Emerging Non-Volatile Memories for Computation-in-Memory

Bin Gao

Institute of Microelectronics, Beijing Innovation Center for Future Chips (ICFC),
Tsinghua University, Beijing, 100084, China.
E-mail: gaob1@tsinghua.edu.cn

Abstract - This paper first introduces the principles of different emerging non-volatile memory (NVM) devices. The device structures, working mechanisms, as well as typical performance of these devices are discussed. The technologies for enhancing data storage density, such as three-dimension integration, multi-level cell, are also discussed. Then different approaches of computation-in-memory (CIM) based on emerging NVM will be presented, specially focus on vector-matrix-multiplication. Later, the paper will summary the performance requirements and key challenges on the device level to realize the CIM. Finally, this paper will provide some possible research directions in the future development on emerging NVM for CIM applications.

I. Introduction

Over years, the memory market is dominated by NAND Flash and DRAM. As the demand of data storage increases, these mainstream memories face challenges such as device scaling issue and low programming speed. In this case, many emerging NVM devices are proposed, including resistive random access memory (RRAM), phase change memory (PCM), spin-transfer-torque magnetic random access memory (STT-MRAM), etc. The emerging NVMs exhibit better potential in high-density integration, and also show fast speed, low operation voltage, and long retention time [1]. More important, the emerging NVM devices make the idea of CIM possible. The CIM concept provides a new solution for processing data-intensive tasks with very high efficiency.

II. Emerging NVM Devices

A. Two Terminal Device

Most of the emerging NVM devices are two-terminal cells, which have the most compact structure, as show in Fig. 1a-f [2]. These devices all have switching dielectric layers between two electrodes. By applied a relative large voltage on the two electrodes, the resistance of the switching dielectric layers can be changed according to the voltage amplitude and polarity. The resistance states are used to storage information. During readout process, a relative small voltage is applied, which cannot influence the resistance state.

The resistive switching of PCM devices is based on amorphous to crystalline transition (Fig. 1a) or Mott transition (Fig. 1b). While the switching in RRAM is caused by ion migration. The ion can be either anion (oxygen ions or oxygen vacancy, so also called OxRAM, Fig. 1c) or cation (metal ions, also called CBRAM, Fig. 1d) [3]. Under different direction of electric field, the ions can migrate to form or de-form a conductive filament. Sometime, RRAM is also called "memristor". Some tunneling junction-based devices, like ferroelectric tunneling junction (FTJ, Fig. 1e) and magnetic tunneling junction (MTJ, Fig. 1f) also show resistive switching behaviors. In this device, the direction of ferroelectric polarization or spin polarization is changed under the control of external electric signal, and finally inducing the change of tunneling current through the device. Among these two-terminal devices, PCM and RRAM have demonstrated relative better performance when considering all the key metrics at the same time.

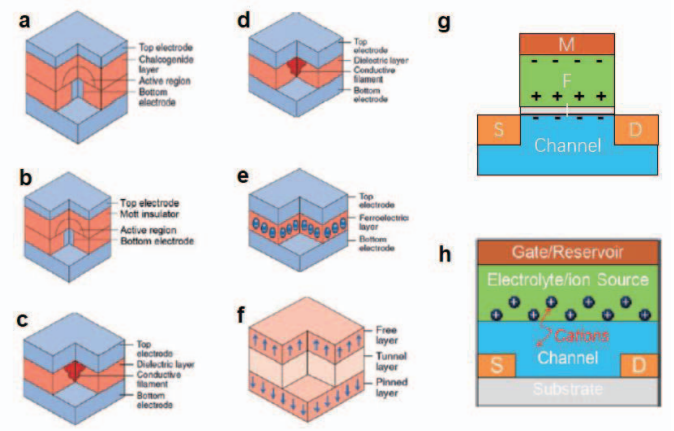

Fig. 1. Different types of emerging NVM devices: (a) PCM, (b) Mott PCM, (c) Oxide RRAM, (d) CBRAM, (e) FTJ, (f) STT-MRAM, (g) FeFET, (h) ECRAM. [2]

B. Three Terminal Device

Comparing with two terminal devices, the three terminal NVM devices separate the control path and readout path. Thus, the resistance tuning ability of the three terminal devices are improved, although the integration density is decreased. Ferro-FET (FeFET) is a type of three terminal NVM. The gate oxide of FeFET is replaced by a ferroelectric layer (Fig. 1g). Besides, some ion migration based transistors were proposed. An electrochemical RAM (ECRAM) is developed by controlling the ion concentration in the channel with a gate voltage (Fig. 1h). FeFET and ECRAM show better analog conductance tuning ability and linearity, but the switching speed and reliability need more optimization in the future.

C. Multilevel and Analog Storage

All the NVM devices have at least two stable states, namely high resistance state (HRS) and low resistance state (LRS). With the two stable states, the device is capable of storing information "0" and "1". To enhance the storage density, a

978-1-7281-4123-7/20/$31.00 ©2020 IEEE

multilevel per cell technique is developed. Different from the binary resistive switching (Fig. 2a), the multilevel resistive switching shows multiple stable states (Fig. 2b). Each state can be obtained by varying the programming conditions.

Recently, the CIM application calls for a new type of switching, called analog resistive switching (Fig. 2c) [4]. In this type of devices (analog NVMs), the resistance can be tune continuously within a window. At the current stage, although many analog NVMs are proposed, this technology is not mature enough for massive production.

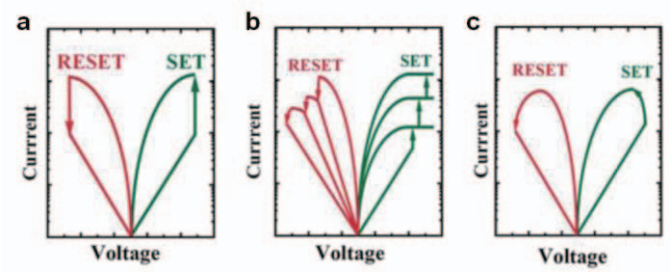

Fig. 2. Different types of resistive switching behaviors: (a) Binary switching, (b) Multilevel switching, (c) Analog switching. [4]

D. Array Integration

The two terminal NVM devices can be integrated into a high-density crossbar structure. The crossbar arrays can be stacked in a 3D form to further improve the density (Fig. 3a). However, this 3D stacked crossbar suffers from high cost issue due to the complex fabrication process. Therefore, vertical 3D NVM arrays, with either line-based (Fig. 3b) or plane-based (Fig. 3c) structure were proposed [5]. With these vertical structures, the number of critical lithography during the fabrication process can be reduced to less than three, and thus the fabrication cost can be reduced significantly.

In the crossbar array, developing a good selector device is a crucial task [6]. Otherwise, the sneak path issue may damage the successful operation in the array. However, up to now, a mature selector device with both excellent reliability and large selective ratio is still lack. Therefore, in most case, the 1T1R array structure is preferred. In a 1T1R array, the two terminal NVM device is located on top of the drain of the transistors. The transistor works as a selector device.

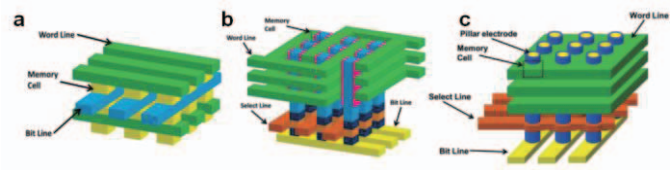

Fig. 3. Different types of 3D NVM arrays: (a) 3D-XPoint structure, (b) Line-based 3D-Vertical structure, (c) Plane-based 3D-Vertical structure. [5]

III. CIM BASED ON NVM

Data-intensive computing tasks become more and more common in the present big data era. However, the current computers are all based on traditional von Neumann architecture, in which the computing unit and memory unit are separated. During data processing, data need to transfer between the computing unit and memory unit, which cost huge energy and latency, especially for the data-intensive computing tasks.

The NVM array can naturally process the vector-matrix-multiplication (VMM), which is a key part for many data-intensive computing tasks, such as deep neural network (DNN), video signal processing, linear and partial differential equations, etc. As shown in Fig. 4, read voltage on each word line represents the input vector, and the conductance of NVM devices in the array represents the weight matrix. The multiplication results can be naturally got via Ohmic law and Kirchhoff law, represented by the readout current on each bit line. Since all the weights are stored at the computing site, there is no need to transfer most of the data during computing. With this CIM approach, the computing speed and energy efficiency of VMM can be reduced by several orders.

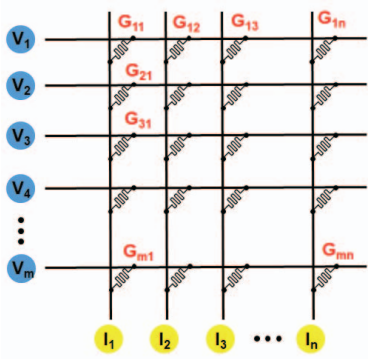

Fig. 4. Schematic of vector-matrix-multiplication with a NVM array.

DNN is the most widely used artificial intelligence (AI) algorithm and data-intensive computing task. A DNN can be compiled into many VMMs, and thus is quite suitable to be processed on NVM arrays [7]. The neural network algorithm includes two parts: inference and training. Some NVM based DNN accelerators only support inference, so they are called inference-only chips. For these inference-only chips, the weight data are pre-programmed into the array, and only VMM is performed on-chip. Sometimes, training capability is required. In this case, the conductance of the NVM should be *in situ* tuned to perform the weight update process. For NVM based DNN accelerators, the training capability is crucial, because the *in situ* training can help to tolerant some variability and reliability degradation issues [8].

The performance requirements of NVM devices for inference-only and *in situ* training applications are different, which will be discussed in the next section.

IV. PERFORMANCE REQUIRMENTS OF NVMS

A. Basic Memory Performance

Basic memory performance metrics include switching speed, operation voltage/energy, variability, retention and endurance, etc. Most of the NVM devices show fast speed. The typical switching time is usually less than 10ns, so speed is never a problem for NVM. Operation voltage is important for advanced semiconductor process. Compared with Flash and DRAM, emerging NVMs have much smaller program

voltage. However, the typical program voltage of NVM, usually around 1~2V, is still not low enough. In the future, reducing program voltage is one of the key task for NVM device, especially for high-density embedded memory application. Read voltage of NVM can be varied between 0V to hundreds of mV, which is an ideal factor for both memory and CIM applications. Program and read energy will influence the energy efficiency of the system. The operation energy can be reduced by limiting the switching current of the NVM device. Currently, many NVM devices can be read with an energy consumption below 0.01pJ, which can result in a >100TOPs/W energy efficiency.

Binary state retention is never a problem for emerging NVM devices. To evaluate the retention ability, the NVM is usually baked under a high temperature, like 200ºC. However, for CIM application, the conductance drift of multiple levels should be taken into account, which brings great challenges for NVM devices [9]. Generally, although some states are quite stable, it is difficult to make all the intermediate resistance level stable. There is always some resistance fluctuation or drift over time. Further research from device level and architecture level is highly required to solve the analog state retention issue.

Endurance requirement is different for different applications. There is no clear criterion to judge how many endurance cycles is enough for training or inference-only application [10]. For NVM devices, STT-MRAM shows the best endurance performance. Other NVM devices usually show 10^5~10^8 endurance cycles.

Variability refers to the distribution of operation voltage and resistance of different states. Variations include cycle-to-cycle variation and device-to-device variation. For multilevel storage and CIM application, the distribution should be kept tight enough, which also need more efforts in the future.

B. Device Performance of Training

Training application brings more requirements for the NVM devices, including dynamic ratio, level number, weight update linearity and symmetry, pulse-to-pulse fluctuation, and I-V linearity [11]. The conductance of an analog NVM can be tuned continuously with identical pulse trains. The ideal device shows linear and smooth conductance tuning behavior (Fig, 5a), while there are always some non-ideal effects in the real device (Fig. 5b), including nonlinearity, fluctuation, etc.

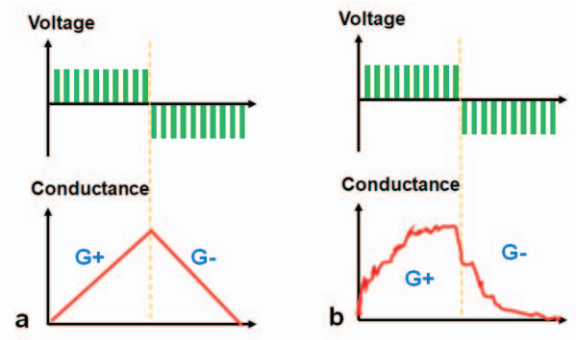

Fig. 5. Schematic of analog conductance tuning behaviors of NVM devices for training application: (a) Ideal behavior, (b) Real behavior with some non-ideal effects.

Dynamic ratio means the ON/OFF conductance ratio in the analog switching region. Some NVM devices have a large full switching ratio great than 1000. However, the dynamic ratio of analog switching region is less than 10. The dynamic ratio reflects the ability to map a floating-number weight to the conductance of a device. A >10 dynamic ratio is preferred for the training application. Level number means the pulse number capacity to program the NVM device from the lowest conductance state to the highest conductance state or from the highest conductance state to the lowest conductance state. If the level number is not enough, the conductance change under one weight update pulse would be too large and the neural network is difficult to converge. Simulation results have shown that for MNIST dataset training with a fully connected neural network, at least 100~200 level number is required [11].

Weight update linearity means whether the conductance change is constant. As shown in Fig. 5b, the conductance usually changes faster at the beginning and then tends to saturation, resulting in a nonlinear weight update behavior. Weight update symmetry refers to whether the weight increase process and weight decrease process are similar. The main asymmetric factors are linearity and level number. Besides weight update linearity/symmetry, pulse-to-pulse fluctuation also affect the training accuracy significantly. Pulse-to-pulse fluctuation is another type of variability. It refers to the deviation of the conductance to the mean value curve during one weight increase/decrease process, as illustrated with the fluctuations in Fig. 5b.

Furthermore, I-V linearity is also important for both training and inference-only applications. During VMM computing, the input vector is expected to encode as pulse amplitude to be applied on the NVM array. However, if the NVM device has large I-V nonlinearity, the readout conductance might be different when the input voltage changes, which means the weight value is not fixed during inference. Therefore, nonlinear I-V will definitely cause the degradation of computing accuracy. To fight against the I-V nonlinearity, sometimes, the input vector is encoded as pulse number instead of pulse amplitude. In this case, the pulse amplitude is fixed, and thus the weight value also keeps constant. However, the pulse number encoding scheme suffers from the significant increase of computing latency and energy. As a result, many device-level researches have focused on improving the I-V linearity. However, due to the non-ohmic conduction mechanisms of NVM devices, improving I-V linearity is also a challenging task.

V. Outlook

Although NVM based CIM has been already regarded as a promising technology for the future high-performance computing, the current research progress is still limited to single device level or simple small-scale array/macro level. There are three key challenges for the realization of a large-scale NVM based CIM chips. The first is the requirement of improving analog switching performance, as discussed in the previous section. The device optimization is a persistent work and may need tens of years. However, to develop a full CIM chip and push it to the practical application

as soon as possible, at least, the stability of multilevel state should be improved. The second challenge is the fabrication process in foundry. Most of the NVM devices needs new materials or new process beyond the conventional CMOS fabrication process. Therefore, the foundries have to spend more time to develop new process for the integration of NVM and try to get a good yield. The third challenge, which is the most important, is the lack of multi-scale modeling techniques for NVM devices and systems.

Different from the conventional memory system, the CIM system highly requires cross-layer co-design. Therefore, the design tool chain should contain at least device level, architecture level and algorithm level. On the device level, due to the complex resistive switching mechanism, there is still lack of a wide-accepted device model for NVMs. Physical based numerical models (Fig. 6a) can well capture some of the resistive switching behaviors, but requires very long simulation time [12]. Compact models (Fig. 6b) and SPICE models (Fig. 6c) of NVM device can be used for circuit simulation or architecture analysis, but the model precision cannot meet the industry standard at the present stage. On the system level, an end-to-end simulator for CIM is demanded (Fig. 6d). With this simulator, the computing accuracy and chip performance of different designs can be evaluated [7].

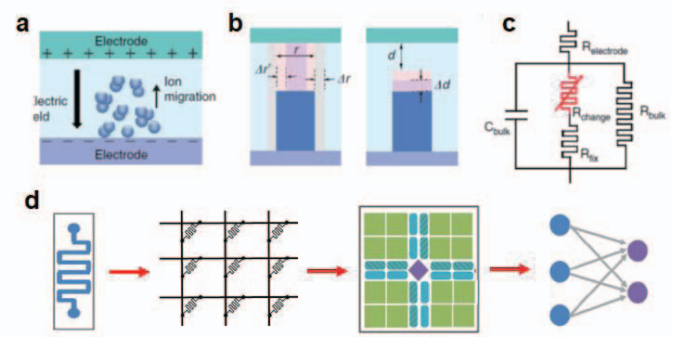

Fig. 6. Schematic of multi-scale modeling and simulation of NVM devices and NVM based CIM system: (a) Physical-based numerical model, (b) Physical-based compact model, (c) SPICE model, (d) End-to-end simulator framework for a CIM system.

Once the above-mentioned challenges have been solved, the NVM based CIM technique will become mature and the large-scale CIM chips will be developed to process many practical computing tasks. Furthermore, the brain-inspired neuromorphic computing will make the CIM chips more powerful by exploiting some neuro-like behaviors on the NVM devices. In the long-term future, the 3D integration technology will make the TB-level on-chip integration density possible, and the CIM chips might run as faster as today's supercomputer while only consume ~W level power.

VI. Conclusion

Emerging NVM devices have exhibited great potential for the future CIM application. This paper summarizes the key technology for the NVM devices and arrays. The challenges in the future research are also discussed.

Acknowledgments

This paper is supported in part by the National Key R&D Program of China (2016YFA0201801), the National Natural Science Foundation of China (61874169), Beijing Municipal Science and Technology Project (Z191100007519008), and Beijing National Research Center for Information Science and Technology (BNRist).

References

[1] H. Wu, et al, "Resistive Random Access Memory for Future Information Processing," *Proceedings of the IEEE*, Vol. 105, pp. 1770-1789, 2017.

[2] J. Tang, et al, "Bridging Biological and Artificial Neural Networks with Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges," *Advanced Materials*, p.1902761, 2019.

[3] B. Gao, et al, "Understanding memristive switching via in situ characterization and device modeling," *Nature Communications*, Vol. 10, p.3453, 2019.

[4] W. Zhang, et al, "Analog-Type Resistive Switching Devices for Neuromorphic Computing," *Physics Status Solidi Rapid*, p.1900204, 2019.

[5] Y. Deng, et al, "Design and Optimization Methodology for 3D RRAM Arrays," *IEEE International Electron Device Meeting* (*IEDM*), pp. 629-632, 2014.

[6] Q. Hua, et al, "A Threshold Switching Selector Based on Highly Ordered Ag Nanodots for X-Point Memory Applications", *Advanced Science*, Vol. 6, p.1900024, 2019.

[7] W. Zhang, et al, "Design Guidelines of RRAM based Neural-Processing-Unit: A Joint Device-Circuit-Algorithm Analysis," *Design Automation Conference* (*DAC*), 2019.

[8] P. Yao, et al, "Face classification using electronic synapses," *Nature Communications*, Vol. 8, p.15199, 2017.

[9] M. Zhao, et al, "Investigation of Statistical Retention of Filamentary Analog RRAM for Neuromophic Computing," *IEEE International Electron Device Meeting* (*IEDM*), pp. 872-875, 2017.

[10] M. Zhao, et al, "Characterizing Endurance Degradation of Incremental Switching in Analog RRAM for Neuromorphic Systems," *IEEE International Electron Device Meeting* (*IEDM*), pp. 468-451, 2018.

[11] H. Wu, et al, "Device and circuit optimization of RRAM for Neuromorphic computing," *IEEE International Electron Device Meeting* (*IEDM*), pp. 274-277, 2017.

[12] B. Gao, et al, "Modeling Disorder Effect of the Oxygen Vacancy Distribution in Filamentary Analog RRAM for Neuromorphic Computing," *IEEE International Electron Device Meeting* (*IEDM*), pp. 91-94, 2017.