Extended Abstracts of the 2018 International Conference on Solid State Devices and Materials, Tokyo, 2018, pp17-18

Ultralow-Power and Compact Nonvolatile Brain-Inspired VLSIs Based on CMOS/MTJ Hybrid Technology

Tetsuo Endoh [1,2,3,4] and Yitao Ma[1,2,3]
[1] Center for Innovative Integrated Electronic Systems (CIES), Tohoku University
[2]Graduate School of Engineering, Tohoku University
[3]Center for Science and Innovation Spintronics (Core Research Cluster), Tohoku University
[4]Center for Spintronics Research Network, Tohoku University
468-1 Aramaki-aza-Aoba, Aoba-ku, Sendai, Miyagi 980-0845, Japan

Abstract

In this invited paper, it is discussed that CMOS/MTJ hybrid VLSI technology has an impact in both brain-inspired computing and neuromorphic computing. Our previous researches show that CMOS/MTJ hybrid VLSI technology is very effective approach to overcome both the power issue and the large device number issue of AI chip technology field.

1. Introduction

Benefiting from the rapid progress of CMOS technology, von-Neumann VLSIs may still remain as the mainstream of fast, symbolic and number-crunching calculators in the near future. However, they face the insurmountable problems on intelligent applications, such as image recognition and automotive car control, to build systems that can learn, reason and help humans make better decisions. To solve this problem, the dedicated VLSIs mimicking functions of the brain (brain-inspired) or operation of the neuron (neuromorphic) have been widely investigated relying on conventional CMOS technology to achieve excellent computational speed. However, they generally require the large-capacity volatile embedded memory to store the data on-chip for highly concurrent processing, remaining the large power consumption as inescapable issue for practical uses in power-critical systems such as smartphones and sensor networks.

2. Superiorities of CMOS/MTJ hybrid VLSI Technology

The emerging nonvolatile memories are highly expected instead of the conventional SRAM to fulfill the large on-chip memory requirement of the brain-inspired VLSIs and reduce the power consumption for full operation period at the same time. There are various kinds of nonvolatile memories developed in recent year, such as FeRAMs, PCRAMs, ReRAMs and STT-MRAMs. The performance benchmark for each memory is summarized in Table I with NAND flash and SRAM also added for references.

The most prospective candidate in these nonvolatile memories for the use in brain-inspired VLSIs requests small cell size at first to realize the basic recognition functions with big data. The high read/write speed is also necessary to achieve not only the high-speed computing but also the high-speed power gating for power reduction. Furthermore, the high Endurance performance is indispensable in order to realize the learning function with heavy data overwriting, which becomes more and more important in recent research of advanced brain-inspired VLSIs. From all above requests, the STT-MRAM with small cell size, high access speed and high endurance shows great superiority for brain-inspired VLSI applications. In addition, the appropriate operation voltage makes STT-MRAM more compatible than FeRAM for combing with CMOS, and the relative low write current of STT-MRAM makes it more low-power than PCRAM and ReRAM. Therefore, Mass-production of STT-MRAM [1] was announced by many companies.

3. CMOS/MTJ hybrid Nonvolatile Brain-Inspired VLSIs

A) Brain-Inspired Computing Approach

Focusing on the indispensable nearest neighbor search (NNS) function of the brain, we have developed a nonvolatile object recognition processor with NSS full-adaptive to any data format, employing nonvolatile memories base on our IPMA type perpendicular MTJ (p-MTJ) [2]. The prototype object recognition processor is fabricated under 90nm-CMOS/70nm-MTJ hybrid process on 300mm-wafer [3]. The 4-Transistor (4T) 2-MTJ memory cells are adopted to completely eliminate standby power. A self-directed power-gating technique leveraging the non-volatility, high access speed and unlimited endurance features of the p-MTJs is employed to shut down idle circuit blocks during not only the standby periods but also the full operation periods. The measured peak operation power consumption of the prototype chip is only 130μW and can be further optimized corresponding to the format of reference data (Fig. 1).

Table I: Specification Comparisons of Various Nonvolatile Memories (* SLC: Single-level cell)

Performance	FeRAM	PCRAM	ReRAM	STT-MRAM	NAND flash	SRAM
Ideal Cell size (SLC)	15 ~ 35 F^2	4 ~ 19 F^2	6 ~ 10 F^2	6 ~ 14 F^2	4 F^2	160 ~ 280 F^2
Operation voltage	~ 1.8 V	1.5 ~ 1.8 V	3.3 ~ 6.5 V	0.8 ~ 1.8 V	~ 20 V	0.6 ~ 1.1 V
Write current	~ 10^{-6} A	~ 10^{-4} A	~ 10^{-4} A	~ 10^{-5} A	~ 10^{-7} A	~ 10^{-5} A
Write time	< 10 ns	~ 100 ns	~ 50 ns	< 10 ns	~ 1 ms	≤ 2 ns
Read time	< 5 ns	< 5 ns	< 5 ns	< 5 ns	~ 100 μs	≤ 2 ns
Retention	> 10 yrs	> 10 yrs	> 10 yrs	> 1month />10 yrs	> 10 yrs	(volatile)
Endurance	10^{13}	10^9 ~ 10^{12}	~ 10^6	~ 10^{15}/~ 10^{12}	~ 10^5	~ 10^{15}

Compared to the latest conventional researches [4-8], the significant improvements of both power performance and circuit density are achieved as shown in Fig. 2 [3, 9].

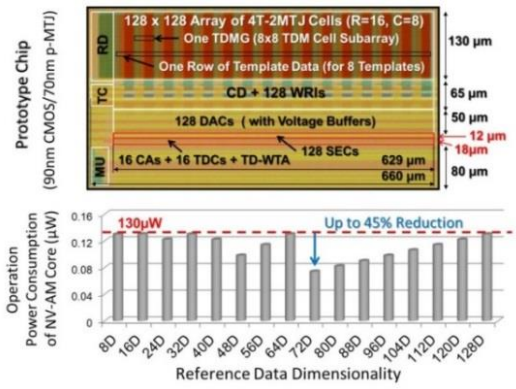

Fig. 1 Prototype chip micrograph of 90nm-CMOS/70nm-MTJ NV-Brain-inspired object recognition processor and its results of self-directed optimization of operation power consumption for different dimensionality of reference data.

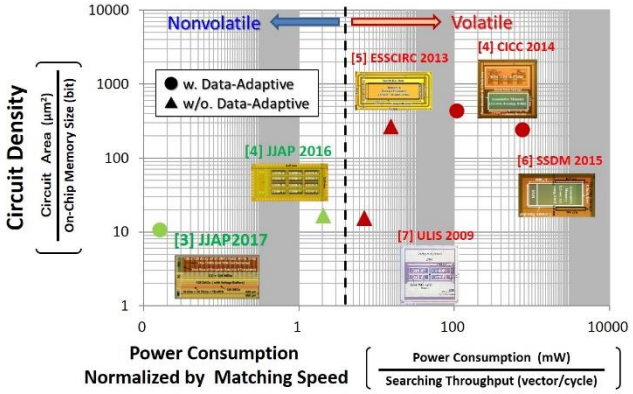

Fig. 2 Further performance comparisons of the operation power efficiency (normalized by matching throughput) and circuit density, among the developed NV-brain-inspired object recognition processor and recent reported processors.

B) Neuromorphic Computing Approach

As the next-generation intelligent VLSIs, neuromorphic processors inspiring closer to the brain/neuron operation have attracted significant attentions, where spike signals are utilized for learning and perception. The results of our recent works show that leveraging p-MTJs as memristive devices in the neuromorphic chips is a prospective approach for resolving the issues of both power consumption and circuit complexity. We have proposed a novel neuron circuit employing 8 nonvolatile synapses (NVS) with very compact 17T-2MTJ circuitry as in Fig.3 [10, 11], where the dual-state synaptic weights are recorded into a 2T-2MTJ latch circuitry during learning and recalled for recognition. An ultimate power gating method is realized using the spike driven scheme, where input spike signals are also utilized as the dynamic power source of NVS without any other static power supply. Thus, almost no power is consumed when the spike keeps the low voltage. The prototype circuit is implemented with 90nm-CMOS / 70nm-MTJ and 90nm-CMOS / 35nm- MTJ technology, respectively. The high-stable spike-driven operations for object recognition using 17T-2MTJ synapses are verified even with 5% variation in MTJ resistances, and this operation stability is proved to be further improved with the scaled-down p-MTJs. More than 68.5% reduction of the gate count is achieved comparing to recent conventional systems (Fig. 4).

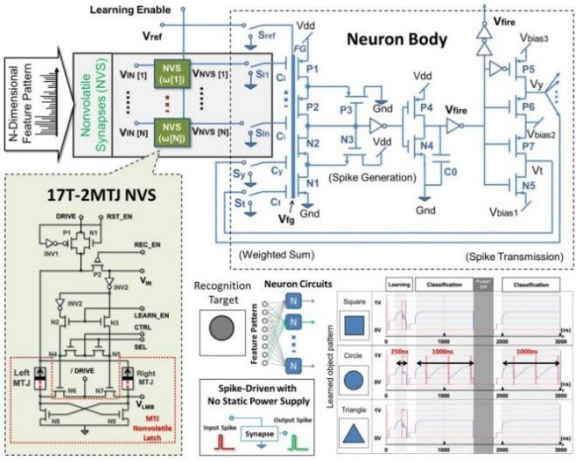

Fig. 3 Architecture of developed 90nm-CMOS/35nm-MTJ neuron circuit using 17T-2MTJ nonvolatile synapses with spike-driven scheme and its operation verification results for object recognition.

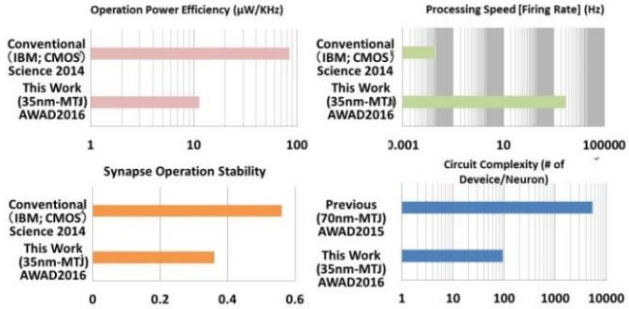

Fig. 4 Comparisons of power, speed, operation stability and circuit complexity performances between developed neuron circuit and the conventional neuromorphic circuit.

4. Conclusions

From above all, the CMOS/MTJ hybrid VLSI technology can be highly expected to hold the center stage of future development of both brain-inspired AI VLSIs and neuromorphic AI VLSIs.

Acknowledgements

This work is supported by CIES's Industrial Affiliation on STT-MRAM program, ACCEL and OPERA under JST.

References

[1] T. Endoh et al., IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol. 6, 109 (2016).
[2] S. Ikeda et al, Nature Materials, **9** (2010) 721-724.
[3] Y.Ma, S.Miura, H.Honjo, S.Ikeda, T.Hanyu, H.Ohno, T.Endoh, Jpn. J. Appl. Phys., **56** (2017) 04CF08.
[4] F. An, et al, CICC, (2014) 1.
[5] T. Akazawa, et al, ESSCIRC, (2013) 267.
[6] F. An, et al, SSDM, (2015) 144.
[7] T. Bui, et al, ULIS, (2009) 213.
[8] Y.Ma, et al, Jpn. J. Appl. Phys., **55** (2016) 04EF15.
[9] T.Endoh, et al, IEEE ISCAS,(2018)
[10] Y. Ma and T. Endoh, AWAD, (2015) 273-278.
[11] Y. Ma and T. Endoh, AWAD, (2016) 268-262.