Circuit Design for Beyond Von Neumann Applications Using Emerging Memory: From Nonvolatile Logics to Neuromorphic Computing

Wei-Hao Chen[1], Win-San Khwa[1], Jun-Yi Li [1], Wei-Yu Lin[1], Huan-Ting Lin[1], Yongpan Liu[2], Yu Wang[2], Huaqiang Wu[2], Huazhong Yang[2], and Meng-Fan Chang[1]

[1]National Tsing Hua University, Hsinchu, Taiwan
[2]Tsinghua University, Beijing, China
[1]E-mail: mfchang@ee.nthu.edu.tw

Abstract

Emerging memory devices enable performance improvements in memory applications and make possible chip designs using beyond von Neumann architectures. This paper explores the use of emerging memory devices in applications of nonvolatile logics and neuromorphic computing, and provides a review of several silicon examples of nonvolatile logics. This paper also discusses the challenges involved in the design of circuits for nonvolatile logics and neuromorphic computing systems based on emerging memory devices.

Keywords

Emerging memory, ReRAM, RRAM, STT-MRAM, PCM, memristor, nonvolatile Logics, neuromorphic computing
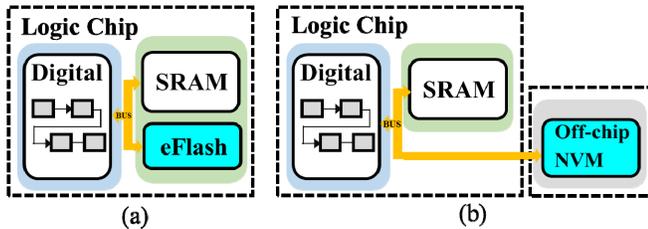
1. Introduction

Fig. 1 Von Neumann-based structure with (a) on-chip and (b) off-chip NVM for intelligent power on-off

Many von Neumann-based energy-efficient systems (Fig. 1) employ on-chip or off-chip nonvolatile memory (NVM) in intelligent power on-off schemes (Fig. 2) aimed at reducing system standby power. This is a particularly important issue in battery-powered or energy-harvester-powered devices equipped with nanometer chips, which are particularly susceptible to current leakage. In these systems, NVM is employed for the storage of programs and critical data in power-off mode [1]-[5].

However, the serial (word-by-word) movement of data between NVM and volatile devices (SRAM, flip-flops) during power off/on operations results excessive power consumption and long access times. This underlines the need for a new circuit architecture (beyond von Neumann architecture) to accommodate intelligent power interruption schemes capable of providing fast speeds and low power consumption.

In the following sections, we discuss two emerging-memory-based approaches to the development of beyond von Neumann architectures: (1) nonvolatile logics (nvLogics) and (2) neuromorphic computing.
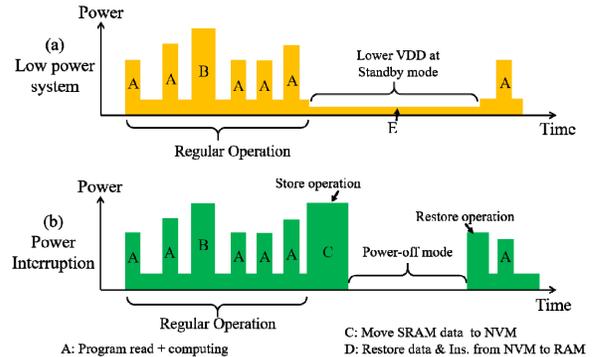

Fig. 2 Standby power reduction schemes: (a) sleep-mode with low VDD; (b) NVM-based intelligent power interruption
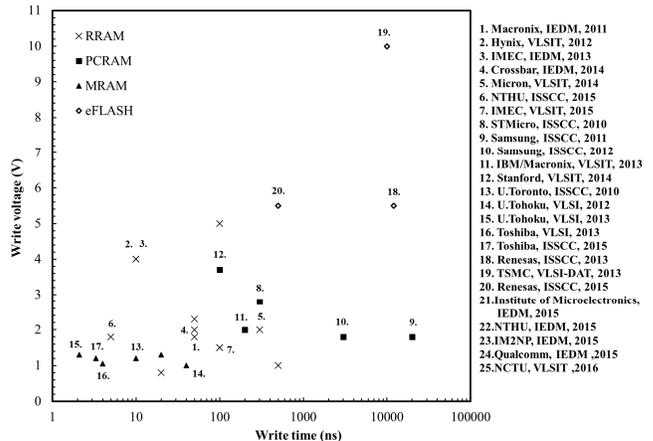
2. Recent Emerging Memory

Fig. 3 Write performance of recent emerging memory devices

Fig. 3 illustrates the performance of recent emerging memory devices, including resistive RAM (ReRAM, RRAM, Memristor) [6]-[17], phase-change memory (PCM) [18]-[23], and spin-transfer-torque magnetic RAM (STT-MRAM) [24]-[28]. These memory devices have much faster write times and lower write voltages than conventional flash memory. This makes it possible for emerging memory

18th Int'l Symposium on Quality Electronic Design

devices to achieve write energy far lower than that of conventional flash memory.
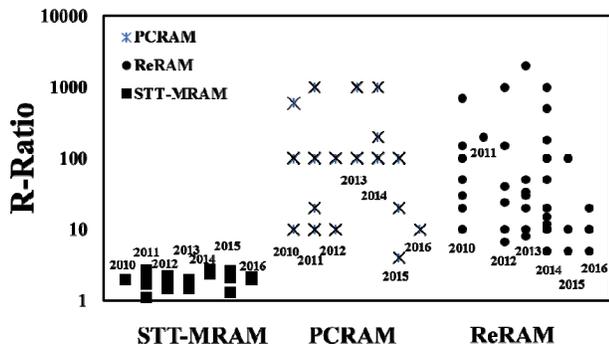

Fig. 4 R-ratio of recent emerging memory devices

Fig. 4 illustrates the resistance-ratio (R-ratio) between the two stored logic-states in recent emerging memory devices. R-ratios have been shrinking in recent years, due to lower write energy and smaller device dimensions.

Fig. 5 presents an example of write-time variations in a ReRAM device [29]. In emerging memory devices, the mean values and distribution of write times (SET and RESET operations) vary with write conditions. Due to process variations, the difference in the period for SET (T_{SET}) or RESET (T_{RESET}) operation between the fastest and slowest cells can exceed 10x or even 100x.
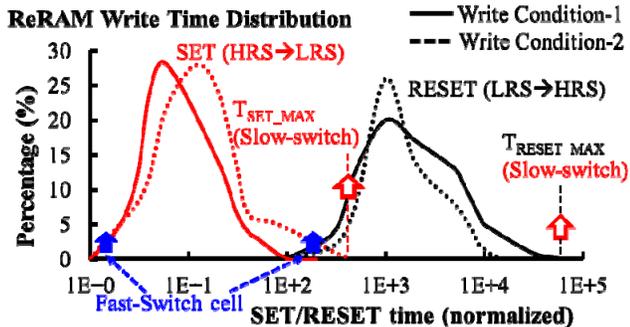

Fig. 5 Example of write-time distribution in ReRAM device

3. Non-Volatile Logics and Nonvolatile Processors

3.1. Concept of nonvolatile logics

Fig. 6 illustrates the concept of nonvolatile logic (nvLogic) and nonvolatile processors (nvProcessors). In conventional von Neumann-based SoC chips, all of the critical computing states in flip-flops (FF) and critical data in SRAM are moved to NVM macros/chips via a shared system bus during power-off operations. The word-by-word and block-by-block sequential movement of data between FF/SRAM and NVM in conjunction with the long NVM-write time of eFlash results in extended power-down latency (T_{STORE}) and large store (power-off) energy consumption (E_{STORE}). Moreover, the movement of data between FF/SRAM and NVM must be controlled by a centralized control unit or the CPU.
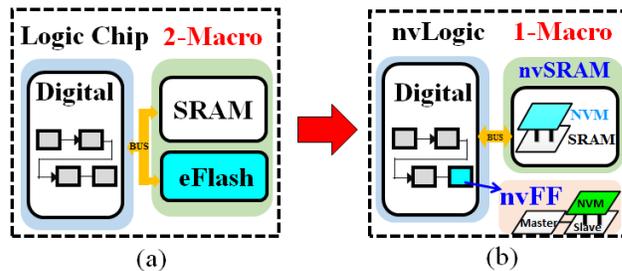

Fig. 6 Concept of nonvolatile processor and nonvolatile logics (a. modified: add a common bus between dig./SRAM/eFlash, b. remove nvTCAM)

As shown in Fig. 6(b), the placement of emerging memory devices above the CMOS devices in CMOS FF or SRAM cells allows for the movement of data between volatile CMOS circuitry and NVM devices within each nvLogic cell; i.e., without using a shared data bus in von Neumann architecture. Unlike the central control scheme in conventional von Neumann-based logic chips, the control of data movement operations is distributed among each nvFF cluster or nvSRAM macro. The parallel movement of data in nvLogic cells enables higher bandwidths, reduced power consumption, and faster store operations than conventional von Neumann-based chips/systems.

3.2. Examples of nvSRAMs and nvFFs

Structure	4T2R	7T2R	8T2R	8T2R	7T1R	7T1R
Schematic						
Cell Area	0.6x	1x	1.55x	1.13x	1.18x	1x
tech. node	MTJ	ReRAM	ReRAM	ReRAM	ReRAM	ReRAM

Fig. 7. Recent silicon-verified nvSRAM cells

Fig. 7 presents examples of recent silicon-verified nonvolatile SRAM (nvSRAM) cells [30]-[36]. Most nvSRAM devices have three operating modes: SRAM, store, and restore. SRAM mode is used for the read and write operations of high-speed or low VDDmin applications, as in regular SRAM devices. Store mode is used for data movement from the storage nodes (Q/QB) of SRAM cells to NVM devices. Restore mode is used for data movement from NVM devices to SRAM storage nodes (Q/QB).

Fig. 8 presents recent silicon-verified nonvolatile Flip-Flops (nvFFs) [37]-[40]. Most nvFFs have three operating modes: Flip-Flop, store, and restore. Flip-Flop mode is used for the regular flip-flop operations of high-speed or low VDDmin applications. Store mode is used for data movement from the storage node (Q/QB) of the slave-stage in FFs to NVM devices. Restore mode is used for data movement from the NVM device to the Q/QB in the slave-stage of FFs.

Structure	22T+2C	18T+4C	17T+2R	15T+2R
Schematic				
tech. node	130nm	130nm	90nm	65nm
Device	FeRAM	FeRAM	MTJ	ReRAM

Fig. 8. Recent silicon-verified nvFF cells

3.3. Challenges in Circuit Design for Emerging memory-based nvLogics

The small R-ratio and wide distribution of cell resistance values in emerging memory devices results in a small difference in signals between nodes Q and QB during restore operations (power-on). This lowers the yield of data restore operations in nvFF and nvSRAM in cases where CMOS transistors are subject to significant process variation.

As illustrated in Fig. 5, nvLogics commonly employ the maximum write time (T_W) for NVM-write (store) operations to ensure sufficient T_W for the emerging memory devices against a wide distribution in NVM write-times. However, a long T_W causes (a) wasted power consumption in fast-switch cells, (b) reliability degradation due to long stress times in fast-switch cells, and (3) write failure in slow-switch cells due to power integrity degradation (lower voltage) induced by the large consumption of DC current by fast-switch cells.

4. Emerging Memory Based Neuromorphic Computing

4.1. Application of Emerging Memory in Neuromorphic Computing

As shown in Fig. 9, many recent studies have employed emerging memory (memristors) as synapses for the storage of weights in biologically inspired neural networks (NN) and neuromorphic computing [41]-[45]. As shown in Fig. 9(b), most memristor-based neuromorphic computing systems employ ReRAM and PCM devices with a high R-ratio and multi-level-cell (MLC) behavior and resistance characteristics [46]-[47]. Unlike the use of SRAM for storing weights, memristor-based synapse implementations enable energy-efficient parallel computing for matrix-vector multiplication (weight sum) using memristor arrays to accumulate a combination of input signals and synapse weightings, as shown in Fig. 9(c).

Modifications to memristor resistance (weight) during learning and training operations is achieved mainly by pulse (spike) control, such as spike-timing-dependent-plasticity (STDP) training [48]-[51].

As shown in Fig. 10, several memristor array structures [42], [52]-[54] can be used for matrix-vector multiplication operations, including (a) cross-bar structures without selectors, (b) cross-bar arrays with selectors, and (c) 1T1R arrays.
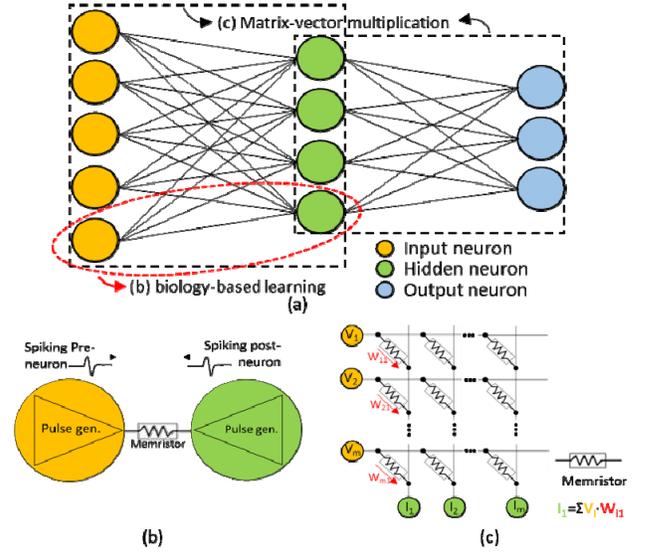

Fig. 9 (a) Simplified model of neural network (NN); (b) using a memristor as a synapse in NN; (c) using a memristor-array for matrix-vector multiplication
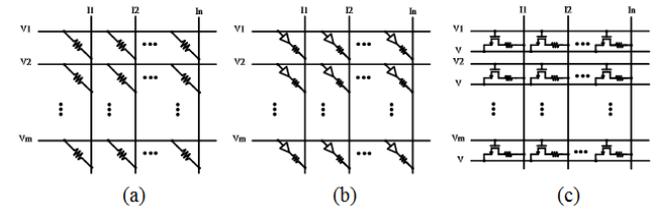

Fig. 10 Memristor array structures

4.2. Challenges in Circuit Designs for Emerging Memory based Neuromorphic Chips

The use of emerging memory devices as memristors for neuromorphic computing involves a number of challenges.

(1) Cell resistance behavior:

Many emerging memory devices have a small R-ratio, do not present linear behavior across a wide range of resistance values, and are limited in the number of cell resistance levels. This underlines the need to develop devices that are suitable for implementation as synapses in neuromorphic circuits. Furthermore, emerging memory devices tend to have wide distribution in cell resistance and write times. This makes it particularly difficult to develop circuits for the implementation of synapses.

(2) Area and power overhead:

The write voltage of most memristor devices is higher than the VDD of core devices. Many emerging memory devices require a larger write (SET/RESET) current to achieve a longer data-retention time. As a result, most of the circuits used for weight training require large area and power consumption.

Moreover, due to analog computing behavior, reading a memristor array requires high-accuracy sense amplifiers for analog-to-digital conversion and integrate-and-fire circuits (IFC)[55]-[56]. This results in large area overhead and high

power consumption for small-offset analog-to-digital conversion, whether using ADC or sense amplifiers.

.

5. Summary

Low write voltage, fast cell switching, and low write power make emerging memory devices highly conducive to beyond von Neumann computing architectures. This paper explores the application of nonvolatile logics and neuromorphic computing based on emerging memory devices. The design of circuits for nvLogics and neuromorphic components involves a number of challenges associated with the characteristics of NVM devices. Novel circuit designs are required to achieve high yields, reduce area overhead, and suppress power consumption.

7. References

[1] Y. Yano, et al. "Take the Expressway to Go Greener", *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers,* pp.24-30, 2012.

[2] M. Hatanaka, et al. "Value creation in SOC/MCU applications by embedded nonvolatile memory evolutions", *IEEE Asia Solid-State Circuits Conf. (ASSCC)*, pp. 38–42, 2007.

[3] H. Hidaka, "Evolution of embedded flash memory technology for MCU", *IEEE International Conference on IC Design & Technology (ICICDT)*, pp. 1–4, 2011.

[4] M. Zwerg, et al. "An 82µA/MHz microcontroller with embedded FeRAM for energy- harvesting applications", *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers*, pp. 334-336, 2011.

[5] Y. Wang, et al. "A 3us wake-up time nonvolatile processor based on ferroelectric flip-flops", *Proceedings of the European Solid-State Circuits Conference (ESSCIRC),* pp. 149-152, 2012.

[6] 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, pp.183-193, 2015.

[7] M.-F. Chang et al. "Read circuits for resistive memory (ReRAM) and memristor-based nonvolatile Logics," *IEEE Asia South Pacific Design Automat. Conf.*, pp. 569–574, 2015.

[8] M.-F. Chang et al. "Challenges at circuit designs for resistive-type nonvolatile memory and nonvolatile logics in mobile and cloud applications," *IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT),* pp. 1–4, 2014.

[9] M.-F. Chang et al. "Endurance-aware circuit designs of nonvolatile logic and nonvolatile SRAM using resistive memory (Memristor) device," *IEEE Asia South Pacific Design Automat. Conf.*, pp. 329–334, 2012.

[10] M.-F. Chang, et al. "Challenges and trends in low-power 3D die-stacked IC designs using RAM, memristor logic, resistive memory (ReRAM)," *IEEE Int. Conf. ASIC*, pp. 299–302, 2011.

[11] M.-F. Chang, et al. "Circuit design challenges in embedded memory and resistive RAM (RRAM) for mobile SoC and 3D-IC," *IEEE Asia South Pacific Design Automat. Conf.*, pp. 197–203, 2011.

[12] L.-Y. Huang, et al. "ReRAM-based 4T2R nonvolatile TCAM with 7 NVM-stress reduction, 4 improvement in speed-wordlength-capacity for normally-off instant-on filter-based search engines used in big-data processing," *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, pp. 1–2, 2014.

[13] K. Eshraghian, et al. "Memristor MOS content addressable memory (MCAM): Hybrid architecture for future high performance search engines," IEEE Trans. Very Large Scale Syst., vol. 19, no. 8, pp. 1407–1417, Aug. 2010.

[14] Q. Luo, et al. "Demonstration of 3D Vertical RRAM with Ultra Low-leakage, High-selectivity and Self-compliance Memory Cells" *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* p. 10.2.1-10.2.4, 2015

[15] H.-W. Pan, et al. "1Kbit FINFET Dielectric (FIND) RRAM in Pure 16nm FinFET CMOS Logic Process" *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* p. 10.5.1-10.5.4, 2015

[16] G. Piccolboni, et al. "Investigation of the potentialities of Vertical Resistive RAM (VRRAM) for neuromorphic applications" *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* p. 17.2.1-17.2.4, 2015

[17] X. Xu, et al. "Fully CMOS Compatible 3D Vertical RRAM with Self-aligned Self-selective Cell Enabling Sub-5nm Scaling" *IEEE Symp. VLS Technology Dig. Tech. Papers*, p. 1-2, 2016

[18] W-S. Khwa, et al. "A Retention-Aware Multilevel Cell Phase Change Memory Program Evaluation Metric", IEEE Electron Device Letters, no.37, pp. 1422-1425, Nov. 2016.

[19] W.-S. Khwa, 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, Dec. 2014.

[20] C.-Y. Wen et al. "A non-volatile look-up table design using PCM (phase-change memory) cells," *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, pp. 302–303, 2011.

[21] M. Rizzi et al. "Statistics of set transition in phase change memory (PCM) arrays," *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 29.6.1–29.6.4, 2014.

[22] H. Pozidis et al. "Reliable MLC data storage and retention in phase-change memory after endurance cycling," *IEEE Int. Memory Workshop*, pp. 100–103, 2013.

[23] M. Boniardi et al. "Optimization metrics for Phase Change Memory (PCM) cell architectures," *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 29.1.1–29.1.4, 2014.

[24] W. J. Kim et al. "Extended scalability of perpendicular STT-MRAM towards sub-20 nm MTJ node," *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 24.1.1–24.1.4, 2011.

[25] E. Kitagawa et al. "Impact of ultra-low power and fast write operation of advanced perpendicular MTJ on power reduction for high-performance mobile CPU", *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers* pp. 29.4.1–29.4.4, 2012.

[26] Y.-H. Wang, et al. "Impact of stray field on the switching properties of perpendicular MTJ for scaled MRAM," *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 29.2.1–29.2.4, 2012.

[27] K. Tsunoda, et al. "Highly manufacturable multi-level perpendicular MTJ with a single top-pinned layer and multiple barrier/free layers", *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers s*, pp. 3.3.1–3.3.4, 2013.

[28] Y. Lu, et al. "Fully Functional Perpendicular STT-MRAM Macro Embedded in 40 nm Logic for Energy-efficient IOT Applications," *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* pp. 26.1.1-26.1.4, 2015

[29] C.-P. Lo, W.-H. Chen, …., M.-F. Chang*, " 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, Dec. 2016.

[30] T. Ohsawa, et al. "A 1 Mb Nonvolatile Embedded Memory Using 4T2MTJ Cell With 32 b Fine-Grained Power Gating Scheme", IEEE Journal of Solid-State Circuits, vol. 48, no. 6, pp. 1511-1520, June 2013.

[31] W. Wang, et al. "Nonvolatile SRAM Cell", *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 1-4, 2006.

[32] S. S. Sheu, et al. "A ReRAM integrated 7T2R non-volatile SRAM for normally-off computing application", *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 245-248, 2013.

[33] S. Yamamoto, et al. "Nonvolatile SRAM (NV-SRAM) using functional MOSFET merged with resistive switching devices", *IEEE Custom Integrated Circuits Conference*, pp. 531-534, 2009.

[34] P.-F. Chiu, et al. "Low Store Energy, Low VDDmin, 8T2R Nonvolatile Latch and SRAM With Vertical-Stacked Resistive Memory (Memristor) Devices for Low Power Mobile Applications", IEEE Journal of Solid-State Circuits, vol. 47, no. 6, pp. 1483-1496, June 2012.

[35] W. Wei, et al. "Design of a Nonvolatile 7T1R SRAM Cell for Instant-on Operation", IEEE Transactions on Nanotechnology, vol. 13, no. 5, pp. 905-916, Sept. 2014.

[36] Lee, et al. "RRAM-based 7T1R nonvolatile SRAM with 2x reduction in store energy and 94x reduction in restore energy for frequent-off instant-on applications",

[37] M. Qazi, et al. "A 3.4pJ FeRAM-enabled D flip-flop in 0.13μm CMOS for nonvolatile processing in digital systems", *IEEE International Solid-State Circuits Conference Dig. Tech. Papers (ISSCC)*, pp.192-193, 2013

[38] S. C. Bartling, et al. "An 8MHz 75μA/MHz Zero-Leakage Non-Volatile Logic-Based Cortex-M0 MCU SoC Exhibiting 100% Digital State Retention at VDD=0V with <400ns Wakeup and Sleep Transitions", *IEEE International Solid-State Circuits Conference Dig. Tech. Papers (ISSCC)*, pp. 432-433, 2013

[39] N. Sakimura, et al. ""A 90nm 20MHz fully nonvolatile microcontroller for standby-power-critical applications," *IEEE International Solid-State Circuits Conference Dig. Tech. Papers (ISSCC)*, pp. 184-185, 2014."

[40] Y. Liu, et al. "4.7 A 65nm ReRAM-enabled nonvolatile processor with 6x reduction in restore time and 4x higher clock frequency using adaptive data retention and self-write-termination nonvolatile logic," *IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers*, pp. 84-86, 2016.

[41] Eryilmaz, S. Bur, et al. "Experimental demonstration of array-level learning with phase change synaptic devices." *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* pp. 25.5.1-25.5.4., 2013.

[42] Yu, Shimeng, et al. "Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect." *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 17.3.1-17.3.4, 2015.

[43] Prezioso, M., et al. "Modeling and implementation of firing-rate neuromorphic-network classifiers with bilayer Pt/Al2O3/TiO2??? x/Pt Memristors." *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* pp. 17.4.1-17.4.4, 2015.

[44] Lee, Daeseok, et al. "Oxide based nanoscale analog synapse device for neural signal recognition system." *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 4.7.1-4.7.4, 2015.

[45] Burr, G. W., et al. "Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power)", *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 4.4.1-4.4.4, 2015.

[46] Engel, Jesse H., et al. "Capacity optimization of emerging memory systems: A shannon-inspired approach to device characterization." *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* pp. 29.4.1-29.4.4, 2014.

[47] Bichler, Olivier, et al. "Visual pattern extraction using energy-efficient "2-PCM synapse" neuromorphic architecture." IEEE Transactions on Electron Devices 59.8, pp. 2206-2214, 2012.

[48] KIM, S., et al." NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with

On-chip neuron circuits for continuous in-situ learning"*IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers*, pp. 17.1.1-17.1.4, 2015.

[49] Moon, Kibong, et al. "High density neuromorphic system with Mo/Pr0. 7Ca0. 3MnO3 synapse and NbO2 IMT oscillator neuron." *IEEE International Electron Devices Meeting (IEDM) Dig. Tech. Papers,* pp.17.6.1-17.6.4, 2015.

[50] Kuzum, Duygu, et al. "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing." Nano letters 12.5, pp.2179-2186, 2011.

[51] Kuzum, et al. "Synaptic electronics: materials, devices and applications." Nanotechnology 24.38, 382001, 2013.

[52] Yao, Peng, et al. "The Effect of Variation on Neuromorphic Network Based on 1T1R Memristor Array." *Non-Volatile Memory Technology Symposium (NVMTS), 2015 15th.*

[53] Bandyopadhyay, Subhankar, et al. "Arbitrary Waveform Generation Using Memristive Cross Bar Array" *Advances in Computing and Communications (ICACC)*, 2014.

[54] Wang, Zhao, et al. "Ferroelectric tunnel memristor-based neuromorphic network with 1T1R crossbar architecture" *International Joint Conference on Neural Networks (IJCNN)*, 2014

[55] Wang, Yu, et al. "Energy efficient RRAM spiking neural network for real time classification." *Proceedings of the 25th edition on Great Lakes Symposium on VLSI. ACM*, pp. 189-194, 2015.

[56] Li, Boxun, et al. "Merging the interface: Power, area and accuracy co-optimization for rram crossbar-based mixed-signal computing system." *Proceedings of the 52nd Annual Design Automation Conference. ACM*, pp.13, 2015.

[57] Chi, Ping, et al. "PRIME: A Novel Processing-In-Memory Architecture for Neural Network Computation in ReRAM-based Main Memory" Proceedings of ISCA. Vol. 43. 2016.

[58] Tang, Tianqi, et al. "Spiking neural network with rram: Can we use it for real-world application?." *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition. EDA Consortium (DATE)*, pp. 860-865, 2015.

[59] Liu, Chenchen, et al. "A spiking neuromorphic design with resistive crossbar." *Proceedings of the 52nd Annual Design Automation Conference. ACM*, pp. 14., 2015.