A Monolithic 3D Integration of RRAM Array with Oxide Semiconductor FET for In-memory Computing in Quantized Neural Network AI Applications

Jixuan Wu[1*], Fei Mo[2], Takuya Saraya[2], Toshiro Hiramoto[2], and Masaharu Kobayashi[1,2]
[1]System Design Research Center (*d.lab*), [2]Institute of Industrial Science, The University of Tokyo
* jixuanwu@nano.iis.u-tokyo.ac.jp

Abstract

We have monolithically integrated RRAM array with oxide semiconductor channel access transistor in 3D stack, achieved uniform memory characteristics of 1T1R cells at each layer, and demonstrated basic functionality of XNOR operation as in-memory computing for binary neural network AI applications, for the first time. The impact of RRAM bit error rate on neural network is also investigated. 3D neural network built by this architecture has high potential to enable area-efficient, low-power and low-latency computing.

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 [2-4]. Binary neural network (BNN) has been proposed for its simple implementation in digital hardware [5]. RRAM-based BNN has advantages such as stability, noise margin, and testability. XNOR operation for weighted sum calculation in BNN can be simply realized by RRAM cells [6-7] as in-memory computing. One challenge of BNN is the network size. Because of the low expression ability of binary weight and activation, network size needs to be large (Fig. 1). For massive parallel input/output, 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 (Fig. 2).

RRAM-only network suffers from the sneak current and programing disturbance if appropriate selector is not used. So far, 1T1R cell is the most robust structure. To stack 1T1R RRAM array, we need access transistor which can be fabricated by low temperature process in BEOL (Fig. 3). Moreover, the access transistor must have sufficiently high mobility to drive RRAM cell (Fig. 4). Oxide semiconductor such as IGZO is a promising channel material because of its high mobility and low temperature process [8-10].

In this work, we propose and develop a monolithic integration of RRAM array with IGZO access transistor in 3D stack. Then we demonstrate basic functionality of in-memory computing in the 3D neural net. The recognition accuracy of the BNN is estimated as a function of bit error rate of RRAM.

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. In neural network, the layer's output is typically connected to the next layer's input. This architecture avoids interconnect wiring overhead (Fig.5(a)).

Device fabrication flow is designed as simple as possible for proof-of-concept in the university lab (Fig. 6). In each layer, IGZO FET is formed by bottom gate structure and HfO_2 gate insulator. RRAM is formed in the stack of $TiN/Ti/HfO_2/TiN$ [11]. The process of 1T1R RRAM array is repeated 3 times. Fig. 5(b-d) show the top down images of FETs after completing 1st, 2nd, and 3rd layer. Process temperature is limited to 400°C. From TEM images in Fig. 7(a-e) and Fig. 7(f-k), we confirmed uniform IGZO FET and RRAM at each layer.

Results and Discussions

A. FET, RRAM, and 1T1R cell characteristics

We characterized IGZO FET and RRAM. Fig. 8 and 9 show I_d-V_g and I_d-V_d 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. 10. 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.11. 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 [12]. The cycle to cycle (C2C) variation of the resistance of 1T1R cell is shown in Fig.12. LRS has uniform distribution but high resistance state (HRS) has large variation. This is typical for HfO_2-based RRAM because of the large variability in filament dissociation in HRS. Fig.13 shows I-V curves of 1T1R cells for all layers. The device to device (D2D) resistance variation is extracted from Fig. 13 in Fig.14. Nearly the same distribution with the on/off ratio of >10 was obtained. Endurance and retention characteristics are shown in Fig.15 and 16. No reliability degradation was found by 3D integration.

B. In-memory computing of XNOR for binary neural net

We demonstrate XNOR operation by a pair of 1T1R cells in Fig.17 (a). We choose voltage sensing scheme [7,13]. Weight bit (W) is complementarily written on RRAMs (R, R'). Input bit (x) is complementarily applied on word lines (V_{WL}, V_{WL}'). Bit line (BL) is precharged. Then, BL is discharged with slow or fast speed depending on the input and weight bit. After certain period, BL voltage is compared with reference voltage. The output bit (y) of XNOR is obtained from the comparator. Fig. 17 (b) shows the fabricated 1T1R array. The operation is performed by using the external peripheral circuit in Fig. 17 (c). Fig.18 shows the waveforms and confirmed XNOR operation. XNOR output is digitally counted [5] or aggregated in voltage sensing at each BL [7] for weighted sum calculation.

Based on the RRAM-based XNOR, we estimate the recognition accuracy of MNIST dataset in BNN using the framework in Fig.19. As shown in Fig. 20, although the accuracy is degraded as RRAM bit error rate (BER) increases, it is not very sensitive to BER up to certain level (10ppm in this case), which indicates the property of error-resilience in BNN.

Summary

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 and error-resilient BNN for 3D neural net.

Acknowledgement

This work was supported by JST CREST (16815651), JSPS KAKENHI Grant Number JP18H01489 and Tokyo Electron Ltd.

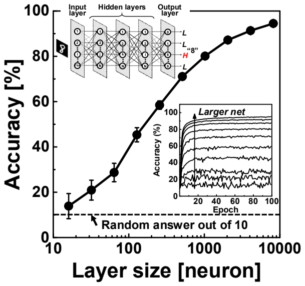


Fig. 1 Simulated recognition accuracy of MNIST dataset in multilayer perceptron of BNN as a function of layer size.

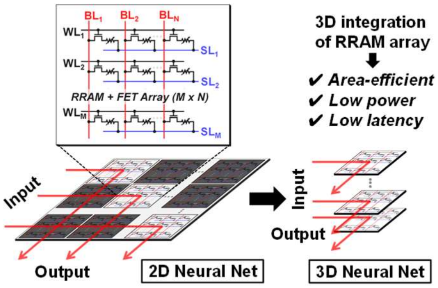


Fig. 2 The proposed concept of 3D neural net in which RRAM array is stacked vertically and in-memory computing is performed at each layer in parallel.

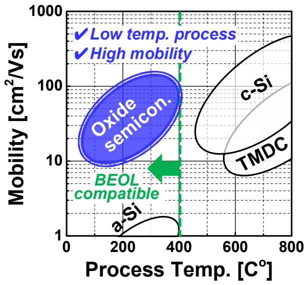


Fig. 3 Benchmark of channel materials for access transistor of RRAM cell regarding mobility and process temp.

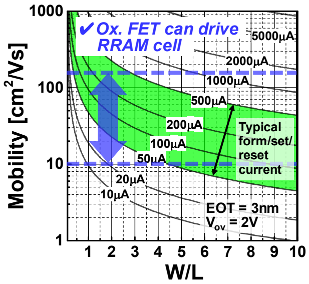
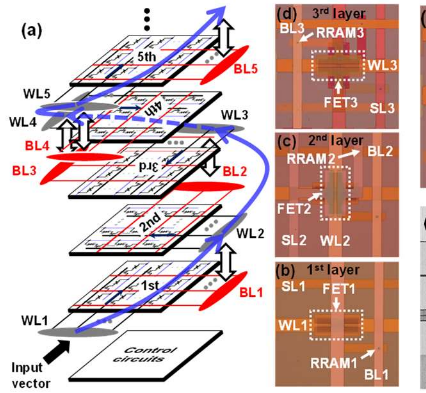


Fig. 4 Calculated contour plot of transistor drive current as a function of mobility and transistor size W/L.



(a)
5th
4th
3rd
2nd
1st
BL5 BL4 BL3 BL2 BL1
WL5 WL4 WL3 WL2 WL1
Input vector
Control circuits
(d) 3rd layer BL3 RRAM3 WL3 FET3 SL3
(c) 2nd layer RRAM2 BL2 FET2 SL2 WL2
(b) 1st layer SL1 FET1 WL1 RRAM1 BL1

Fig. 5 (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.

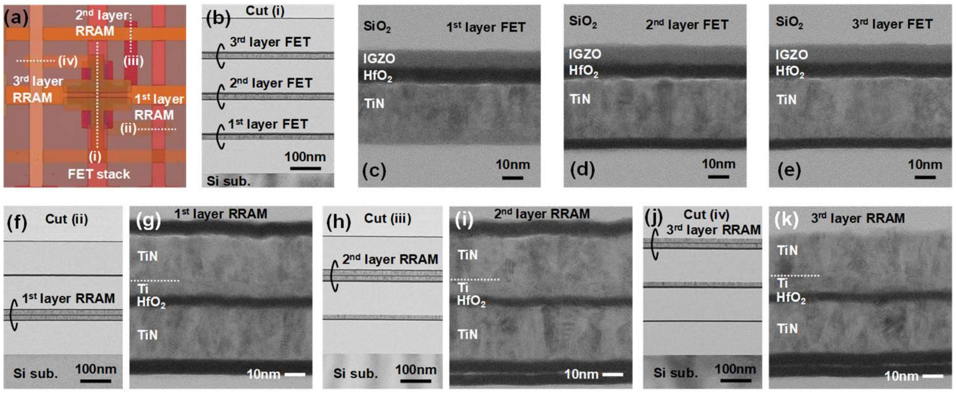
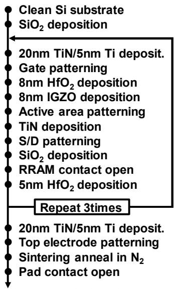


Fig. 7 (a) Top down microscope image of an FET on the 3rd layer. Cross sectional TEM image of (b) full stack of FETs, (c) 1st layer of FET, (d) 2nd layer of FET, (e) 3rd layer of FET. Cross sectional TEM image of (f) low and (g) high mag image of 1st layer RRAM, (h) low and (i) high mag image of 2nd layer RRAM, and (j) low and (k) high mag image of 3rd layer RRAM.



- Clean Si substrate
- SiO2 deposition
- 20nm TiN/5nm Ti deposit.
- Gate patterning
- 8nm HfO2 deposition
- 8nm IGZO deposition
- Active area patterning
- TiN deposition
- S/D patterning
- SiO2 deposition
- RRAM contact open
- 5nm HfO2 deposition
- Repeat 3times
- 20nm TiN/5nm Ti deposit.
- Top electrode patterning
- Sintering anneal in N2
- Pad contact open

Fig. 6 Fabrication process flow.

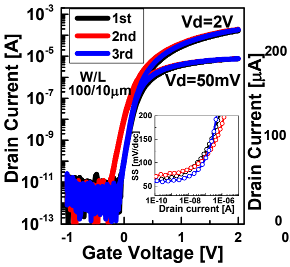


Fig. 8 Measured I_d-V_g curves of 1st, 2nd, and 3rd layer IGZO FET. The insert shows SS vs I_d.

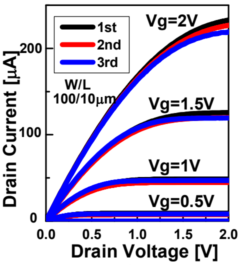


Fig. 9 Measured I_d-V_d curves of 1st, 2nd, and 3rd layer IGZO FET up to V_g=2V.

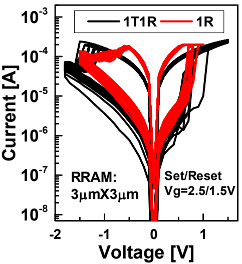


Fig. 10 Measured I-V curves of 1R and 1T1R device and their C2C variation.

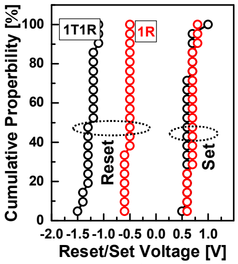


Fig. 11 Measured cumulative probability of set/reset voltages for 1R and 1T1R. C2C variation.

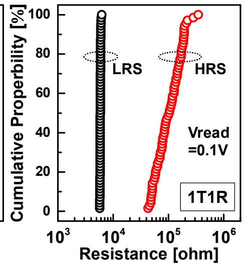


Fig. 12 Measured cumulative probability of resistance at HRS/LRS. C2C variation

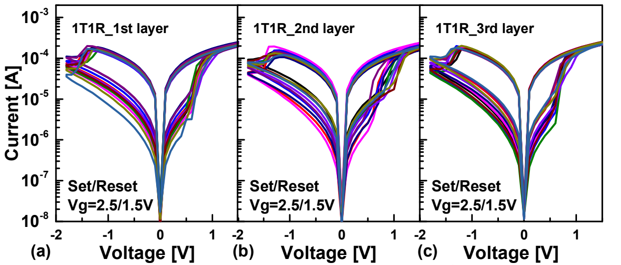


Fig. 13 Measured I-V curves of (a) 1st layer RRAM, (b) 2nd layer RRAM, and (c) 3rd layer RRAM with 1T1R configuration. Each line is an average of repeated measurement of single devise. D2D variation is shown in lines.

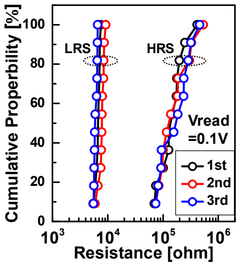


Fig. 14 Measured cumulative probability of resistance of HRS/LRS. D2D variation

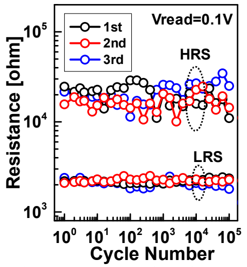


Fig. 15 Measured endurance characteristics of 1st, 2nd, and 3rd layer RRAM at room temp.

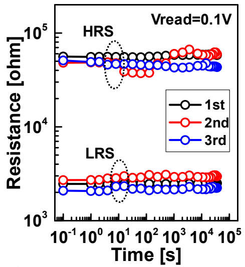
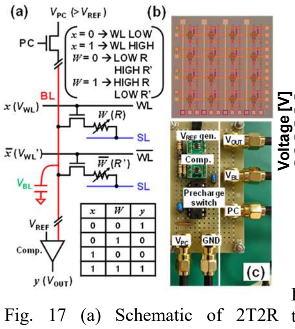


Fig. 16 Measured retention characteristics of 1st, 2nd, and 3rd layer RRAM at room temp.



x	W	y
0	0	1
0	1	0
1	0	0
1	1	1

Fig. 17 (a) Schematic of 2T2R XNOR cell. (b) Fabricated 1T1R array. (c) External peripheral circuit.

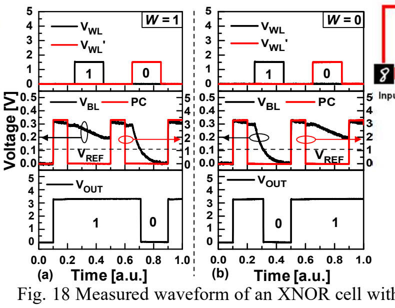
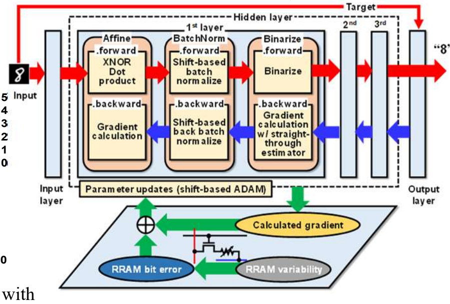


Fig. 18 Measured waveform of an XNOR cell with the peripheral circuit of Fig. 17 (c) for (a) (R, R') = (HIGH, LOW) and (b) (R, R') = (LOW, HIGH). V_{PC}=0.3V, V_{REF}=0.1V, V_{WL}=1.5V. 3.3V for circuit.



Target
Hidden layer
1st layer
2nd
3rd
Input
"8"
Affine .forward XNOR Dot product
BatchNorm .forward Shift-based batch normalize
Binarize .forward Binarize
.backward Gradient calculation
.backward Shift-based back batch normalize
.backward Gradient calculation w/ straight-through estimator
Input layer
Output layer
Parameter updates (shift-based ADAM)
Calculated gradient
RRAM bit error
RRAM variability

Fig. 19 Schematic of the digitally implementable framework of the BNN [5] incorporating RRAM BER due to RRAM cell variability.

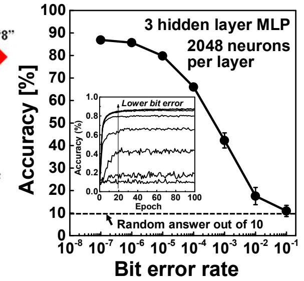


Fig. 20 Estimated recognition accuracy of MNIST dataset in BNN as a function of RRAM cell BER for layer size 2048.

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] M. Bocquet et al., IEDM, 484 (2018), [7] Y. Zha et al., VLSI Symp., 206 (2019), [8] K. Nomura et al., Nature, 432, 488 (2004), [9] M. Oota et al., IEDM, 50 (2019), [10] C.-C. Chang et al., APL, 112, 172101 (2018), [11] H. Y. Lee et al., IEDM, 297 (2008), [12] R. Yang et al., IEDM, 477 (2017), [13] M. F. Chang et al., JETCAS, 5, 2, 183 (2015).