3D Synaptic Architecture with Ultralow sub-10 fJ Energy per Spike for Neuromorphic Computation

I-Ting Wang, Yen-Chuan Lin, Yu-Fen Wang, Chung-Wei Hsu, and Tuo-Hung Hou*

Department of Electronics Engineering and Institute of Electronics, National Chiao Tung University, Hsinchu, Taiwan

*Tel: +886-3-5712121 ext 54261; E-mail: thhou@mail.nctu.edu.tw

Abstract

A high-density 3D synaptic architecture based on self-rectifying Ta/TaOx/TiO2/Ti RRAM is proposed as an energy- and cost-efficient neuromorphic computation hardware. The device shows excellent analog synaptic features that can be accurately described by the physical and compact models. Ultra-low energy consumption comparable to that of a biological synapse (<10 fJ/spike) has been demonstrated for the first time.

Introduction

Existing IT systems based on Si devices and von Neumann architecture suffer from the fundamental problems of device scaling and low energy efficiency, which may ultimately impede the realization of true human-level artificial intelligence. As a result, many believe that neuromorphic computation, similar to how our brains operate, is a promising direction to pursue for future artificial intelligence systems. However, present neuromorphic computations are implemented using complex Si-based CMOS circuits or software algorithms, and cannot overcome the limitations of low density and high energy consumption. To achieve true human-level artificial intelligence, a low-power two-terminal electronic synapse—a fundamental neuromorphic computation hardware that emulates the functions of a biological synapse—must be developed. RRAM, sometimes called memristor, is now being actively developed as a low-power electronic synaptic device [1-3]. It also inspired intriguing applications, such as pattern recognition and auditory processing [4-7]. RRAM-based synapses exhibit excellent scalability, compact 4F² cell size, full CMOS compatibility, and ultralow pJ energy consumption per spike [6], outperforming other electronic synapse candidates. However, two critical challenges remain. First, high-density 3D synaptic networks rather than 2D ones are essential to realize innumerable connections among neurons in the brain (Fig. 1). However, a cost-effective 3D architecture that overcomes the sneak current problem [8] has yet to be developed. Second, reducing the energy consumption to that of a biological synapse (~10 fJ per spike [3]) remains challenging. Recently, a self-rectifying Ta/TaOx/TiO2/Ti cell has shown great potential for implementing 3D vertical RRAM (V-RRAM) arrays over 10 Mb [9-11]. In this paper, we further demonstrate promising analog synaptic features in this device, and develop complete physical-based simulation and analytical compact model to facilitate future neuromorphic system design. Furthermore, the self-rectifying characteristic enables the first high-density 3D synaptic architecture using easily fabricated V-RRAM. We also first demonstrate the ultralow energy consumption of less than 10 fJ per spike that represents 100x reduction as compared with the previous best synaptic device [3]. This extremely energy- and cost-efficient high-density 3D synaptic architecture is a breakthrough hardware for future neuromorphic computation.

Ta/TaOx/TiO2/Ti Cell

The device fabrication of the 2D Ta/TaOx (20 nm)/TiO2 (60 nm)/Ti cell was similar to that described in [9]. The device area was 10⁴ μm². Figs. 2(a) and (b) illustrate its SET-controlled and RESET-controlled multi-level-cell (MLC) operations, respectively. The gradual SET and RESET operations are crucial for analog synaptic devices where the synaptic weight (conductance) is affected by the input (training) strength. The conductance values can be read at -2 V, while very low current with no difference among different states was observed at a positive voltage bias.

Synaptic Functions

Fig. 3 illustrates alternating potentiating and depressing cycles in the Ta/TaOx/TiO2/Ti device by using consecutive 2000 training pulses. In contrast to the previous filamentary RRAM-based synapse [6], this device shows monotonically synaptic weight changes with little variation and robust cycling endurance. The potentiating and depressing ability in an identical device allows simultaneous implementation of both excitatory and inhibitory synapses. In addition, the asymmetric current readout at positive and negative biases provides a high self-rectifying ratio that suppresses sneak current through inactive synapses in the network. Furthermore, the synaptic weight function is tunable using the number of potentiating and depressing pulses (Fig. 4). Fig. 5 illustrates an extremely wide tunable range, showing no saturation after 800 pulses. A large number of analog synaptic weight states are known to improve capacity and robustness of neuromorphic systems [3]. Fig. 6 shows the measurement result of spike-timing-dependent plasticity (STDP). An action potential-like waveform (Fig. 7) was used to emulate STDP in biological synapses. STDP is a critical function for learning and memory in our brains based on the Hebbian learning rule where synaptic weight change depends on the relative timing of pre- and post-synaptic spikes. Fig. 8 illustrates the measurement result of paired-pulse facilitation (PPF). PPF is a...
short-term synaptic plasticity where reducing time interval between two sequential potentiating pulses enhances synaptic weight. The fitting time constants agree with that of biological synapses [12]. Fig. 9 illustrates memory retention characteristics. Increasing the number of training pulses (frequent rehearsal) facilitates the transition from short-term memory (STM) to long-term memory (LTM).

**Physical Simulation and Compact Model**

The switching and conduction mechanism of the Ta/TaOx/TiO2/Ti cell has been explained elsewhere [9-11] by oxygen ion migration and homogeneous barrier modulation. The physical equations shown in Fig. 10 were used to simulate the SET-controlled (Fig. 11(a)) and RESET-controlled (Fig. 11(b)) MLC, alternating potentiating and depressing cycles (Fig. 12), and STDP (Fig. 13). Furthermore, a compact model of the proposed synaptic device must be developed for future large-scale neuromorphic system design. A set of analytical equations in Fig. 14 is derived from the physical-based simulation, and thus accurately describe potentiating (Fig. 15) and depressing (Fig. 16) characteristics with multiple pulse amplitudes and STDP (Fig. 6).

**3D Synaptic Network**

Cost-effective 3D V-RRAM was fabricated as 3D synaptic network using a simple process described in [10-11]. Fig. 17 shows the cross-sectional TEM image of double-layer V-RRAM with an effective device area of 0.2 μm². The 3D device shows stable potentiating and depressing characteristics (Fig. 18) and biologically equivalent STDP (Fig. 19), similar to those in the 2D device, but the cell conductance in the area-scaled 3D device is drastically reduced because of the homogeneous current conduction [10]. Reducing cell conductance is critical for implementing high-density crossbar array [10] and lowering energy consumption per training pulse. Fig. 20 shows that the energy consumption can be as low as 7 fJ per training pulse for depression. Although the high input voltage requires further improvement, this result first demonstrates that an electronic synapse can be as energy efficient as a biological one. Fig. 21 shows the schematic and fabricated 3D (double-layer) synaptic network prototype. Both 2D feature-size scaling and increasing the number of vertically stacked layers would increase integration density in the future. Commercial 3D vertical NAND technology using a similar fabrication concept is capable of integrating 10¹¹ bits (connections) on a single chip [13].

**Conclusion**

Numerous promising synaptic features in a Ta/TaOx/TiO2/Ti RRAM-based synaptic device have been successfully demonstrated. The analog synaptic plasticity can be precisely simulated using the physical and compact models. Furthermore, the 3D synaptic network with self-rectifying characteristics for high-density integration and ultra-low training power comparable to that in biological synapses have been realized, suggesting promising potential of the 3D synaptic architecture for future neuromorphic computation.

**Acknowledgments**

This work was supported by National Science Council of Taiwan, Republic of China, under grant NSC 102-2221-E-009-188-MY3 and NCTU-UCB I-RiCE program, under grant NSC-102-2911-I-009-302.

**References**


Fig. 1 (a) Intricate 3D synaptic network connects billions of neurons in human brains. (b) Integration density is low in 2D synaptic network of present neuromorphic systems. (c) 3D implementation of synaptic network in (b) significantly increases integration density. (d) High-density 3D synaptic network emulates that in biological systems.

Fig. 2 (a) SET-controlled and (b) RESET-controlled MLC operations in the 2D Ta/TaOx/TiO2/Ti device. By varying $V_{SET}$ or $V_{RESET}$, multiple resistance states can be read out at -2 V, while very low current was observed at a positive voltage bias.

Fig. 3 Alternating potentiating (P) and depressing (D) cycles for 2000 training pulses ($P$: +9 V 50 μs/12, -8 V 50 μs) in the 2D Ta/TaOx/TiO2/Ti device. The response of the first and last 100 pulses are enlarged.

Fig. 4 Potentiating and depressing characteristics as a function of various numbers of training pulses. This phenomenon mimics biological memory enhancement by increasing stimuli.

Fig. 5 Continuous potentiating and depressing characteristics over 800 training pulses, showing monotonically synaptic weight (conductance) changes without saturation.

Fig. 6 STDP measurement shows a similar trend as that of a biological synapse, and is in good agreement with the compact model.

Fig. 7 An action potential-like waveform used to measure STDP (Fig. 6). The interval $\Delta t$ and net voltage drop are the relative timing and voltage difference between pre- and post-spike.

Fig. 8 PPF measurement (blue open square) and fitting (red line) using the listed formula ($\tau_1$=10 ms, $\tau_2$=175 ms). Inset shows the sequence of potentiating and read pulses.

Fig. 9 Memory retention (MR) as a function of various numbers of training inputs (N). Increasing the number of training pulses facilitates the transition from STM to LTM.

Fig. 10 Equations used in the physical-based simulation [11]. Equation (1) is the continuity equation used to calculate time-dependent evolution of $O^{2+}$ concentration $N_{O^2^+}$. In equation (2), $a$ is the effective hopping distance, $f$ is the attempt-to-escape frequency, $E_a$ is the activation energy of $O^{2+}$ migration, and $kT$ is the thermal energy. In equation (3), $\gamma$ is the fitting parameter for field dependence, $q$ is the elementary charge, and $F$ is the local electric field. In equation (4), $N_{O^2^+}$ is the oxygen vacancy concentration, $\varepsilon$ is the dielectric constant of TaOx or TiO2, and $\varepsilon_0$ is the permittivity of vacuum.
Fig. 11 Simulated I-V characteristics of (a) SET-controlled and (b) RESET-controlled MLC using physical equations in Fig. 10.

Fig. 12 Physical simulation of alternating potentiating and depressing characteristics, showing excellent agreement with the measurement.

Fig. 13 Physical simulation of STDP shows a similar trend as that of a biological synapse.

Fig. 14 Analytical equations derived from the physical-based simulation. Equation 5 describes the relation between current and voltage where $I_0$ and $V_0$ are fitting parameters, $g$ is the tunneling gap distance at the Ta/TaOx Schottky barrier, and $V_{READ}$ is the read voltage. Equation 6 defines the time dependence of function $g$. $V$ is the programming voltage, and $E_a$ and $\gamma$ are $g$-dependent fitting parameters.

Fig. 15 Experimental potentiating characteristics (symbols) by using various voltage amplitudes. The initial resistance state is fixed approximately at 2.4 M$\Omega$. Experiments and compact model calculation (color lines) are in good agreement.

Fig. 16 Experimental depressing characteristics (symbols) by using various voltage amplitudes. The initial resistance state is fixed approximately at 1.2 M$\Omega$. Experiments and compact model calculation (color lines) are in good agreement.

Fig. 17 Cross-sectional TEM image of 3D Ta/TaO$_x$/TiO$_2$/Ti double-layer V-RRAM. Inset illustrates the cutline where the TEM image was taken. The effective device area is 0.2 $\mu$m$^2$.

Fig. 18 Alternating potentiating (P) and depressing (D) cycles for 500 training pulses (P: +10 V 50 $\mu$s/D: -7 V 50 $\mu$s) in the 3D Ta/TaO$_x$/TiO$_2$/Ti device. The response of the first and last 100 pulses are enlarged.

Fig. 19 STDP measurement result in the 3D Ta/TaO$_x$/TiO$_2$/Ti device. The result shows a similar trend as that of a biological synapse, and is in good agreement with the compact model. The measurement input waveform is shown in Fig. 7.

Fig. 20 Depressing characteristics (-18 V/10 ns) in the 3D device. The initial resistance is ~450 M$\Omega$, and the energy consumption per spike is ~7 fJ. The required high voltage with a short pulse is in part caused by unoptimized RC delay in the device structure and measurement setup.

Fig. 21 (a) Schematic illustration and (b) top-view optical microscope image of the 3D double-layer V-RRAM array fabricated using a four-mask process. High-density 3D synaptic network in Fig. 1(d) can be realized by further increasing the number of vertical layers and 2D feature-size scaling.