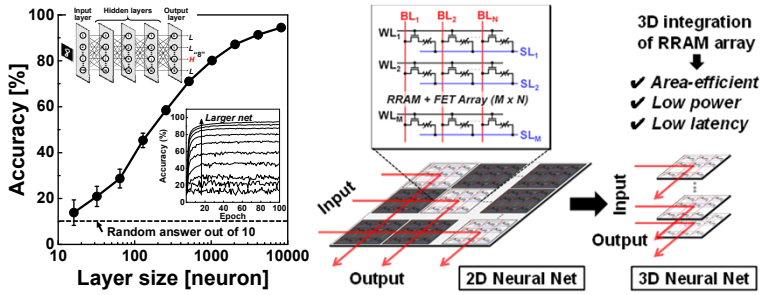
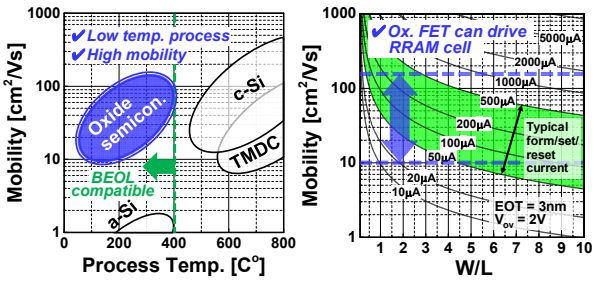
A Monolithic 3D Integration of RRAM Array with Oxide Semiconductor FET for In-memory Computing in Quantized Neural Network AI Applications

Background and Motivation



- In-memory computing has attracted worldwide attention for deep neural network applications because of its high energy efficiency. 3D neural net is a new direction enabling area-efficient, low power, and low latency computing.

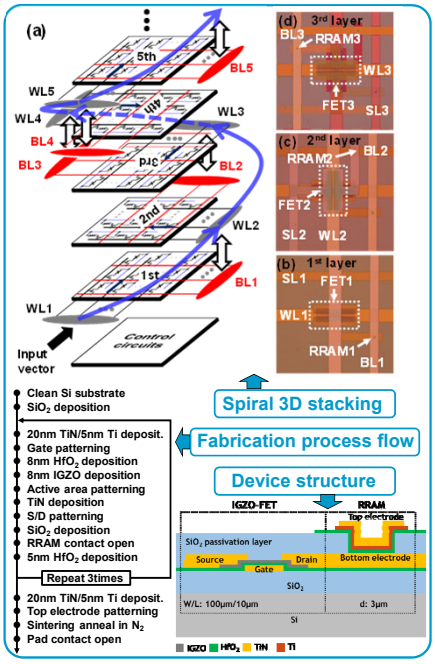
Benchmark of Channel Materials for FET

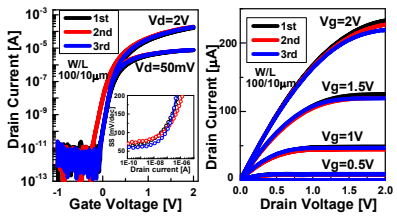


- Oxide semiconductor such as InGaZnO (IGZO) which has low temperature process and high mobility is a promising channel material to drive RRAM cell.

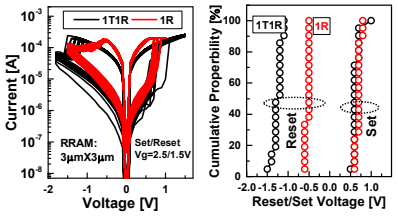
Objectives: we propose and develop a monolithic integration of RRAM array with IGZO access transistor in 3D stack.

Device Structure, Fabrication and Characteristics





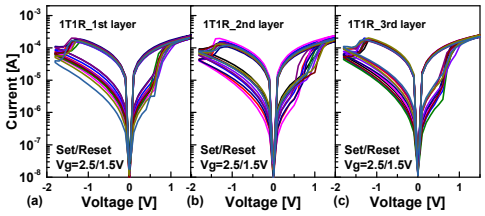
Characteristics of IGZO-FET devices at each layer with normally-off operation, nearly ideal subthreshold slope, and >200μA drive current.



Characteristics of RRAM and 1T1R device at single layer

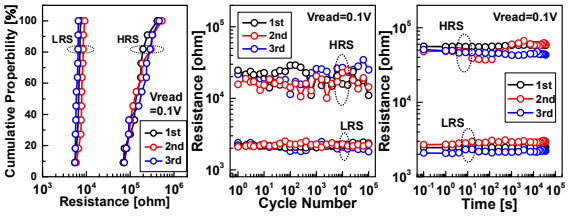
- 1T1R cell has higher reset voltage than 1R cell caused by series resistance of IGZO FET.

Device-to-Device Variation of Three Layer 1T1R Devices



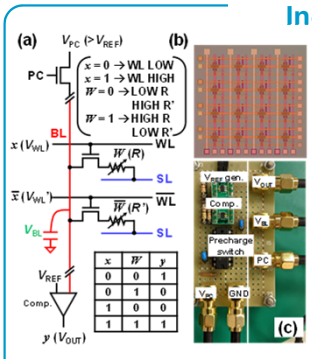
I-V Curves of 1st-3rd layer 1T1R devices

- Nearly the same distribution with the on/off ratio of >10.

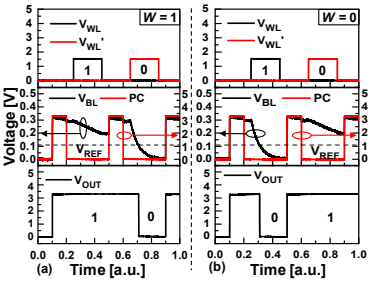


- The LRS and HRS variability of each layer is almost identical.
- No reliability degradation (endurance & retention) was found by 3D integration.

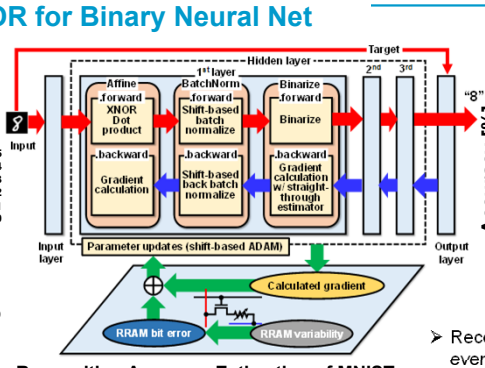
In-memory Computing of XNOR for Binary Neural Net



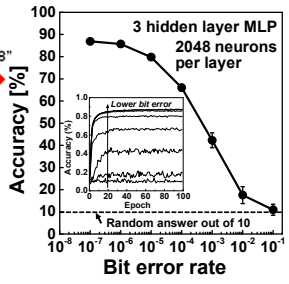
Schematic of 2T2R XNOR cell and external peripheral circuit



Waveforms and Confirmed XNOR Operation



Recognition Accuracy Estimation of MNIST Dataset in BNN



- Recognition accuracy keeps high even with high BER up to 10ppm which indicates the property of error-resilience in RRAM BNN.

Summary

1) We have developed monolithic 3D integration of RRAM array with IGZO access transistor in 3D stack.
2) Functionality of in-memory computing of XNOR and error-resilient BNN for 3D neural net are demonstrated.
3) 3D neural network built by this architecture has high potential to enable area-efficient, low-power and low-latency computing.

This work was supported by JST CREST, JSPS KAKENHI, and Tokyo Electron Ltd.

東京大学生産技術研究所