MONOLITHIC 3D INTEGRATION OF DENDRITIC NEURAL NETWORK WITH MEMRISTIVE SYNAPSE, DENDRITE AND SOMA ON SI CMOS

Tingyu Li, Jianshi Tang, Junhao Chen, Xinyi Li, Han Zhao, Yue Xi, Wen Sun, Yijun Li, Qingtian Zhang, Bin Gao, He Qian and Huaqiang Wu*

School of Integrated Circuits, Beijing Innovation Center for Future Chips (ICFC), BNRist, Tsinghua University, Beijing, China

*E-mail: jtang@tsinghua.edu.cn

ABSTRACT

We report a monolithic three-dimensional integration of dendritic neural network (M3D-DNN) with memristors-based artificial synapse, dendrite and soma on top of Si-based CMOS logic. The Si CMOS layer served as control logic fabricated in foundry. A 1k-bit artificial synaptic array was built with HfO_2-based nonvolatile memristors to implement computing-in-memory (CIM). In addition, TiO_x-based memristive artificial dendrite and NbO_xN_y-based memristive artificial soma were adopted to implement the dendritic neuron (DN) layer to process postsynaptic signals. Both the CIM and DN layers were fabricated using a BEOL-compatible process. The structural integrity and proper function of each layer in the M3D-DNN were verified. Our work demonstrates a promising architecture to efficiently implement bio-plausible artificial neural networks (ANNs).

INTRODUCTION

The rise of artificial intelligence (AI) with deep learning demands for ever increasing computing power and energy efficiency. This imposes critical challenges for conventional computing hardware based on von Neumann architecture. Inspired by human brain, neuromorphic computing with bio-mimicking devices, such as memristors, emerges as a promising paradigm to break the von Neumann bottleneck and build energy-efficient AI chips [1]. Tremendous progress has been made in the past decade to use various memristors to implement ANNs with orders of magnitudes higher energy efficiency than CPU and GPU [2-4]. Most prior works have been focused on memristor-based artificial synapses with the advantage of CIM. It should be noted that, besides synapse, dendrite and soma also play vital roles in the signal processing in biological neural networks. Their functions, such as the signal filtering and nonlinear integration of dendrite, are indispensable for the extremely low power of human brain. Recently, a novel dendritic neural network (DNN) with memristors-based artificial synapse, dendrite and soma was proposed as a more bio-plausible ANN. A board-level DNN system was built to demonstrate the classifications of both static images and dynamic human motions [5-6], exhibiting significant advantages in accuracy and energy efficiency by incorporating dendrites.

In this work, inspired by the 3D nature and complex topography of brain, we demonstrate an M3D-DNN with memristors-based artificial synapse, dendrite and soma on top of Si-based CMOS logic. The ultra-dense inter-layer vias (ILVs) in M3D could facilitate the high bandwidth data transfer across different layers with significantly reduced latency and power consumption. The memristor devices were carefully optimized for these three critical computing units to fulfill their functions.

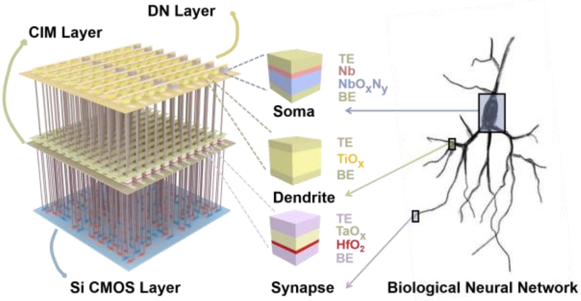

Figure 1: The architecture of M3D-DNN and its correspondence with biological neural network.

FABRICATION OF M3D-DNN

To start, the Si CMOS layer was fabricated using a standard 130nm CMOS foundry process for logic and control. The process was stopped at M4 with W vias exposed after chemical mechanical polishing (CMP). The CIM layer was then fabricated with $TiN/TaO_x/HfO_2/TiN$ memristors. The 8nm HfO_2 serving as the resistive switching layer was deposited by atomic layer deposition (ALD) at 300 ℃, followed by sputtering 45nm TaO_x serving as the thermal enhanced layer. The TiN top and bottom electrodes (TE and BE) were deposited by sputtering. The memristors were then patterned by reactive ion etching (RIE) and passivated.

After that, the DN layer consisting of $Ti/TiO_x/Pd$ dendrite devices and $Pt/Nb/NbO_xN_y/Pd$ soma devices was fabricated. First, 50nm Pd was evaporated as the BE of dendrites. 30nm TiO_x was sputtered using Ti target in Ar/O_2 followed by deposition of 30nm Ti as the TE of dendrite. Next, Pd was evaporated as the BE of soma. Then, 50nm NbO_xN_y was sputtered using Nb target in $Ar/O_2/N_2$ followed by the deposition of 10nm Nb. The proportion of Ar, O_2 and N_2 was strictly controlled and the thin Nb interface layer improved the yield of soma devices

[6-7]. Finally, Pt was deposited as the TE of soma. The process flow and chip images are presented in Figure 2.

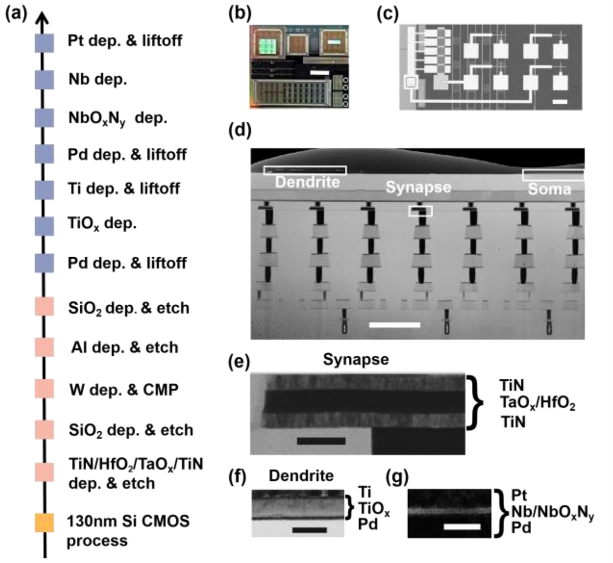

Figure 2: (a) The fabrication process flow of M3D-DNN. (b) Photo of the fabricated chip. Scale bar is 2 mm. (c) Optical image of memristors. Scale bar is 100 um. (d) Cross-sectional image of each layer. Scale bar is 2 um. (e) TEM image of HfO_2-based artificial synapse in the CIM layer (f) TEM image of TiO_x-based artificial dendrite and (g) NbO_xN_y-based artificial soma in the DN layer. Scale bars in (e), (f), (g) are 50 nm.

CHARACTERIZATIONS OF M3D-DNN

Electrical properties of each layer in the chip were measured. Figure 3a shows the analog switching characteristics of the memristive synapse in the CIM layer. Figure 3b presents the programmability of 8 representative conductance states (~3bits) read for 100 cycles, where 64 devices were measured for each state. Figure 3c-d show the mapping result and corresponding error of a 32×32 matrix using the 1k-bit one-transistor-one-resistor (1T1R) synaptic array. These results confirm the excellent analog switching characteristics of artificial synapses.

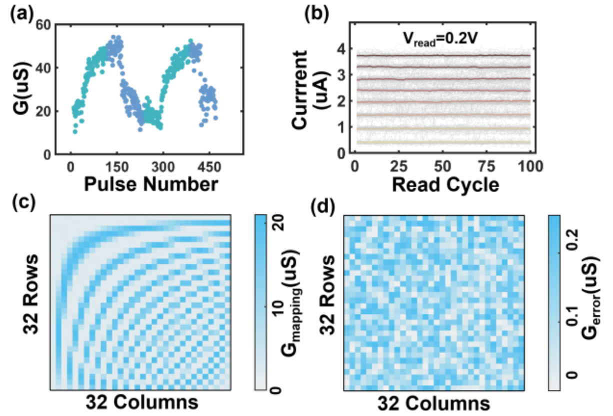

Figure 3: (a) Analog switching characteristics of a typical cell in the 1T1R synaptic array. (b) Read noise of 8 states. (c) Mapping result $G_{mapping}$ and (d) corresponding error G_{error} when mapping a 32×32 matrix in the 1k-bit array.

For the DN layer, we first characterized the dendrite device as shown in Figure 4. The device remained off when applying a voltage below the threshold (~3V), and turned on when the bias went above the threshold (e.g. 4V) [5]. It also exhibited a nonlinear current integration behavior resembling biological dendrite (Figure 4b). Figure 5a-b illustrate the relatively small cycle-to-cycle and device-to-device variability. Figure 5c depicts the dendrite device size dependence of the current response, confirming the interfacial switching mechanism. It also provided a knob to tune the device resistance to match with the soma device in the DN unit.

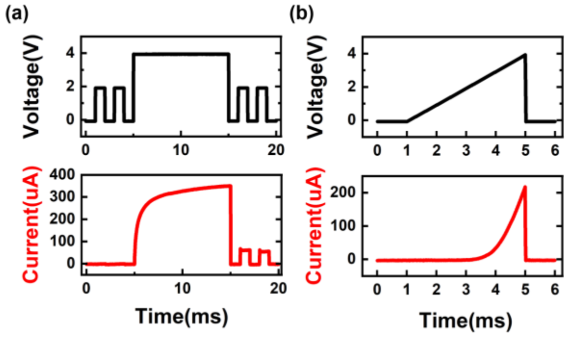

Figure 4: (a) Current response of the artificial dendrite device in the off and on states, exhibiting a filtering property. (b) Current response of the dendrite device, showing a nonlinear integration behavior.

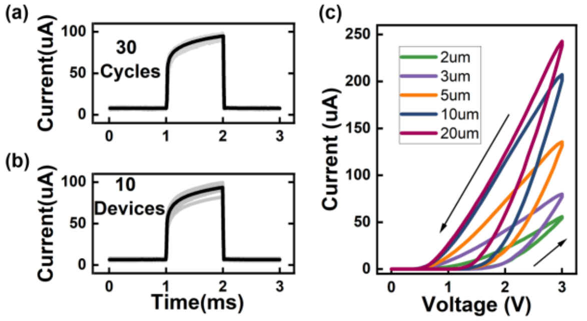

Figure 5: (a) Measured 30 cycles of a typical dendrite device with the size of 3um×3um under a 1ms pulse. (b) Measured 10 devices with the same size of 3um×3um. (c) I-V characteristics of dendrite devices with different sizes.

Furthermore, Figure 6a shows the threshold firing property of the soma device with a large window (~1V) between the threshold voltage (V_{th}) and hold voltage (V_{hold}). Besides, oscillation neuron characteristics were also demonstrated in Figure 6b-c. As the bias voltage increased, the oscillation frequency also increased linearly. The stable oscillations indicate low variability in V_{th} and V_{hold}, thanks to the N dopants in the NbO_xN_y layer that help confine the migration of oxygen vacancies [7].

NEURAL NETWORK SIMULATION

Using the above characterized artificial synapse, dendrite and soma units, a bio-plausible DNN can be implemented as illustrated in Figure 7a-b. Figure 7c shows the neuron-firing rate using the street-view house numbers (SVHN) dataset to benchmark the performance of our M3D-DNN. A high accuracy of ~89.7% was achieved by incorporating artificial dendrite, which helps improve the accuracy and reduce power consumption [5]. Figure 8 reveals that M3D-DNN could achieve 2923 × lower power consumption than GPU and 4.3× faster speed than 2D baseline owing to ultra-dense ILVs and ultra-high on-chip bandwidth in M3D architecture [5,8-10].

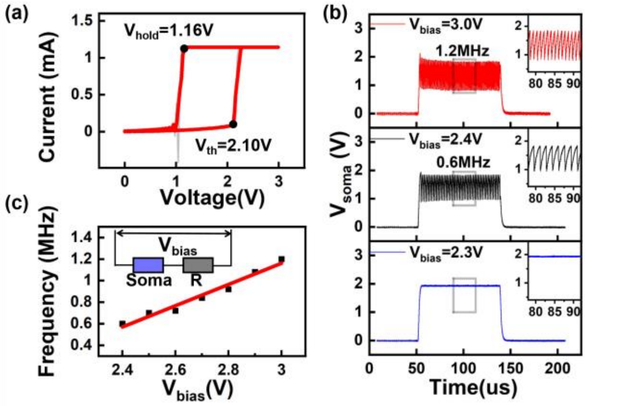

Figure 6: (a) I-V characteristic of a soma device with the size of 10um×10um. (b) Output waveforms of the oscillation neuron circuit under 100us-wide pulses with amplitudes of 3.0 V, 2.4V and 2.3V. (c) Oscillation frequency of the neuron circuit (R=3kΩ) under different V_{bias}.

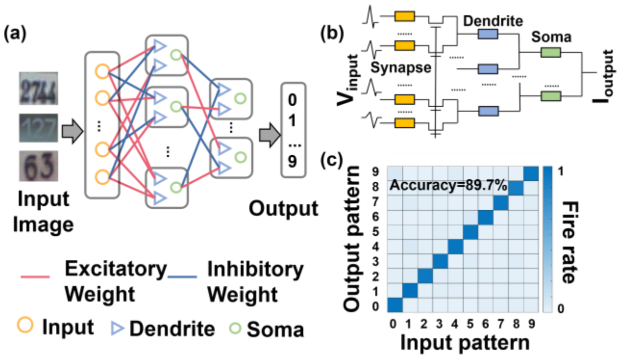

Figure 7: (a) Schematic of the implemented DNN where synapses represent tunable weights, dendrites process hierarchical post-synaptic information and somas provide the integration and firing function to yield the final output. (b) The equivalent circuit model. (c) Firing rate and recognition accuracy of SVHN dataset for M3D-DNN.

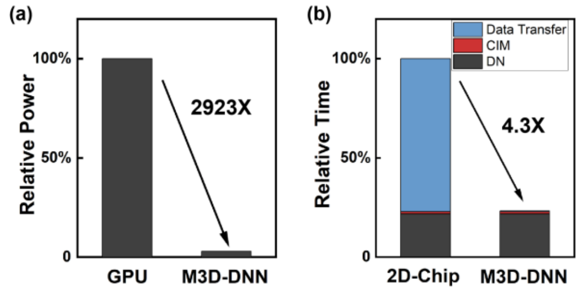

Figure 8: (a) Power consumption benchmark of M3D-DNN and GPU executing the same network. (b) Execution time of M3D-DNN and 2D baseline.

CONCLUSION

To sum up, we have designed and fabricated an M3D-DNN chip with HfO_2-based memristive synapse in the CIM layer, TiO_x-based dendrite and NbO_xN_y-based soma in the DN layer on top of Si CMOS logic layer. These three different types of memristors were carefully engineered to be compatible with BEOL process. Structural integrity and electrical properties were characterized to verify the performance of M3D-DNN. The presented M3D-DNN architecture could endow neuromorphic computing hardware with enhanced performance as well as significantly reduced energy consumption and latency.

ACKNOWLEDGMENT

This work was in part supported by the National Key Research and Development Program of China (2021YFF1200803), National Natural Science Foundation of China (92264201).

REFERENCES

[1] J. Tang et al., *Adv. Mater.*, vol. 31, pp. 1902761, 2019.

[2] M. Prezioso et al., *Nature*, vol. 521, pp. 61-4, 2015.

[3] P. Yao et al., *Nature*, vol. 577, pp. 641-646, 2020.

[4] W. Wan et al., *Nature*, vol. 608, pp. 504-512, 2022.

[5] X. Li, J. Tang, Q. Zhang et al., *Nat. Nanotechnol.*, vol. 15, pp. 776-782, 2020.

[6] X. Li et al., *Adv. Mater.*, pp. 2203684, 2022.

[7] J. Chen et al., *in IEEE Trans. Electron Devices*, vol. 69, pp. 6686-6692, 2022.

[8] Y. -J. Lee, P. Morrow and S. K. Lim, *2012 IEEE/ACM ICCAD*, pp. 539-546, 2012.

[9] M. Shulaker, G. Hills, R. Park et al., *Nature*, vol. 547, pp. 74-78, 2017.

[10] Y. Li et al., *2021 IEEE IEDM*, pp. 21.5.1-21.5.4, 2021.

[11] R. An et al., *2022 IEEE IEDM*, pp. 18.1.1-18.1.4, 2022.