3D Integration of RRAM Array with Oxide Semiconductor FET for In-Memory Computing

IIS, Univ. of Tokyo ○Jixuan Wu, Fei Mo, Takuya Saraya, Toshiro Hiramoto, and Masaharu Kobayashi

E-mail: jixuanwu@nano.iis.u-tokyo.ac.jp

1. Introduction: In-memory computing has attracted worldwide attention for deep neural network applications because of its high energy efficiency [1]. In particular, RRAM-based neural network has been extensively studied from device to system level [2-5]. 2D neural net suffers from large energy and delay in long interconnect wires. 3D neural net is a new direction enabling area-efficient, low power, and low latency computing. In this work, we propose and develop a monolithic 3D integration of RRAM array with IGZO access transistor. Then we demonstrate basic functionality of in-memory computing in the 3D neural net.

2. Device structure and Fabrication: 1T1R RRAM array with IGZO FET are integrated in spiral 3D stacking architecture where each layer is rotated by 90° from previous layer (Fig. 1(a)). Device fabrication flow is designed as simple as possible for proof-of-concept in the university lab. In each layer, IGZO FET is formed by bottom gate structure and HfO_2 gate insulator. RRAM is formed in the stack of TiN/Ti/HfO_2/TiN [6]. The process of 1T1R RRAM array is repeated 3 times for three layers devices. Fig. 1(b-d) show the top down images of FETs after completing 1st, 2nd, and 3rd layer. Process temperature is limited to 400°C.

3. Result and Discussion: We characterized IGZO FET and RRAM. Fig. 2(a) shows the I_d-V_g curves of IGZO FET for all layers. Each layer shows almost identical characteristics. Normally-off operation, nearly ideal subthreshold slope, and >200μA drive current were obtained. I-V curves of 1R cell and 1T1R cell are compared in Fig. 2(b). On-current of 1T1R cell is smaller than that of 1R cell because of the series resistance by IGZO-FET. Set and reset voltage of 1R and 1T1R cell are extracted in Fig.2(c). While 1T1R cell has almost the same set voltage as 1R cell, 1T1R cell has higher reset voltage than 1R cell. This is because series resistance by IGZO FET is relatively larger than the resistance of RRAM when RRAM is in low resistance state (LRS) before reset. Note that reducing the resistance of access transistor by higher mobility is crucial for low voltage operation and small cell area of 1T1R cell [7]. Fig.2(d) shows resistance distribution of 1T1R cells for all layers. Nearly the same distribution with the on/off ratio of >10 was obtained. No reliability degradation was found by 3D integration. We also demonstrate XNOR operation by a pair of 1T1R cells. We choose voltage sensing scheme as shown in Fig.3 [8-9].

4. Conclusion: We developed monolithic 3D integration of RRAM array with IGZO access transistor in 3D stack, confirmed each layer has uniform and almost identical device characteristics without degradation, and demonstrated functionality of in-memory computing of XNOR for 3D neural net.

References

[1] V. Sze et al., Proc. IEEE, 105, 12, 2295 (2017), [2] S. Yu, Proc. IEEE, 106, 2, 260 (2018), [3] R. Mochida et al., VLSI Symp., 175 (2018), [4] B. Yan et al., VLSI Symp., 86 (2019), [5] M. Courbariaux et al., arxiv: 1602.02830v3 (2016), [6] H. Y. Lee et al., IEDM, 297 (2008), [7] R. Yang et al., IEDM, 477 (2017), [8] Y. Zha et al., VLSI Symp., 206 (2019), [9] M. F. Chang et al., JETCAS, 5, 2, 183 (2015).

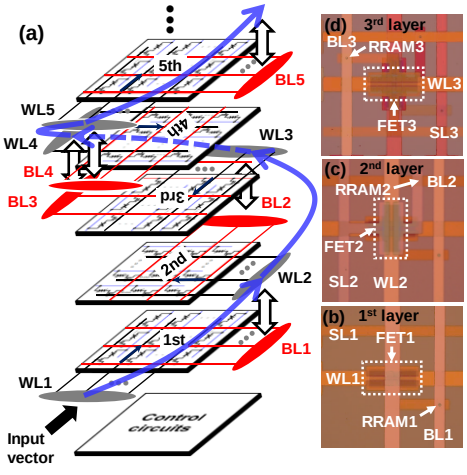

Fig. 1 (a) Schematic of the proposed spiral stacking of RRAM array. (b)~(d) Top down microscope image of fabricated IGZO FETs on 1st, 2nd, 3rd layer, respectively.

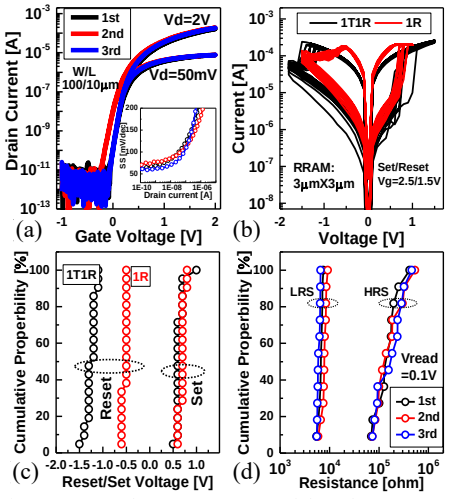

Fig. 2 Measured (a) I_d-V_g curves of three layers IGZO FET, (b) I-V curves of 1R and 1T1R device and their C2C variation, (c) cumulative probability of set/reset voltages for 1R and 1T1R. C2C variation, (d) cumulative probability of resistance of HRS/LRS. D2D variation.

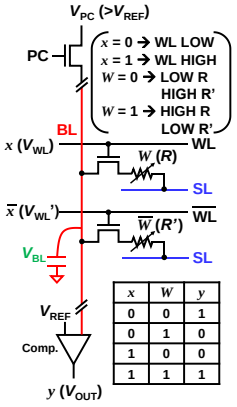

Fig. 3 Schematic of XNOR operation by a pair of 1T1R.