RRAM for Edge AI: from Device, Circuit to System

Bin Gao*, Yuyao Lu, Xueqi Li, Pufan Xu, Jianshi Tang

School of Integrated Circuits (SIC), Tsinghua University

The demand for higher intelligence and thus more complicated AI algorithm on edge devices is explosively increasing. Consequently, the energy efficiency of the current computing paradigm requiring massive amounts of data transfer between computing and memory units, has been overwhelmed by the surging calculation workload. In sharp contrast, analog RRAM-based computing-in-memory (CIM) paradigm could significantly reduce data movement and achieve superior energy efficiency, which is critical for edge devices. We have explored the entire analog RRAM-based technological path from RRAM device to RRAM-based CIM system.

At device level, good analog switching property enables single RRAM device to represent high precision number, which is the key to high energy efficiency in edge AI devices. To modify analog switching property, we proposed a thermal enhanced layer (TEL) RRAM [1] (Fig. 1a) to induce multiple weak conductive filaments instead of single strong filament (Fig. 1b) in the switching layer. With multiple weak filaments, the resistive state of TEL RRRAM could be modulated gradually and continuously, demonstrating outstanding analog switching property.

At circuit level, we designed a fully integrated analog RRAM-based CIM chip [2] (Fig. 2a) featuring high energy efficiency. To reduce accumulative current and IR drop, we adopted a sign-weighted 2T2R cell with current subtraction technique. To alleviate the power overhead of interface circuit blocks, we designed low power and resolution-adjustable ADC enabling tradeoff between accuracy and energy consumption.

At system level, we proposed a sign- and threshold-based learning (STELLAR) architecture [3], enabling high energy-efficient on-chip learning in RRAM-based CIM edge devices. To circumvent high cost precise weight tuning and RRAM device nonideality, STELLAR adopts an efficient verification-free weight update scheme (Fig. 2b), where only weight update direction is calculated, and accordingly a fixed set or reset pulse is alternately applied to the 2T2R weight cells in row-wise parallel fashion. Furthermore, to lower weight update frequency and improve learning convergency, STELLAR predefines a threshold filtering out small updates. By defining an appropriate threshold, STELLAR could achieve dramatic energy efficiency advantage with little accuracy loss compared with conventional backpropagation algorithm (Fig. 2c). Finally, we implemented STELLAR on the aforementioned RRAM-based CIM chip, and demonstrated several improvement learning tasks, including motion control for a light-chasing car, image classification, and audio recognition and all exhibited good accuracy and high energy efficiency.

References

[1] W. Wu, H. Wu, B. Gao, N. Deng, S. Yu, and H. Qian, IEEE EDL, 38 (2017) 1019-1022.

[2] Q. Liu, B. Gao, P. Yao, et al., IEEE ISSCC, San Francisco, CA, USA, (2020) 500-502.

[3] W. Zhang, P. Yao, B. Gao, et al., Science, 381 (2023) 1205–1211.

* Corresponding author: email: gaob1@tsinghua.edu.cn

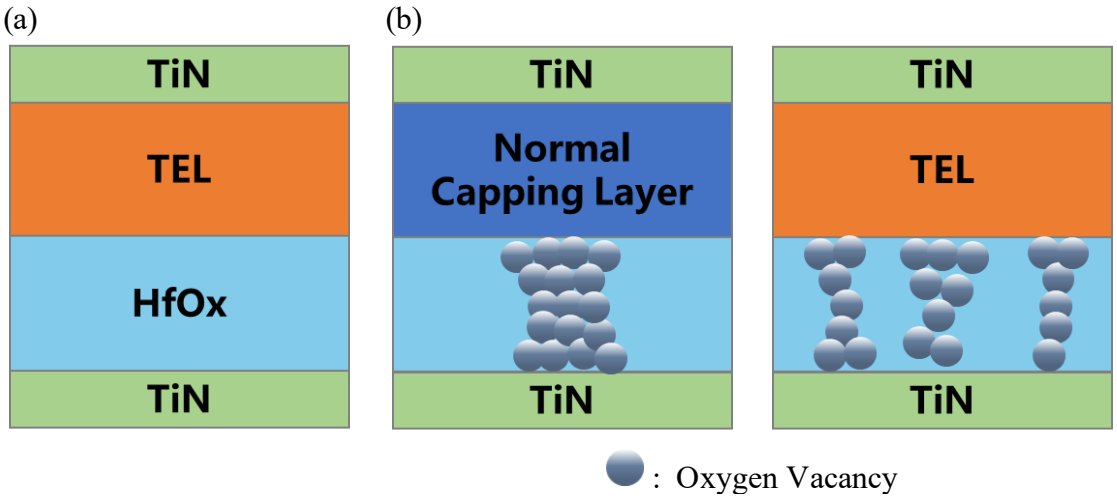

Fig. 1. TEL RRAM device [1]: (a) device material stack; (b) single strong filament (with normal capping layer) vs. multiple weak filaments (with TEL).

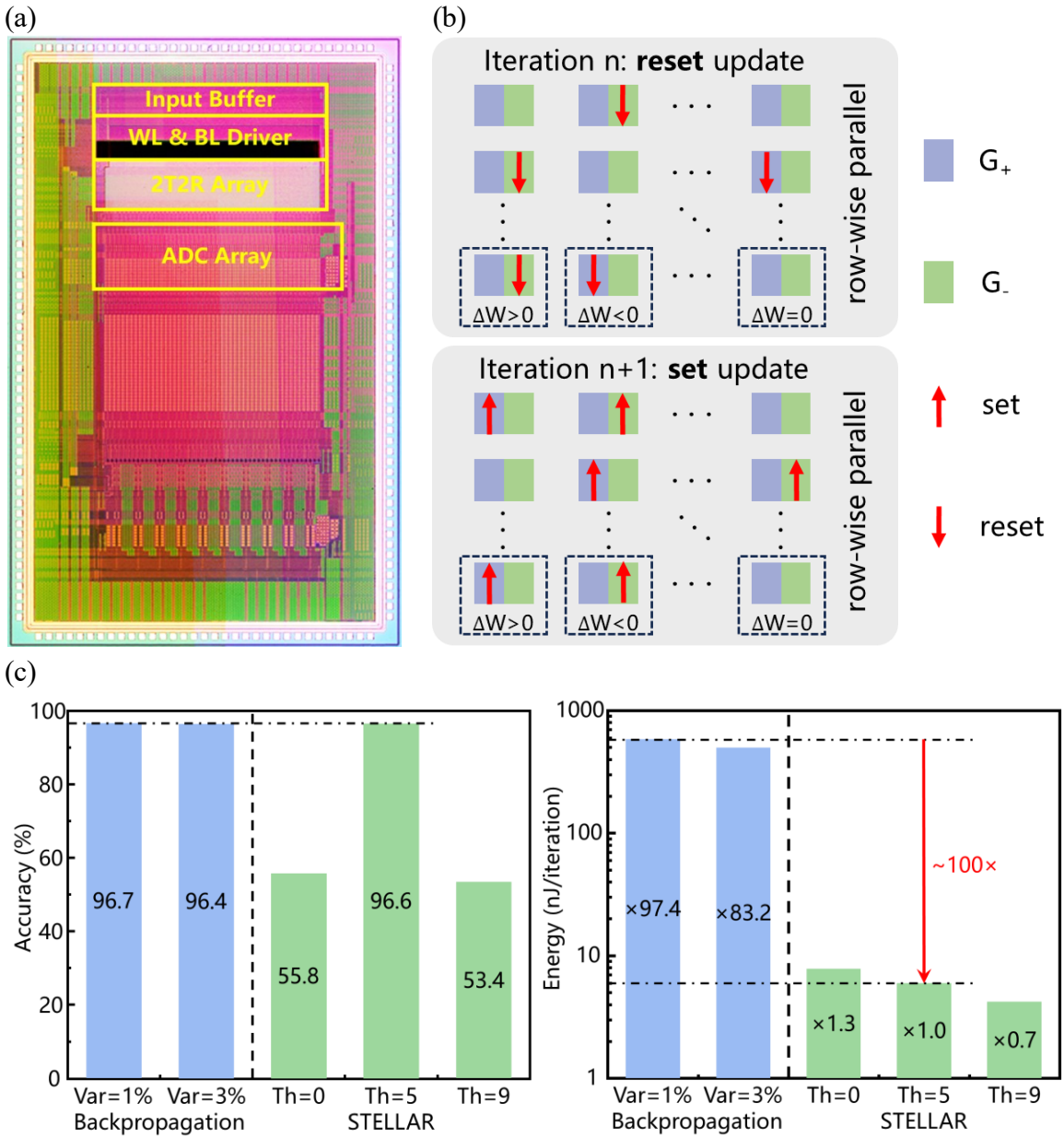

Fig. 2. (a) Fully integrated analog RRAM-based CIM chip [2]; (b) efficient verification-free weight update scheme [3]; (c) accuracy and energy consumption: backpropagation algorithm vs. STELLAR architecture [3].