, V_o^{2+}, and V_o^{2-}) on the two output electrodes. e) The $1.0\ V$ output pulses generated from the two "neurons" (V_f^1 and V_f^2) are shown in color gradients versus time. f) A "no" word triggers $1.75\ V$ V_i^{2+} and V_i^{4+} pulses on the 2nd and 4th input electrodes and a $1.75\ V$ V_o^{2+} pulse on the 2nd output electrode concurrently, which decrease w^{22} and w^{42}, the conductances of the synstors connected with the electrodes. A "no" word triggers $-1.75\ V$ V_i^{1-} and V_i^{3-} pulses on the 1st and 3rd input electrodes, and a $-1.75\ V$ V_o^{1-} pulse on the 1st output electrode concurrently, which increase w^{11} and w^{31}. A "yes" word triggers $1.75\ V$ V_i^{1+} and V_i^{3+} pulses on the 1st and 3rd input electrodes, and a $1.75\ V$ V_o^{1+} on the 1st output electrode concurrently, which decrease w^{11} and w^{31}. A "yes" word triggers $-1.75\ V$ V_i^{2-} and V_i^{4-} pulses on the 2nd and 4th input electrodes, and a $-1.75\ V$ V_o^{2-} pulse on the 2nd output electrode concurrently, which increase w^{22} and w^{42}. The duration of all the pulses is 10 ns. The dashed lines represent the moments when the paired pulses are triggered. The speech signals and pulses triggered by "yes" and "no" words are displayed in blue and red colors, respectively in d), e), and f).