Monolithic 3D Integration of Logic, Memory and Computing-In-Memory for One-Shot Learning

Yijun Li[1], Jianshi Tang[1,*], Bin Gao[1], Jian Yao[2], Yue Xi[1], Yuankun Li[1], Tingyu Li[1], Ying Zhou[1], Zhengwu Liu[1], Qingtian Zhang[1], Song Qiu[2], Qingwen Li[2], He Qian[1], and Huaqiang Wu[1,*]
[1]School of Integrated Circuits, Beijing Innovation Center for Future Chips, BNRist, Tsinghua University, Beijing, China;
[2]Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Science, Suzhou, China
*Email: jtang@tsinghua.edu.cn; wuhq@tsinghua.edu.cn

***Abstract*—We demonstrate a monolithic 3D integration of Si-based CMOS logic, resistive random-access memory (RRAM) based computing-in-memory (CIM) and ternary content-addressable memory (TCAM) layers, namely M3D-LIME, to implement one-shot learning. The first layer of Si MOSFETs was designed and fabricated using a standard CMOS process and served as control logic. The second layer of 1T1R array was fabricated with $HfAlO_x$-based analog RRAM using a low-temperature (≤ 300 °C) back-end-of-line (BEOL) process to implement CIM for feature extractions. The third layer of 2T2R-based TCAM was fabricated with carbon nanotube field-effect transistors (CNTFETs) and Ta_2O_5-based RRAM to perform template storing and matching. Extensive structural analysis and electrical measurements were carried out to validate the integrity and proper function of the fabricated M3D-LIME chip. As a demonstration, GPU-equivalent classification accuracy up to 97.8% was achieved in the one-shot/few-shot learning task on the Omniglot dataset with 162× lower energy consumption. Our work demonstrates the feasibility and great potential of M3D chips consisted of logic, memory and CIM for emerging applications such as artificial intelligence (AI) and high-performance computing (HPC).**

I. Introduction

Amid the slowdown of conventional Moore's law scaling, M3D emerges as an appealing technology to continuously increase the integration density, enrich the functionality and boost the performance of a single chip. Using BEOL-compatible fabrication processes, M3D monolithically integrates multiple layers of different functions which are connected through fine-grain and dense vertical inter-layer vias (ILVs) for high-bandwidth data transfer between different layers [1]. As a result, M3D chips are expected to show significant improvements in terms of speed, power and energy efficiency for data-abundant applications like AI [2]. However, the design and fabrication of M3D chips is very challenging, for which multi-layer BEOL-compatible high-performance logic and memory devices are needed. For this purpose, CNTFETs and RRAM are promising candidates for M3D due to their low-temperature substrate-independent fabrication processes and outstanding electrical performance. For example, high-speed complementary metal–oxide–semiconductor (CMOS) digital circuits and microprocessors have been demonstrated with CNTFETs, which can potentially provide nearly ten times performance improvement over Si counterparts [3-5]. On the other hand, different types of RRAM devices, such as $HfAlO_x$-based analog-type RRAM with multi-level switching characteristics and Ta_2O_5-based digital-type RRAM with large on/off ratios, have been developed for various applications.

With high data bandwidth and low latency communication between logic and memory layers, M3D could play important roles in the era of big data and AI, where computing becomes increasingly memory-centric. For example, as a bio-plausible learning strategy with minimum training cost, one-shot or few-shot learning recently attracts more and more research attention, and it can be efficiently implemented in M3D chips. One promising approach to implement one-shot/few-shot learning is a memory-augmented neural network (MANN), where features extracted from a neural network can be stored and retrieved from an attentional memory, for example, TCAM [6].

In this paper, we demonstrated a M3D chip (M3D-LIME, as illustrated in **Fig. 1**) that is consisted of Si MOSFET-based CMOS logic (1st layer), $HfAlO_x$ RRAM-based CIM (2nd layer) and Ta_2O_5 RRAM-based TCAM (3rd layer). The latter two layers were fabricated using a low-temperature BEOL process without affecting the performance of prior layers, which was carefully verified by structural analysis and electrical measurements. Furthermore, the M3D-LIME chip was used to implement one-shot/few-shot learning task on the Omniglot dataset, where features were extracted using a convolution neural network (CNN) implemented in the 2nd layer, and then stored and retrieved using TCAM implemented in the 3rd layer. A high classification accuracy up to 97.8%, which was very close to the value of GPU, was achieved by M3D-LIME.

II. Fabrication of M3D-LIME Chip

The 1st layer of CMOS control logic was designed and fabricated on an 8-inch Si wafer using a standard 130 nm CMOS foundry process. Si transistors for the one-transistor-one-resistor (1T1R) cells of the 2nd layer were also fabricated in this layer. The wafer processing was stopped at M4 with exposed W top vias after chemical mechanical polishing (CMP) for the fabrication of subsequent layers.

The 2nd layer of $HfAlO_x$-based RRAM crossbar array was fabricated to implement CIM. Firstly, 30 nm TiN was sputtered as the bottom electrode (BE), followed by atomic layer deposition (ALD) of 8 nm $HfAlO_x$ (Hf:Al=3:2) at 300 °C as the resistive switching layer. Then, 45 nm TaO_x and 30 nm TiN were sputtered as the thermal enhanced layer (TEL) [7] and the top electrode (TE), respectively. Here the Al doping and TEL were used to improve the RRAM analog switching property and reliability. After depositions, the RRAM stack was patterned by lithography and reactive ion etching (RIE). Next, 400 nm SiO_2 was deposited as the passivation layer using plasma-enhanced chemical vapor deposition (PECVD) at 300 °C and then contact

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720534

holes were opened by lithography and RIE. After that, W vias were deposited using electroplating and the surface was planarized using CMP. 400 nm Al was sputtered and patterned using lithography and RIE as metal interconnects. Finally, 100 nm SiO_2 and 900 nm Si_3N_4 were deposited using PECVD at 300 °C, followed by contact opening using lithography and RIE and then the wafer surface was planarized again using CMP.

The 3rd layer of CNTFETs and Ta_2O_5-based RRAM was fabricated to construct 2T2R unit cells for TCAM. First, 20 nm Pd was deposited as the local back gate of CNTFETs using evaporation. Secondly, 10 nm Al_2O_3 and 5 nm HfO_2 were deposited by ALD at 220 °C as the gate dielectric, and contact vias were opened by buffered oxide etch. High-density CNTs were then transferred on top using high-purity CNT solution. Next, 80 nm Pd was deposited as the source and drain contacts of CNTFETs. After that, CNTs were patterned and etched using O_2 plasma to define the transistor channels. Next, 45 nm Al_2O_3 was deposited as the passivation layer by ALD at 150 °C and contact vias were opened by RIE to define the BE of RRAM. Then, 20 nm TaO_x (oxygen-deficient) and 10 nm Ta_2O_5 were deposited by sputtering as the oxygen reservoir and the resistive switching layer, respectively. Finally, 130 nm Pt was deposited as the TE of RRAM, and the TaO_x/Ta_2O_5 RRAM stack was etched by RIE using Pt TE as the hard mask, followed by interconnect formation. Transmission electron microscopy (TEM) images of the M3D-LIME chip are shown in **Fig. 2**, and optical images during the fabrication are shown in **Fig. 3**.

III. Device Characterizations of M3D-LIME

Fig. 4 shows DC I-V sweeps of a 1T1R cell in the 2nd layer before and after the fabrication of the 3rd layer, showing no degradation in the electrical performance. This result confirms that the low-temperature fabrication of the 3rd layer caused no damage on the previously fabricated 1st and 2nd layers.

Multi-level programming capability and analog switching characteristics of RRAM are critical to implement CIM on the 1T1R array. **Fig. 5** shows the RRAM conductance in the 2nd layer of 1T1R array as a function of set and reset pulses, exhibiting good analog resistive switching characteristics. **Fig. 6** shows the cumulative probability distribution of 128 cells in the 1T1R array with 32 equally distributed conductance states, equivalent to 5 bits. **Fig. 7** shows the retention test of $HfAlO_x$-based RRAM device mapped to different conductance states as a function of baking time at 125°C, exhibiting a multi-state long retention over 10^4 s. These results demonstrate excellent analog switching characteristics and reliability of 1T1R array for CIM.

Fig. 8 plots the I_D-V_{GS} transfer curves of 13 measured CNTFETs in the 3rd (TCAM) layer, showing a typical on/off ratio $I_{on}/I_{off} > 10^4$. **Fig. 9** plots the I_D-V_D characteristics of a typical CNTFET with a large on-state current density $I_{on}/W >$ 20 μA/μm. **Fig. 10** shows the histograms of $\log_{10}(I_{on}/I_{off})$ and I_{ds}/W of 103 measured CNTFETs. The good uniformity of CNTFETs suggests the feasibility of large-scale integration.

Besides CNTFETs, one of the most important metrics of RRAM for TCAM is the ratio of high/low resistance states (HRS/LRS), which determines the length of search lines [8]. We measured the HRS/LRS ratio of a typical 1T1R half-cell of TCAM with multiple set/reset cycles. The result is shown in **Fig. 11**, where a large HRS/LRS ratio of 560× was observed. Furthermore, 10 more 1T1R half-cells were measured for 10 set/reset cycles each to obtain the statistics of HRS/LRS ratio as shown in **Fig. 12**, where a large HRS/LRS ratio > 520× was observed. **Figs. 13** and **14** further demonstrated the excellent retention (>10^4 s at 125 ºC) and endurance (>5×10^5 cycles) of the TaO_x/Ta_2O_5-based RRAM in the TCAM layer, which are consistent with our previous studies [9].

Fig. 15 shows the search schematic of the 2T2R-based TCAM. The discharging time of match line decides the search result ("0", "1", or "don't care"). **Fig. 16** shows that repeatedly search of one TCAM cell (match or mismatch) had little influence on the stored data. These results demonstrate the proper function and excellent reliability of the TCAM layer.

IV. Implenmentation of One-Shot Learning

To explore the system functionality of M3D-LIME chip, one-shot learning was implemented. Here the data were input through the 1st CMOS layer and then transferred to the 2nd layer of CIM, where a CNN with 2 convolution layers and 2 fully connected layers was realized by the RRAM array to extract features from the input data. Next, the extracted features were quantized into binary signature vectors and transferred to the 3rd layer of TCAM, where the templates of classes were stored. The Hamming distance between the input binary signature vectors and the templates was calculated in parallel using search operation in TCAM. The schematic of one-shot/few-shot learning is shown in **Fig. 17**, and we used it to learn on the Omniglot dataset. The results in **Fig. 18** indicates that M3D-LIME achieved similar classification accuracy as GPU, where 91.8% and 97.8% can be achieved in 5-way 1-shot and 5-shot learning, respectively. Finally, in the performance benchmark, our M3D-LIME shows 162× lower energy consumption than GPU and also 1.68× faster than 2D baseline (**Fig. 19**).

V. Conclusions

In sum, we have designed and fabricated a novel monolithic 3D integration (M3D-LIME) of Si MOSFETs-based logic, RRAM-based CIM and TCAM. The materials and devices in each layer were carefully engineered to meet the requirements of process compatibility. All the devices and cells were characterized to verify that each functional layer worked as designed. Finally, one-shot/few-shot learning was successfully implemented on the M3D-LIME and GPU-equivalent classification accuracy up to 97.8% was achieved.

Acknowledgment

This work was in part supported by National Key R&D Program of China (2019YFB2205104), Beijing Municipal Science and Technology Project (Z191100007519008) and Natural Science Foundation of China (91964104, 61974081).

References

[1] M. Shulaker, et al., Nature, 547, 74, 2017. [2] W. Hwang, et al., ISCAS, 2018. [3] L.-M. Peng, et al., Nat. Electron., 2, 499, 2019. [4] S.-J. Han, et al, Nat. Nanotechnol., 12, 861, 2017. [5] J. Tang, et al., Nat. Electron., 1, 191, 2018. [6] K. Ni, et al., Nat. Electron., 2, 521, 2019. [7] W. Wu, et al., EDL, 38, 1019, 2017. [8] R. Yang, et al., Nat. Electron., 2, 108, 2019. [9] H. Wu, et al., EDL, 35, 39, 2014.

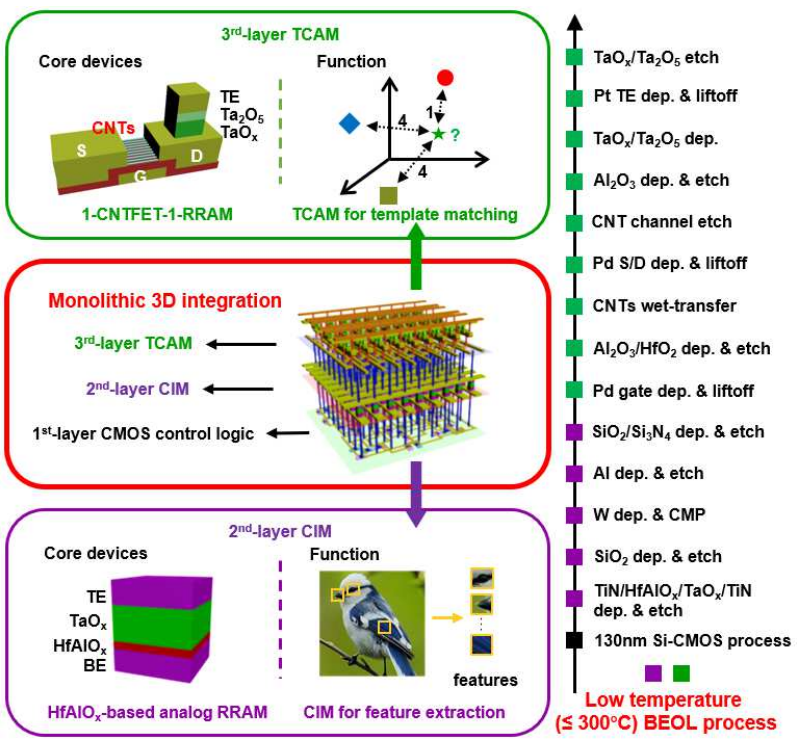

Fig.1. The architecture of monolithic 3D integrated chip (M3D-LIME) and the corresponding fabrication process flow. The chip consists of 3 layers. The 1st Si CMOS layer is fabricated using a standard 130nm process and acts as control logic. The 2nd layer consists of HfAlOx RRAM-based 1T1R array for computing-in-memory (CIM) fabricated using a low-temperature BEOL process. The 3rd layer consists of 2T2R array with CNT FETs and Ta2O5-based RRAMs for ternary content-addressable memory (TCAM).

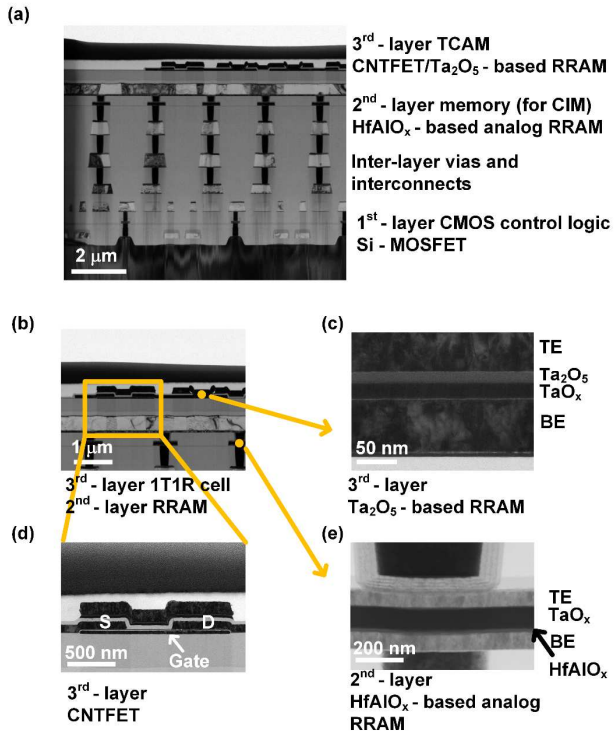

Fig.2. (a) Cross-sectional transmission electron microscopy (TEM) image of the M3D-LIME demonstrated in this work. (b) Zoom-in TEM image of the 2nd layer of CIM and the 3rd layer of TCAM. (c) TEM image of the Ta_2O_5-based RRAM in the TCAM layer. (d) TEM image of the CNTFET in the TCAM layer. (e) TEM image of the $HfAlO_x$-based analog RRAM in the CIM layer.

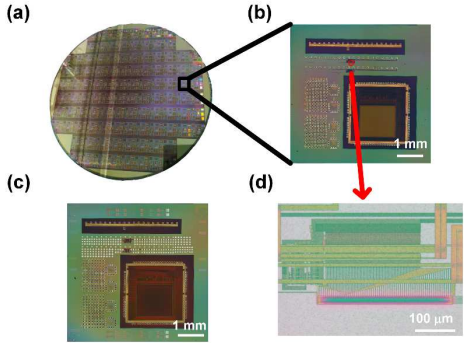

Fig.3. (a) Photo of a 8-inch wafer after the fabrication of the 2nd layer. (b) Optical image of a die on the wafer. (c) Optical image of the same die after the fabrication of all the three layers. (d) Zoom-in view of the 1T1R array in the 2nd layer.

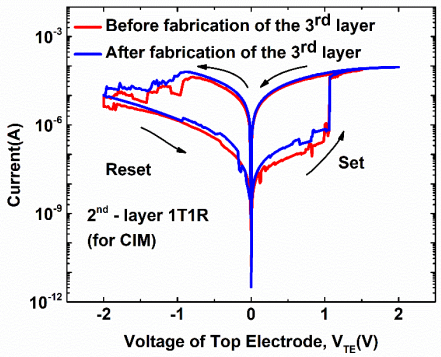

Fig.4. DC I-V sweeps of one 1T1R cell in the 2nd (CIM) layer before and after the fabrication of the 3rd (TCAM) layer, showing no degradation on the RRAM cell performance after the BEOL fabrication process.

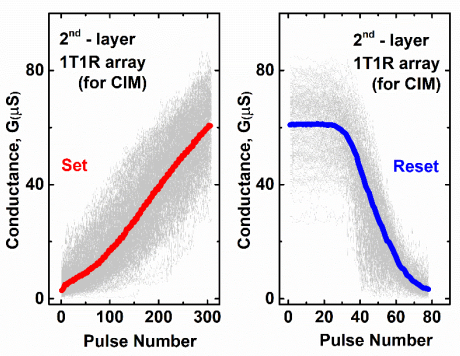

Fig.5. Analog resistive switching characteristics of the 1T1R cell in the 2nd (CIM) layer. 20 devices were measured under a series of set and reset pulses. Gray lines are the raw data. The red and blue lines are the average conductance of 20 devices.

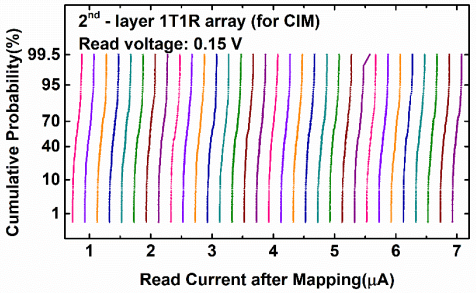

Fig.6. Cumulative probability distribution of 128 cells in the 1T1R array of the 2nd (CIM) layer with 32 equally distributed conductance states, showing the programming capability of 5 bits per cell.

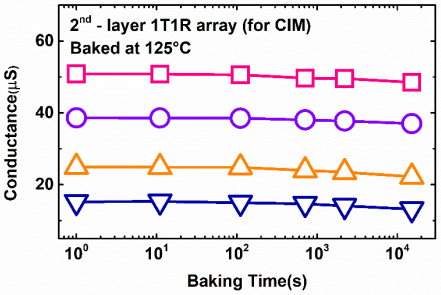

Fig.7. Retention test of 4 representative conductance states of the 2nd (CIM) layer as a function of baking time at 125℃.

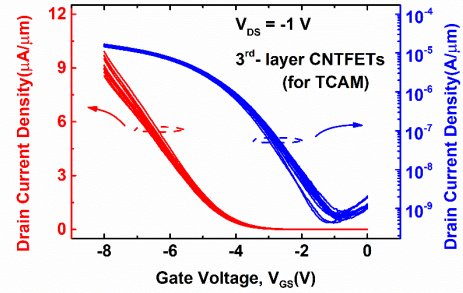

Fig.8. I_D-V_{GS} transfer curves of 13 measured CNTFETs in the 3rd (TCAM) layer. The channel length is 2 μm, and the gate dielectric consists of 10 nm Al_2O_3 and 5 nm HfO_2.

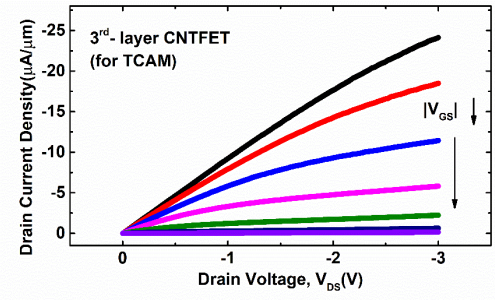

Fig.9. I_{DS}-V_D characteristics of a typical CNTFET in the 3rd (TCAM) layer measured at different gate voltages (V_{GS}) from -8V (top) to -2V (bottom) with steps of 1V.

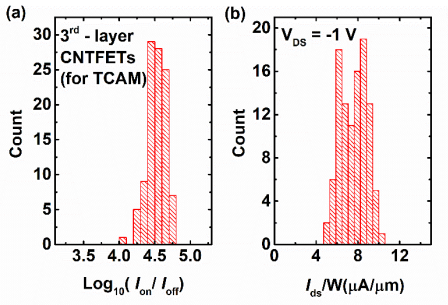

Fig.10. (a) $\mathrm{Log}_{10}(I_{on}/I_{off})$ histogram of 103 CNT FETs measured in the TCAM layer. (b) The corresponding histogram of the current density I_{ds}/W at V_{DS} = -1 V.

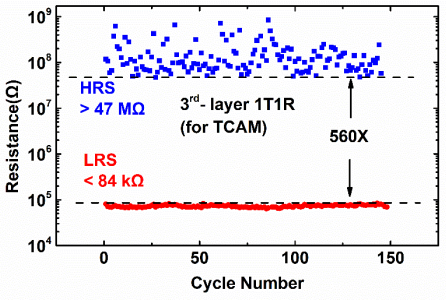

Fig.11. Resistance of a 1T1R cell in the 3rd (TCAM) layer as a function of pulse cycle number. The cell was programmed with 1 μs set and reset pulses (with verify). A large HRS/LRS ratio of 560× was observed.

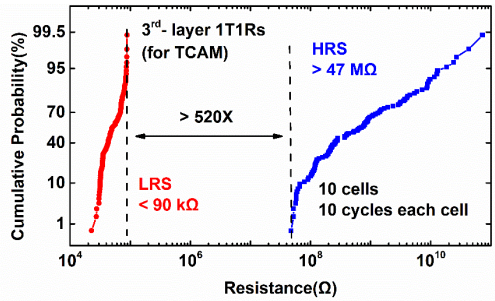

Fig.12. Resistance distribution of 10 1T1R cells in the 3rd (TCAM) layer programmed with 10 set/reset cycles each, showing a large HRS/LRS ratio >520×.

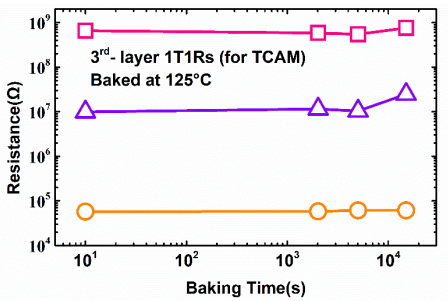

Fig.13. Retention test of three resistance states of 1T1R cells in the 3rd (TCAM) layer as a function of baking time at 125°C.

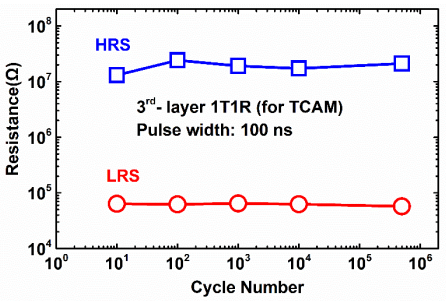

Fig.14. Endurance test of a 1T1R cell in the 3rd (TCAM) layer. The set/reset pulse voltages are 4V and 5V, respectively.

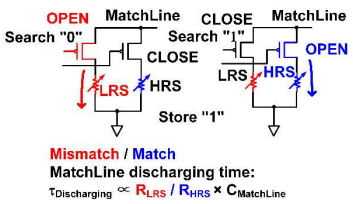

Fig.15. Searching schematic of a 2T2R TCAM cell. By programming the 2 RRAMs to different resistance states, the cell stores "0", "1", and "don't care". The stored data can be searched through the search lines.

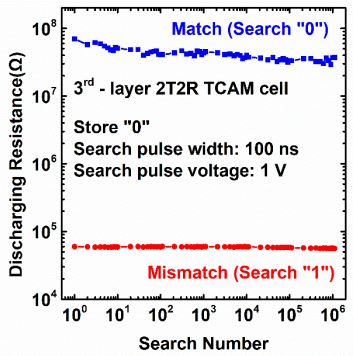

Fig.16. Discharging resistance of a 2T2R TCAM cell depending on the search data. The cell stores "0", so it is a mismatch when searching "1" while a match when searching "0".

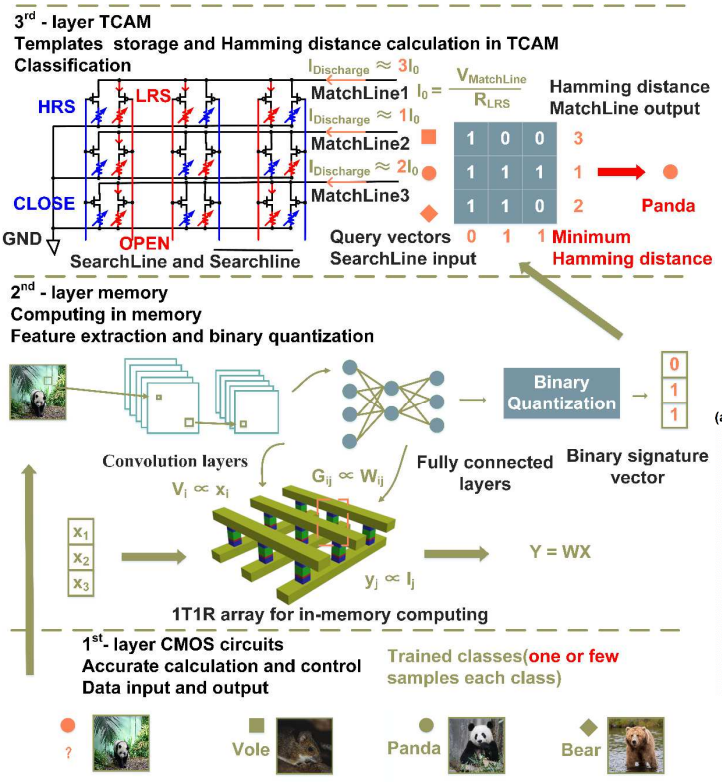

Fig.17. Schematic of implementing one-shot/few-shot learning using the M3D-LIME demonstrated in this work. The 1st layer of CMOS circuits inputs the data, outputs results, and also controls the learning process. The 2nd layer of RRAM array extracts features from the input data by taking the advantage of CIM. The 3rd layer of TCAM stores templates and calculates the Hamming distance.

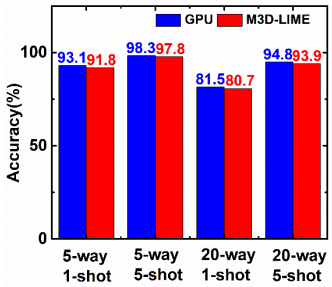

Fig.18. The classification accuracy of one-shot/few-shot learning on the Omniglot dataset using GPU and the M3D-LIME. The accuracy is the average of randomly selected N classes (N-way) in the dataset.

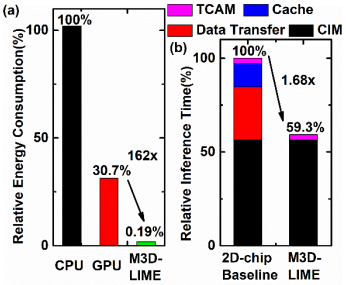

Fig.19. (a) Energy consumption benchmark on the M3D-LIME, CPU and GPU. (b) Inference time benchmark on the M3D-LIME and 2D-chip baseline (in which CIM and TCAM are not monolithically integrated). 162× energy reduction compared to GPU and 1.68× speedup compared to 2D baseline were achieved by the M3D-LIME.