Circuit Design Challenges in Computing-in-Memory for AI Edge Devices

Xin Si[1,2], Cheng-Xin Xue[1], Jian-Wei Su[3], Zhixiao Zhang[1], Sih-Han Li[3], Shyh-Shyuan Sheu[3], Heng-Yuan Lee[3], Ping-Cheng Chen[4], Huaqiang Wu[5], He Qian[5], 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] University of Electronic Science and Technology of China, Sichuan, China

[3] Industrial Technology Research Institute, Hsinchu, Taiwan

[4] I Shou University, Kaohsiung City, Taiwan

[5] Tsinghua University, Beijing, China

Abstract—Computing-in-memory (CIM) structures are meant to overcome the memory bottleneck and improve energy efficiency for artificial intelligence (AI) edge devices. In this article, we review recent trends in the development of CIM macros for the Internet of Things and AI applications. We also look at recent advances in the development of CIMs based on SRAM and nonvolatile memory for AI edge devices as well as the challenges involved in circuit design.

Keywords-Artificial Intelligence (AI), Internet of Things (IoT), SRAM, Nonvolatile memory (NVM), computing-in-memory (CIM)

1. Introduction

Rapid development in the Internet of Things (IoT) and artificial intelligent (AI) has increased the amount of data to be moved between the CPU and memory. This issue is referred to as the "memory bottleneck" associated with the conventional von Neumann computing architecture [1]-[6]. Considerable research has gone into the development of high performance memory devices and beyond-von Neumann computing architectures, such as the computing-in-memory (CIM) structure, which is meant to reduce power consumption and latency through improved parallelism and energy efficiency [7]-[15].

In this work, we discuss recent trends in high-performance volatile and nonvolatile memory devices and outline some of the recent advances in the development of CIM macros based on these schemes. We also look at the challenges involved in circuit design.

2. Recent trends in high-performance memory devices

As shown in Fig.1, most existing research on volatile memory, such as SRAM, has focused on (a) faster read/write speeds and wider bandwidth for embedded applications, and (b) reduced power consumption for IoT and AI applications. Rapid advancements in technology nodes (from 90nm to 7nm) have enabled a 35+x decrease in SRAM bit cell size and 1.25+x reduction in the minimum operating supply voltage. Considerable developments have also been made in the design of circuits for write-assist and read-assist schemes [39]-[40].

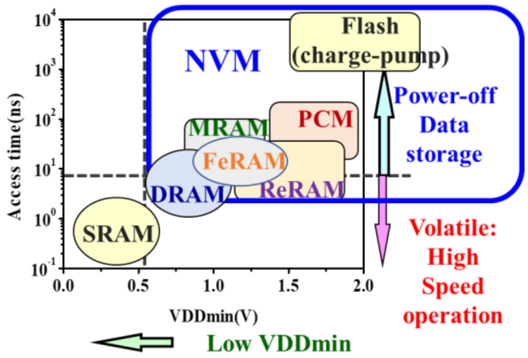

Fig. 1 Trends and performance comparison of volatile and nonvolatile memory devices

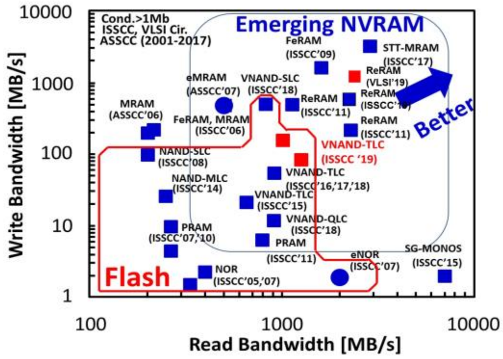

Fig. 2. Trends in write and read bandwidth between Flash and emerging nonvolatile memory (nvRAM)

Emerging nonvolatile memory (eNVM) devices, such as spin-transfer torque magnetic random-access memory (STT-MRAM) [7], resistive random-access memory (ReRAM) [36] and phase change memory (PCM) [8] are compatible with CMOS BEOL processes. They also

978-1-7281-0735-6/19/$31.00 ©2019 IEEE

allow high array density and support lower operating voltages, which makes them ideal for battery-less energy harvesting systems. As shown in Fig. 2, emerging nonvolatile memory technologies have considerable potential in further expanding read and write bandwidth [7]-[33], [44]-[45].

3. Computing in Memory (CIM)

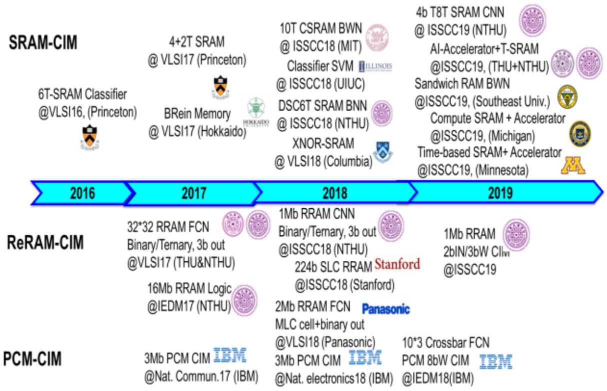

Fig. 3. Roadmap of the development of CIM structures

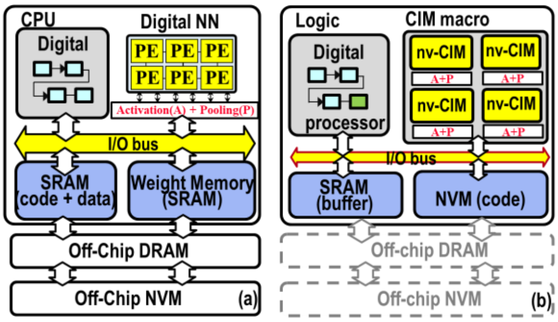

Fig. 4. Comparison of (a) conventional von Neumann and (b) CIM architectures

Deep neural networks (DNNs) [1]-[4] require the movement of large amounts of data between memory devices and the CPU. Multiply and accumulate (MAC) operations dominate the computational workload in convolutional neural networks (CNN) and fully connected neural networks (FC). The computing in memory (CIM) structure was developed to overcome the memory wall in the development of DNN processors for AI edge devices, with the aim reducing power consumption, increasing inference speeds, and enhancing energy efficiency [34]-[35]. Fig. 3 presents a roadmap showing recent developments in the CIM structure.

3.1 Concept of CIM

Computing-in-memory (CIM) overcomes the von-Neumann bottleneck by performing computational processing directly within the memory macros. This enables highly parallel computation, suppresses the amount of intermediate data, and allows the completion of MAC operations within fewer cycles. Fig. 4 presents block diagrams of the conventional von-Neumann and CIM architectures. Numerous silicon proven CIMs have also been developed using SRAM (SRAM-CIM) and NVM (nvCIM); however, they impose a different set of circuit design challenges.

3.2 Circuit Design Challenges of SRAM-CIM

Recent advances in SRAM-CIM include a 6T SRAM-based error adaptive classifier for MNIST recognition [38], a dual-wordline-control 6T SRAM-based SRAM-CIM macro for fully connected layers [20], a 10T Conv-SRAM for binary weight neural networks [37], a deep-in-memory machine learning classifier with on-chip training [43], and a twin-8T SRAM-CIM for multi-bit neural networks [41]. This research has clearly demonstrated the benefits of the CIM architecture in terms of functionality and energy efficiency.

Multibit CNNs are required to improve the inference accuracy of machine learning applications under the increased data complexity following the shift from the MNIST to Image-Net datasets [1]-[6]. Researchers face several challenges and tradeoffs in the design of multi-bit SRAM-CIM: (a) write disturb when performing MAC operations, (b) limited signal margin with an increase in the number of MAC operations, (c) excessive area overhead in signed weight implementation, and (d) large area overhead and power consumption in generating reference signals for multi-bit readouts.

Several solutions have been proposed to overcome these issues, including the use of larger SRAM bit cells (e.g., twin-8T [41] and 10T SRAM cell [37]), the development of small offset sense amplifiers [20], and efficient mapping methods for the CIM structure (e.g., two's complement weight mapping [41]).

3.2 Circuit Design Challenges of nvCIM

Recent developments in CIM architectures based on emerging nonvolatile memory (NVM) devices, such as resistive RAM (ReRAM) [35],[40], and phase change memory (PCM), have greatly improved processing speeds and energy efficiency. The high resistance ratio and ease of mass manufacturing in foundries have made 1T1R single-level-cell (SLC) ReRAM a prime candidate for nonvolatile CIM (nvCIM). Previous ReRAM CIM macros have been used for MAC operations, including binary-input ternary-weight with 3-bit outputs for the MNIST dataset [35], binary-input 8-bit weights with binary-outputs for fully-connected networks, and 2-bit input 3-bit weights with 4-bit outputs for the CIFAR-10 dataset [42].

However, designing circuits for SLC ReRAM-based nvCIM for multi-bit MAC operations imposes a number of challenges: (a) a tradeoff between area cost, speed, and power consumption in the placement of multi-bit inputs and weights within the memory cell array, (b) high input offset, and large parasitic load on the read-path, due to high maximum bitline (BL) current (I_{BL}) across MAC values (MACs), and (c) limited inference accuracy induced by small read margin across various input-weights patterns and variations in memory cell resistance.

4. Summary

Rapid developments in IoT and AI technologies have exposed several fundamental challenges inherent to the von Neumann computing architecture. This article explores recent trends in volatile (SRAM) and nonvolatile memory (including PCM, STT-MRAM, and ReRAM). We also look at recent developments and the challenges ahead in the development of SRAM and nonvolatile memory (NVM)-based CIM structures, the objective of which is to reduce the amount of data that must be moved between processors and memory.

Acknowledgments

We would like to express our sincere gratitude for our collaborators at TSMC, MXIC, and ITRI.

References

[1] M.-F. Chang, Nonvolatile Circuits for Memory, Logic, and Artificial Intelligence, *IEEE International Solid State Circuits Conference (ISSCC) tutorial* (2018)

[2] M.-F. Chang, Challenges of emerging memory and memristor based circuits: Nonvolatile logics, IoT security, deep learning and neuromorphic computing, *IEEE 12th International Conference on ASIC (ASICON)*, pp. 140-143 (2017)

[3] W. H. Chen, and M.-F. Chang et al., Circuit Design for Beyond Von Neumann Applications Using Emerging Memory: From Nonvolatile Logics to Neuromorphic Computing, in *Proc. International Symposium on Quality Electronic Design (ISQED)*, pp. 23-28 (2017)

[4] C. Dou, & M.-F. Chang et al., Emerging Memory Based Circuits for Beyond von Neumann Applications: Nonvolatile-Logic and Computing-in-Memory, *International Conference on Solid State Devices and Materials (SSDM)*, (2018)

[5] F. Su et al., Design of nonvolatile processors and applications, *IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, pp.1-6, (2016)

[6] A. Xu, et al. Scaling for edge inference of deep neural networks, *Nat. Electron* 1, 216-222 (2018)

[7] T.-H. Yang, & M.-F. Chang et al., A 28nm 32Kb Embedded 2T2MTJ STT-MRAM Macro with 1.3ns Read-Access-Time for Fast and Reliable Read Applications, *IEEE International Solid State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 480-481 (2018)

[8] V. Khwa, and M.-F. Chang, et al. A Resistance Drift Compensation Scheme to Reduce MLC PCM Raw BER by Over 100X for Storage Class Memory Applications, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 134-135, (2016)

[9] W. S. Khwa, and M. F. Chang, et al. Novel Inspection and Annealing Procedure to Rejuvenate Phase Change Memory from Cycling-Induced Degradations for Storage Class Memory Applications, *IEEE International Electron Devices Meeting (IEDM)* Dig. Tech. Papers, pp. 29.8.1 - 29.8.4, (2014)

[10] S.-S. Sheu, and M.-F. Chang, et al., A 4Mb embedded SLC Resistive-RAM macro with 7.2ns read-write random access time and 160ns MLC-access capability, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 200-201 (2011)

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

[12] M.-F. Chang et al., Embedded 1Mb ReRAM in 28nm CMOS with 0.27-to-1V Read Using Swing-Sample-and-Couple Sense Amplifier and Self-Boost-Write-Termination Scheme, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 332-333, (2014)

[13] M.-F. Chang, et al., A 0.5V 4Mb Logic-Process Compatible Embedded Resistive RAM (ReRAM) in 65nm CMOS Using Low Voltage Current-Mode Sensing Scheme with 45ns Random Read Time, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 434-435, (2012)

[14] M.-F. Chang, et al., A Low-Voltage Bulk-Drain-Driven Read Scheme for Sub-0.5V 4Mb 65nm Logic-Process Compatible Embedded Resistive RAM (ReRAM) Macro, *IEEE Journal of Solid-State Circuits*, vol. 48, no. 9, pp. 2250-2259 (2013)

[15] M.-F. Chang, et al., A High-Speed 7.2-ns Read-Write Random Access 4-Mb Embedded Resistive RAM (ReRAM) Macro Using Process-Variation-Tolerant Current-Mode Read Schemes, *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 878-891 (2013)

[16] ISSCC Trend, *ISSCC*, vol.10, issue:1, (2019)

[17] J.J. Yang, et al., Memristive devices for computing. *Nat. Nanotech*. 8, 13-24 (2013).

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

[19] D. Ielmini, & H.-S.P. Wong, In-memory computing with resistive switching devices. *Nat. Electron*. 1, 333-343 (2018)

[20] V. Khwa, and M.-F. Chang et al., A 65nm 4Kb Algorithm-Dependent Computing-in-Memory SRAM Unit-Macro with 2.3ns and 55.8 TOPS/W Fully Parallel Product-Sum Operation for Binary DNN Edge Processors, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 496-497, Feb. 2018

[21] X. Sun, Low-VDD Operation of SRAM Synaptic Array for Implementing Ternary Neural Network, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, issue 10, 2962-2965 (2017)

[22] F. Su, & M.-F. Chang et al., A 462GOPs/J RRAM-based nonvolatile intelligent processor for energy harvesting IoE sys-tem featuring nonvolatile logics and processing-in-memory, *Symposium on VLSI Circuits* Dig. Tech. Papers, pp. T260-T261 (2017)

[23] F. Su, et al., A 130nm FeRAM-Based Energy Harvesting Nonvolatile System-On-Chip with 5.2x Higher Performance & 26.9x Faster System Wakeup Using Adaptive Load Balance and Fast Peripheral Startup Schemes, *Symposium on VLSI Circuits* Dig. Tech. Papers, pp. C260-C261 (2017)

[24] X. Li et al., Design of Nonvolatile SRAM with Ferroelectric FETs for Energy-Efficient Backup and Restore, *IEEE Transactions on Electron Devices*, vol. 64, no. 7, pp. 3037-3040 (2017)

[25] X. Li, et al, Enabling Energy-Efficient Nonvolatile Computing With Negative Capacitance FET, *IEEE Transactions on Electron Devices*, vol 64, no. 8, pp.3452-3458 (2017)

[26] P.-F. Chiu, & M.-F. Chang et al., A Low Store Energy, Low VDDmin, Nonvolatile 8T2R SRAM with 3D Stacked RRAM Devices for Low Power Mobile Applications, *Symposium on VLSI Circuits* Dig. Tech. Papers, pp. 229-230 (2010)

[27] A. Lee, & M.-F. Chang et al., ReRAM-based 7T1R Nonvola-tile SRAM with 2x Reduction in Store Energy and 94x Reduc-tion in Restore Energy for Frequent-Off Instant-On Applications, *Symposium on VLSI Circuits* Dig. Tech. Papers, pp. 76-77 (2015)

[28] Y. Liu, & M.-F Chang, et al., Data Backup Optimization for Nonvolatile SRAM in Energy Harvesting Sensor Nodes, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 36, no. 10, pp. 1660-1673 (2017)

[29] C.-P. Lo, & M.-F. Chang 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, *IEEE International Electron Devices Meeting (IEDM)* Dig. Tech. Papers, pp. 16.3.1-16.3.4 (2016)

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

[31] M.-F. Chang, et al, A 3T1R Nonvolatile TCAM using MLC ReRAM for Frequent-Off Instant-On Filters in IoT and Big-Data Processing, *IEEE Journal of Solid-State Circuits*, vol. 52, no. 6, pp. 1664-1679 (2017)

[32] M.-F. Chang, et al, A ReRAM-Based 4T2R Nonvolatile TCAM Using RC-Filtered Stress-Decoupled Scheme for Frequent-OFF Instant-ON Search Engines Used in IoT and Big-Data Processing, *IEEE Journal of Solid-State Circuit*s, vol. 51, no. 11, pp. 2786-2798 (2016)

[33] C.-C. Lin, and M.-F. Chang et al., A 256b-Wordlength ReRAM-based TCAM with 1ns Search-Time and 14x Improvement in Word Length-Energy Efficiency-Density Product using 2.5T1R cell, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 136-137 (2016)

[34] W. –H. Chen, & M.-F. Chang et al., A 16Mb Dual-Mode ReRAM Macro with Sub-14ns Computing-In-Memory and Memory Functions Enabled by Self-Write Termination Scheme, *IEEE International Electron Devices Meeting (IEDM)* Dig. Tech. Papers, pp. 28.2.1 – 28.2.4 (2017)

[35] W.-H. Chen, and M.-F. Chang*, et al., A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processor, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers (2018)

[36] M.-F. Chang, et.al., An offset-tolerant current-sampling-based sense-amplifier for sub-100na-cell-current nonvolatile memory, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers (2011)

[37] Avishek Biswas, Anantha P.Chandrakasan, Conv-SRAM: An Energy-Efficient SRAM with Embedded Convolution Computation for Low-Power CNN-Based Machine Learning Applications, *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers (2018)

[38] Naveen Verma et al., A machine-learning Classifier Implemented in a Standard 6T SRAM Array, *IEEE Journal of Solid-State Circuits (JSSC)* (2017)

[39] M.-F Chang, et al.," A Compact-Area Low-VDDmin 6T SRAM with Improvement in Cell Stability, Read-Speed and Write-Margin Using a Dual-Split-Control Assist Scheme," *IEEE Journal of Solid-State Circuits,* vol. 52, no. 9, pp. 2498-2514, Sept. (2017)

[40] M.-F. Chang, et. al. "A Sub-0.3 V Area-Efficient L-Shaped 7T SRAM With Read Bitline Swing Expansion Schemes Based on Boosted Read-Bitline, Asymmetric-Vth Read-Port, and Offset Cell VDD Biasing Techniques," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2558-2569, Oct. (2013)

[41] X. Si, *et al.*, "A Twin-8T SRAM Computation-In-Memory Macro for Multiple-bits CNN-Based Machine Learning" in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pp. 396-397, Feb. (2019)

[42]C.-X. Xue et al., "A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing time for CNN-based AI Edge Processors," *IEEE International Solid-State Circuits Conference (ISSCC)* Dig. Tech. Papers, pp. 388-390 Feb. (2019)

[43] S. K. Gonugondla et al., "A 42pJ/Decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier with On-Chip Training," IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp. 490-491, Feb. (2018)

[44] M.-F. Chang, et. al., Challenges and trends in low-power 3D die-stacked IC designs using RAM, memristor logic, and resistive memory, in *IEEE International Conference on ASIC*,(2011)

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