Extended Abstracts of the 2018 International Conference on Solid State Devices and Materials, Tokyo, 2018, pp609-610

Emerging Memory Based Circuits for Beyond von Neumann Applications: Nonvolatile-Logic and Computing-in-Memory

Chunmeng Dou[1], Wei-Hao Chen[1], Cheng-Xin Xue[1], Jian-Wei Su[2], Sih-Han Li[2], Ping-Cheng Chen[3]
Huaqiang Wu[4], He Qian[4], and Meng-Fan Chang[1]

[1] National Tsing Hua University, Hsinchu, Taiwan
Phone: +886-3-516-2181 E-mail: mfchang@ee.nthu.edu.tw
[2] Industrial Technology Research Institute, Hsinchu, Taiwan
[3] I Shou University, Kaohsiung City, Taiwan
[4] Tsinghua University, Beijing, China

Abstract

Emerging nonvolatile memories (NVMs), because of their advantages on low power, high speed, and high compatibility with CMOS process, have not only become candidates for future memory technology, but also opened up many opportunities to enable innovative beyond von Neumann architectures. Here, we discuss the advances of applying emerging NVM circuits for nonvolatile logic (nv-Logic) and computing-in-memory (CIM).

1. Introduction

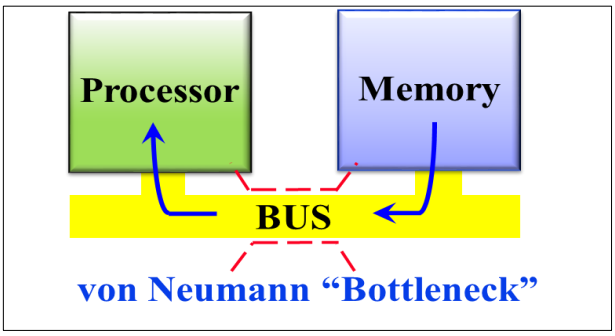

Fig. 1. Illustration of von Neumann Bottleneck

In the relentless pursuit of high performance and low power computing, the data movement between processor and memory has been proven as the main concern that bottlenecks the performance of modern von Neumann architecture (Fig.1) [1]. Figure 2 shows major emerging NVMs, including resistive random access memory (ReRAM), phase change memory (PCM), spin-transfer torque magnetic random access memory (STT-MRAM). They have exhibited strong advantages on power and speed over conventional NVM based on the floating gate structure [2]. Besides, because emerging NVMs are highly compatible to mainstream CMOS processes, they can be easily integrated through back-end-of-line (BEOL) process. Because of these advantages, they have not only aroused extensive attention as candidates next-generation NVM with high energy-efficiency and large bandwidth, but also been considered as key enablers for beyond von Neumann architecture [2-4]. Here, we discuss the frontiers of applying emerging NVM circuits for nv-Logic and CIM.

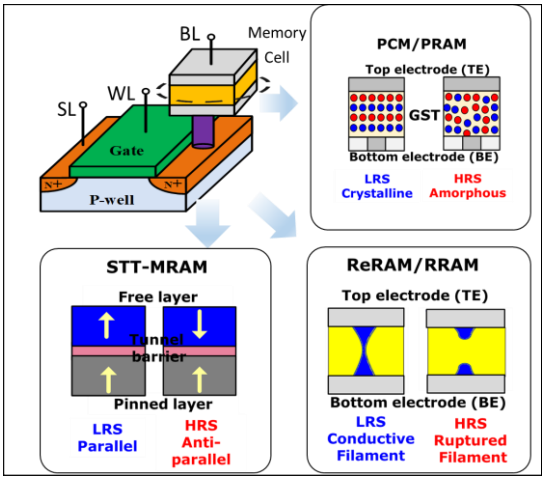

Fig. 2. Major emerging NVM technologies, including PCM, STT-MRAM, and ReRAM.

2. Emerging memory for nonvolatile Logics

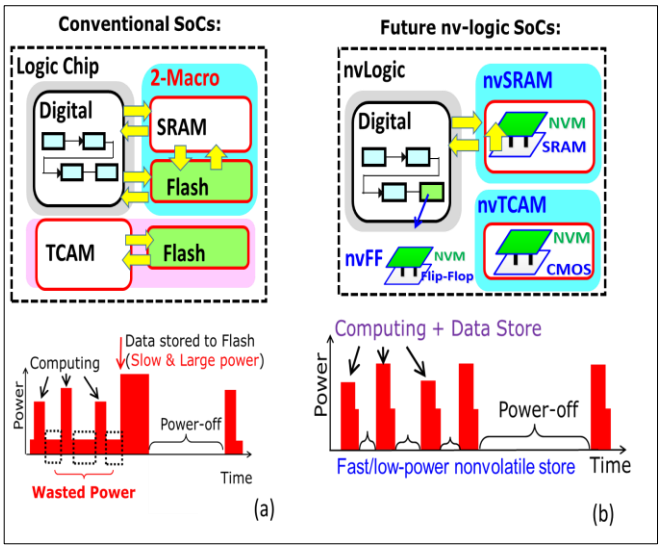

Fig. 3. Conceptual views of SoC chips and their power consumption versus time based on (a) conventional two-macro scheme and (b) nvLogic architecture.

By combining conventional logic circuits and emerging

NVM cells in BEOL, nv-Logic components, such as nonvolatile SRAM (nvSRAM), nonvolatile flip-flops (nvFF), and nonvolatile TCAM (nvTCAM) [6-9], can be achieved. They are particularly energy-efficient in internet-of-thing (IoT) edge devices that requires normally-off and instant-on operations. Figure 4 comparatively show system-on-chip (SoC) chips using a conventional structure (a) and nv-Logic scheme (b). In the conventional scheme, all of the critical data in digital circuits must be serially moved to the flash macro, which incurs considerable power and latency. On the contrary, nv-Logics are able to store their states locally during power-off in a parallel and distributed manner, and thus leads to less consumption of energy and time.

3. Emerging memory for computing-in-memory

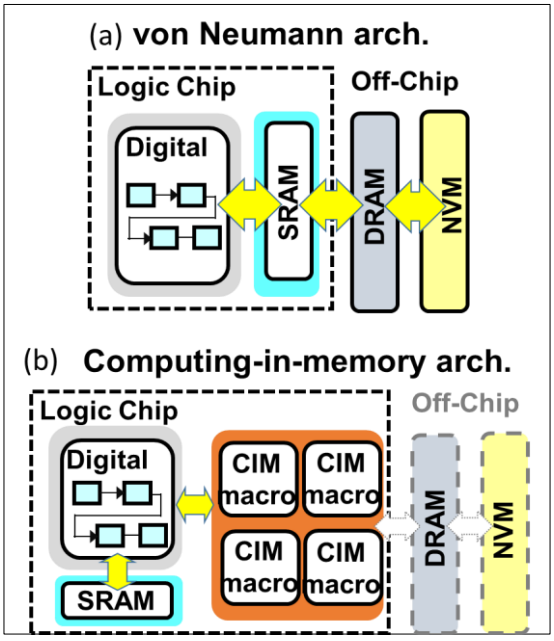

Fig. 4. Conceptual views of von Neumann (a) and CIM architectures.

The recent rise of artificial intelligence (AI) technology requires highly parallel computations, which further exacerbated the von Neumann bottleneck effect. The CIM approach provides a path to circumvent the bottleneck by its capability to carry out computations inside the memory. Figure 4 conceptually compares the von Neumann and CIM system. In a typical von Neumann architecture, the data need to be firstly fetched from NVM through memory hierarchy and thus transferred to digital circuits for computations. On the other hand, cross-memory-hierarchy data movement and the data communication between logic circuits and memory can be effectively reduced using the CIM approach. Besides, the off-chip NVMs are not necessary in the systems with embedded nonvolatile CIM macros based on emerging NVMs [10-12], which further increases the energy- and area-efficiency of the system.

4. Conclusions

In this work, we explore the advances of nv-Logic and CIM leveraging emerging NVM circuits for low-power and small latency computations. They can bypass the von Neumann bottleneck because (1) emerging NVM cells can be operated with low power and fast speed, (2) the cross-memory-hierarchy data access can be reduced, and (3) the serial date movement between logic circuits and memories are effective reduced.

Acknowledgements

We would like to express sincere thanks to our collaborators in TSMC, MXIC, and ITRI.

References

[1] D.A. Patterson and J.L. Hennessy, Computer Architecture: A Quantitative Approach, 2nd ed., Morgan Kaufmann (1996)

[2] M.-F. Chang, Nonvolatile Circuits for Memory, Logic, and Artificial Intelligence, *ISSCC Tutorial* (2018)

[3] C. Dou, et al., Nonvolatile Circuits-Devices Interaction for Memory, Logic and Artificial Intelligence, Symp. *VLSI Tech.*, pp. 171-172, (2018)

[4] H.-S. P. Wong, et al., Memory leads the way to better computing, *Nat. Nanotech*. 10, 191–194 (2015).

[5] M.-F. Chang, et al. Challenges and circuit techniques for energy-efficient on-chip nonvolatile memory using memristive devices, *IEEE J. Emerging and Selected Topics in Circuits and Systems*, 183-193 (2015).

[6] P.-F. Chiu, et al., A Low Store Energy, Low VDDmin, Nonvolatile 8T2R SRAM with 3D Stacked RRAM Devices for Low Power Mobile Applications, *Symp. VLSI Circuits*, pp. 229-230 (2010)

[7] A. Lee, et al., ReRAM-based 7T1R Nonvolatile SRAM with 2x Reduction in Store Energy and 94x Reduction in Restore Energy for Frequent-Off Instant-On Applications, *Symp. VLSI Circuits*, pp. 76-77 (2015)

[8] C.-P. Lo, * et al., A ReRAM-based Single-NVM Nonvolatile Flip-Flop with Reduced Stress-Time and Write-Power against Wide Distribution in Write-Time by Using Self-Write-Termination Scheme for Nonvolatile Processors in IoT Era, *IEDM*, pp. 16.3.1-16.3.4, (2016)

[9] A. Lee, et al., A ReRAM-based Nonvolatile Flip-Flop with Self-Write-Termination Scheme for Frequent-Off Fast-Wakeup Nonvolatile Processors , *IEEE J. Solid-State Circuits*, vol. 52, no. 8, pp. 2194-2207 (2017)

[10] F. Su, et al., A 462GOPs/J RRAM-based nonvolatile intelligent processor for energy harvesting IoE system featuring nonvolatile logics and processing-in-memory, *Symp. VLSI Circuits*, pp. C260-C261 (2017)

[11] W. –H. Chen, et al., A 16Mb Dual-Mode ReRAM Macro with Sub-14ns Computing-In-Memory and Memory Functions Enabled by Self-Write Termination Scheme, *IEDM*, pp. 28.2.1 – 28.2.*4* (2017)

[12] W. –H. Chen, et al., A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processor, *ISSCC*, pp. 494-495 (2018)