Challenges of Emerging Memory and Memristor Based Circuits: Nonvolatile Logics, IoT Security, Deep Learning and Neuromorphic Computing

Chunmeng Dou[1], Wei-Hao Chen[1], Yi-Ju Chen[1], Huan-Ting Lin[1], Wei-Yu Lin[1], Mon-Shu Ho[2], Meng-Fan Chang[1]*

[1] Electrical Engineering Dept., National Tsing Hua University, Hsinchu, Taiwan
[2] Physics Dept., National Chung Hsing University, Taichung, Taiwan
* Email:mfchang@ee.nthu.edu.tw

Abstract

Emerging nonvolatile memory (NVM) devices are not limited to build nonvolatile memory macros. They can also be used in developing nonvolatile logics (nvLogics) for nonvolatile processors, security circuits for the internet of things (IoT), and computing-in-memory (CIM) for artificial intelligence (AI) chips. This paper explores the challenges in circuit designs of emerging memory devices for application in nonvolatile logics, security circuits, and CIM for deep neural networks (DNN). Several silicon-verified examples of these circuits are reviewed in this paper.

1. Introduction

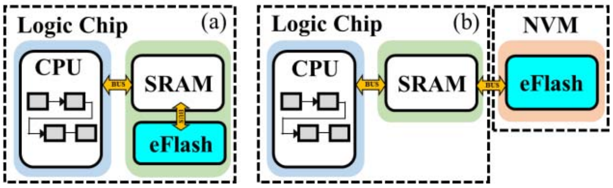

Fig.1 Conventional processor structures involving CPU, SRAM, and (a) on or (b) off-chip NVM macro

Figure 1 illustrates typical processor structures using on- or off-chip NVM macro [1-5]. Considerable time and energy are consumed in the serial movement of data between CPU, SRAM, and NVM and writing data to NVM during power-off operation. Emerging NVM devices can act as key enablers in the development of innovative circuits and advanced computing architectures aimed at improving the performance and energy-efficiency of such systems.

In the following sections, we briefly review the current status of emerging NVM devices and explore the promises and challenges in the further development of emerging NVM-based circuits for nvLogics and nonvolatile processors, security circuitry, and CIM for DNN computing and AI chips.

2. Recent Emerging Memory

Figure 2 illustrates recent emerging NVM technologies, including resistive RAM (ReRAM, RRAM, Memristor) [6]-[17], phase-change memory (PCM) [18]-[23], and spin-transfer-torque magnetic RAM (STT-MRAM) [24]-[28], all of which incur write voltage and write time lower than that of Flash memory. These benefits make emerging NVM devices particularly attractive for energy-efficiency systems.

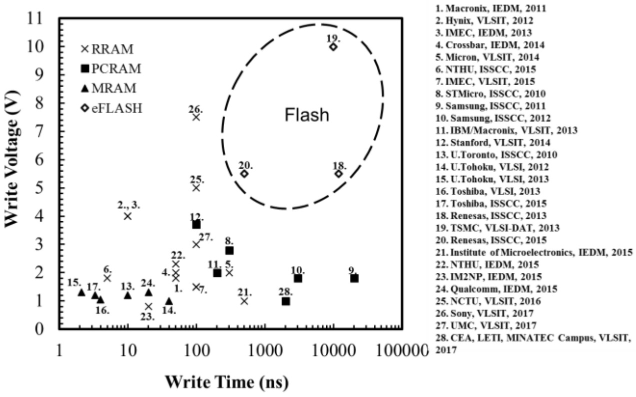

Fig. 2 Write voltages and write times of emerging NVM devices and Flash memory (2010 – 2017)

Figure 3 presents the R-ratio (i.e., the resistance-ratio between high and low resistive states), of emerging NVMs. There has not been a notable increase in R-ratio in recent years, due mainly to the need to reduce the operational power. This has necessitated the development of memory peripheral circuits with tolerance for the small R-ratio.

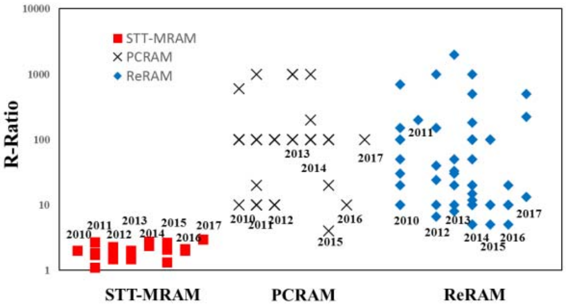

Fig. 3 Comparison of R-ratio values in recent NVMs (2010 – 2017)

978-1-5090-6625-4/17/$31.00 ©2017 IEEE

3. nvLogics and Nonvolatile Processors

3.1 nvLogics: Concept and Application

nvLogics rely on emerging memory-based devices, such as nonvolatile SRAM (nvSRAM) [30-36], nonvolatile flip-flops (nvFF) [37-41], and nonvolatile TCAM (nvTCAM) [42-46]. Figure 4 compares system-on-chip (SoC) processors based on embedded NVM and nvLogics. The TCAM circuits can be used as filters for the removal of redundant data transmission in the IoT application [42-46]. In a conventional system, all of the critical data must be serially moved from the logic circuits (e.g., flip-flops, SRAM, and TCAM) to NVM (eFlash) for power-off, or reversely for power-on.

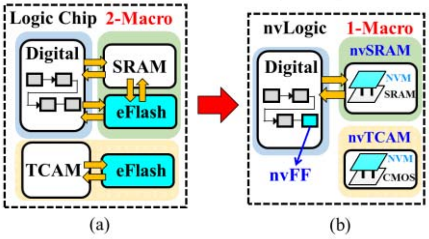

Fig. 4 Energy-efficient systems based on (a) embedded NVM and (b) nvLogics

In contrast, the nvLogics based system is able to store or restore its critical data with the assistance of local NVM devices during power switching. The latency and energy required by the system to power-down and wake-up can be effectively reduced because (1) the data are moved in a parallel and distributed manner, and (2) emerging NVMs have low operational power and fast access time. Hence nvLogics can enable energy-efficient systems under frequent power-off conditions (Fig. 5).

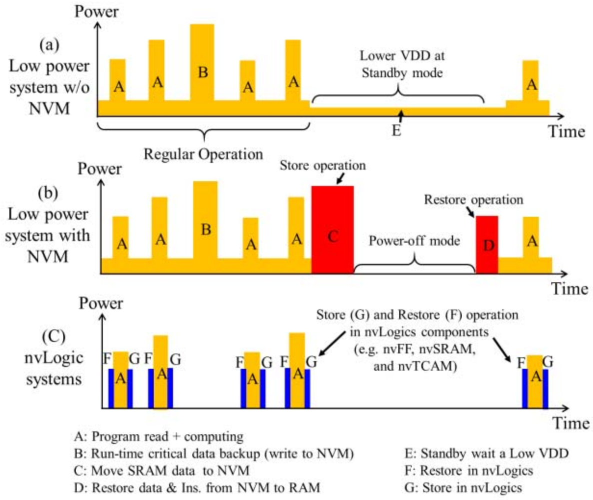

Fig.5 Schematic illustration showing power vs. time relationship of various power-saving schemes based on (a) low stand-by voltage, (b) embedded NVM macros, and (c) nvLogics.

3.2 Examples of nvLogics (nvSRAM and nvTCAM)

nvLogics components can be fabricated by integrating emerging NVM devices with CMOS devices by back-end-of-line (BEOL) processes. Figure 6 presents examples of recent silicon-verified nonvolatile nvSRAM cells [30-36]. nvSRAM can be operated in the normal SRAM mode as well as the store/restore mode by utilizing the NVM cells at storage nodes (Q/QB).

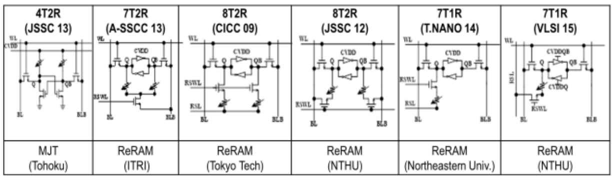

Fig. 6 Recent silicon-verified nvSRAM cells

Figure 7 presents several silicon-verified nvTCAMs [42-46]. The signature pattern can be pre-stored by appropriately biasing local NVM devices during write operations. The search signal is input by search-data lines (DL and DLB) and the output is read from the machine line (ML) during search operations.

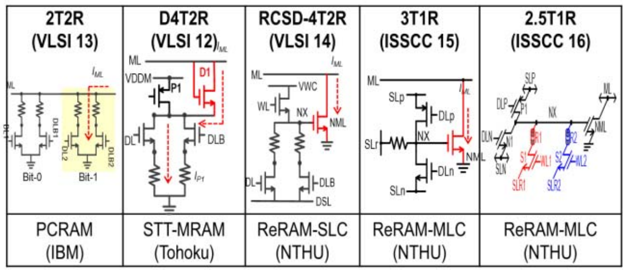

Fig. 7. Recent silicon-verified nvTCAM cells

3.3. Challenges in Circuit Design nvLogics

In nvSRAM and nvFF, the small R-ratio of emerging NVM cells results in only small differences between the signals at Q and QB nodes during restore operations, which leaves the system susceptible to restore errors. Furthermore, the common practice of using the maximum write time for store operations (to ensure the proper operation of even the slowest cells) can further degrade the reliability of the system [29,41]. In nvTCAM, the small R-ratio of NVM cells results in small ML current-ratio between ML mismatch current and ML leakage current, thereby limiting the word length of the TCAM. ML leakage current and ML parasitic load are also major concerns.

4. Emerging memory-based Security Circuits

4.1 Concepts of emerging memory in security circuits

Figure 8 illustrates the concept of physical unclonable function (PUF)-based network security schemes and a brief authentication flow. The security level of PUF depends on the number of unique challenge-response pairs (CPRs) that could be generated. Conventional PUFs exploit random variations in the CMOS process. Emerging NVMs have recently been shown to enhance the randomness of PUFs due to intrinsic variations in their resistive switching processes [47-49].

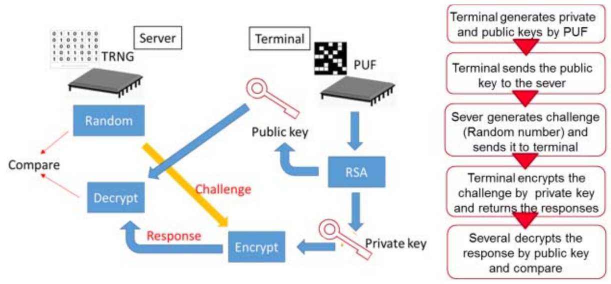
Server
TRNG
Random
Compare
Decrypt
Challenge
Response
Encrypt
Terminal
PUF
Public key
RSA
Private key
Terminal generates private and public keys by PUF
Terminal sends the public key to the sever
Sever generates challenge (Random number) and sends it to terminal
Terminal encrypts the challenge by private key and returns the responses
Several decrypts the response by public key and compare

Fig.8 PUF-based network security scheme and brief authentication flow

4.1 Challenges for emerging memory-based PUF

Because of the high cell resistances and the minor differences in the resistance, the sensing circuits for emerging NVM-based PUFs suffer from small sensing margin and long access time. Therefore, the development of high-resolution sense amplifiers (SAs) with small offset remains a challenge. Finally, the reliability of NVM PUFs depends on keeping the outputs stable after read-out stress or under fluctuating environmental conditions, such as variations in the temperature or voltage supply.

5. Emerging Memory-based CIM for DNN and AI Chips

5.1 Concepts of emerging memory-based CIM

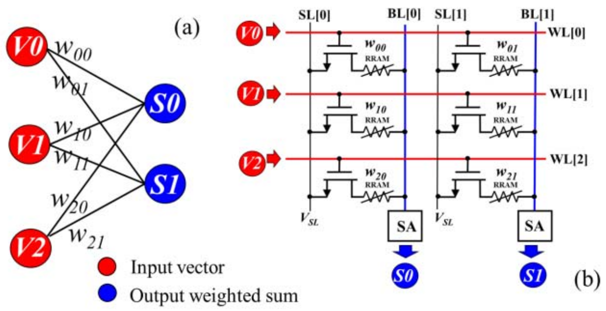

Fig. 9 (a) Simplified NN model and (b) the RRAM array structure for CIM

Many recent studies have employed emerging NVM (memristors)-based circuits for DNN and AI Chips [50-60]. Figure 9 illustrates a simplified model of neural network (NN) computing and its circuit implantation based on RRAM array structure for CIM. Weight information can be stored and trained in the memory array by writing the cell. The matrix-vector multiplication (MVM) operation can then be executed directly within the memory array, as the output signals of the array are weighted sums of the input signals. This approach to CIM accelerates NN computing without requiring the serial movement of data between NVM and digital circuits. Combining multiple memory arrays also makes this approach directly applicable to DNN. The memory array structure for CIM can be realized using either cross-bar or 1T1R configuration [60-63].

5.2 Challenges for Emerging Memory-based CIM

NVM-based CIM introduces challenges beyond the typical problems of NVM outlined above: (1) high performance NVM devices with well-controlled multiple resistive states and low-power operations are still expected; (2) There is a possibility of large sneak current during MVM operations; and (3) The analog nature of CIM processes necessitates the optimization of NVM/CPU interfaces to boost system performance [64-66].

Summary

Due to their low operating voltages, fast access speeds, and compatibility with CMOS devices, emerging NVM devices have enabled many innovative circuits for advanced system architectures and computing model. These circuits have opened the door to further improvements in the energy-efficiency and computing capacity of existing systems used in low-power computing, IoT security, neuromorphic computing and AI. Nonetheless, the design of these circuits imposes a number of challenges associated with the characteristics of NVM devices, such as their small R-ratio and considerable variations on cell behaviors. Novel circuit designs are expected to further achieve high yields, reduce area overhead, and suppress power consumption.

Acknowledgments

The authors would like to thank the support from ITRI, TSMC, MXIC, CIC and MOST of Taiwan.

7. References

[1] Y. Yano, *et al.*, *ISSCC,* pp.24-30 (2012)

[2] M. Hatanaka, *et al.*, *IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 38–42 (2007)

[3] H. Hidaka, *IEEE International Conference on IC Design & Technology (ICICDT)*, pp. 1–4 (2011)

[4] M. Zwerg, *et al.*, *ISSCC*, pp. 334-336 (2011)

[5] Y. Wang, *et al.*, *Proceedings of the European Solid-State Circuits Conference (ESSCIRC),* pp. 149-152 (2012)

[6] M.-F. Chang, *et al.*, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, pp. 183-193 (2015)

[7] M.-F. Chang, *et al.*, *IEEE Asia South Pacific Design Automat. Conf.*, pp. 569–574 (2015)

[8] M.-F. Chang, *et al.*, *IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT),*, pp. 1–4 (2014)

[9] M.-F. Chang, *et al.*, *IEEE Asia South Pacific Design Automat. Conf.*, pp. 329–334 (2012)

[10] M.-F. Chang, *et al.*, *IEEE Int. Conf. ASIC*, pp. 299–302 (2011)

[11] M.-F. Chang, *et al.*, *IEEE Asia South Pacific Design Automat. Conf.*, pp. 197–203 (2011)

[12] S. Kim, *et al.*, *Symp. VLSI Techonlogy*, pp. 155–156 (2012)

[13] K. Eshraghian, *et al.*, *IEEE Trans. Very Large Scale Syst.*, vol. 19, no. 8, pp. 1407-1417 (2010)

[14] Q. Luo, *et al., IEDM,* pp. 10.2.1-4 (2015)

[15] H.-W. Pan, *et al., IEDM,* pp. 10.5.1-4 (2015)

[16] G. Piccolboni, *et al.*, *IEDM,* pp. 17.2.1-4 (2015)

[17] X. Xu, *et al.*, *Symp. VLSI Technology*, pp.1-2 (2016)

[18] W-S. Khwa, *et al.*, *IEEE Electron Device Letters*, no.37, pp. 1422-1425 (2016)

[19] W.-S. Khwa, *et al.*, *IEDM*, pp. 29.8.1-4 (2014)

[20] C.-Y. Wen, *et al.*, *Symp. VLSI Circuit*, pp. 302-303 (2011)

[21] M. Rizzi, *et al.*, *IEDM*, pp. 29.6.1-4 (2014)

[22] H. Pozidis, *et al.*, *IEEE Int. Memory Workshop*, pp. 100-103 (2013)

[23] M. Boniardi, *et al., IEDM*, pp. 29.1.1-4 (2014)

[24] W. J. Kim, *et al.*, *IEDM*, pp. 24.1.1-4 (2011)

[25] E. Kitagawa, *et al.*, *IEDM* pp. 29.4.1-4 (2012)

[26] Y.-H. Wang, *et al.*, *IEDM*, pp. 29.2.1-4 (2012)

[27] K. Tsunoda, *et al.*, *IEDM*, pp. 3.3.1-4 (2013)

[28] Y. Lu, *et al.*, *IEDM,* pp. 26.1.1-4 (2015)

[29] C.-P. Lo, W.-H. Chen, …., M.-F. Chang*, *IEDM*, pp. 16.3.1-4 (2016)

[30] T. Ohsawa, *et al.*, *IEEE Journal of Solid-State Circuits*, vol. 48, no. 6, pp. 1511-1520 (2013)

[31] W. Wang, *et al.*, *IEDM*, pp.1-4, (2006)

[32] S. S. Sheu, *et al.*, *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 245-248 (2013)

[33] S. Yamamoto, *et al.*, *IEEE Custom Integrated Circuits Conference*, pp. 531-534 (2009)

[34] P.-F. Chiu, *et al.*, *IEEE Journal of Solid-State Circuits*, vol. 47, no. 6, pp. 1483-1496 (2012)

[35] W. Wei, *et al.*, *IEEE Transactions on Nanotechnology*, vol. 13, no. 5, pp. 905-916 (2014)

[36] A. Lee, *et al.*, *Symp. VLSI Circuit*, pp. C76-C77 (2015)

[37] M. Qazi, *et al.*, *ISSCC*, pp.192-193 (2013)

[38] S. C. Bartling, *et al.*, *ISSCC*, pp. 432-433 (2013)

[39] N. Sakimura, *et al.*, *ISSCC*, pp. 184-185 (2014)

[40] Y. Liu et al., *ISSCC*, pp. 84-86 (2016)

[41] A. Lee, *et al.*, *IEEE Journal of Solid-State Circuits*, vol. 52, no. 8, pp. 2194-2207 (2017)

[42] J. Li, *et al.*, *Symp. VLSI Circuit*, pp. 104-105 (2013)

[43] S. Matsunaga, *et al.*, *Symp. VLSI Circuit,* pp. 44-45 (2012)

[44] L. -Y. Huang, *et al.*, *Symp. VLSI Circuit*, pp. 99-100 (2014)

[45] M.-F. Chang, *et al.*, *ISSCC*, pp. 318-320 (2015)

[46] M.-F. Chang, *et al.*, *ISSCC*, pp. 136-138 (2016)

[47] A. Chen, 2015, *IEDM*, pp. 10.7.1-4 (2015)

[48] R. Liu, *et al.*, *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, McLean, VA, 2016, pp. 13-18 (2016)

[49] L. Zhang, *et al.*, *IEEE International Symposium on Circuits and Systems (ISCAS),* pp. 2169-2172 (2014)

[50] S. B. Eryilmaz, *et al.*, *IEDM,* pp. 25.5.1-4 (2013)

[51] S. Yu, *et al.*, *IEDM*, pp. 17.3.1-4 (2015)

[52] M. Prezioso, *et al.*, *IEDM,* pp. 17.4.1-4 (2015)

[53] D. Lee, *et al.*, *IEDM*, pp. 4.7.1-4 (2015)

[54] G. W. Burr, *et al.*, *IEDM*, pp. 4.4.1-4 (2015)

[55] J. H. Engel, *et al.*, *IEDM,* pp. 29.4.1-4 (2014)

[56] O. Bichler, *et al.*, *IEEE Transactions on Electron Devices,* vol. 59, no. 8, pp. 2206-2214 (2012)

[57] S. Kim, *et al.*, *IEDM*, pp. 17.1.1-4 (2015)

[58] K. Moon, *et al.*, *IEDM,* pp. 17.6.1-4 (2015)

[59] D. Kuzum, *et al.*, *Nano letters,* vol. 12, no. 5, pp. 2179-2186 (2011)

[60] D. Kuzum, *et al.*, *Nanotechnology*, vol. 24, no. 38, pp. 382001.1-22 (2013)

[61] Y. Peng, *et al.*, *15th Non-Volatile Memory Technology Symposium (NVMTS),* pp. 1-3 (2015)

[62] S. Bandyopadhyay, *et al.*, *Advances in Computing and Communications (ICACC),* pp. 113-116 (2014)

[63] Z. Wang, *et al.*, *International Joint Conference on Neural Networks (IJCNN),* pp. 29-34 (2014)

[64] Y. Wang, *et al.*, *Proceedings of the 25th edition on Great Lakes Symposium on VLSI. ACM*, pp. 189-194 (2015)

[65] B. Li, *et al., Proceedings of the 52nd Annual Design Automation Conference. ACM*, pp. 1-6 (2015)

[66] F. Su, *et al.*, *Symp. VLSI Circuit*, pp.C260-C261 (2017)