Ultra-fast and Energy-efficient Write-Computing Operation for Neuromorphic Computing

Liang Chang[1], Zhaohao Wang[2*], Youguang Zhang[1], and Weisheng Zhao[2]
[1]Fert Beijing Research Institute, BDBC, School of Electronic and Information Engineering, Beihang University.
[2]Fert Beijing Research Institute, School of Microelectronics, BDBC, Beihang-Geortek Joint Microelectronics Institute,
Qingdao Research Institute, Beihang University. *Zhaohao.Wang@buaa.edu.cn

***Abstract*—Emerging Non-volatile memory (NVM) has demonstrated superior performance on the computing-in-memory (CIM) architecture. By re-purposing the peripheral circuits, the certain NVM array can perform both storage and computing operations to accelerate the data-intensive convention Neural Networks (CNNs). However, the parallelism of the NVM-based CIM should be considered. In this paper, we present a CIM architecture developed by the Spin-orbit Torque (SOT) MRAM using both read-out and write-in operations. We highlight the memory structure and control method of the write-in operations. With the excellent write performance of SOT-MRAM, the proposed write-in operation can obtain ultra-fast and energy-efficient computing data operations. The write-in operation works as a complement of the conventional read-out CIM architecture rather than replace it.**

***Index Terms*—Computing-in-Memory, Spin-Orbit Torque MRAM, Write-based Logic, Computing element, Neural Network**

I. Introduction

Recent developments of computing-in-memory (CIM) architecture promote advanced neural network accelerators. Compared with traditional neural network accelerators, the CIM architecture can significantly reduce the data movement between the computation unit (such as Center Process Unit) and the main memory [1], [2]. By performing most of the data-intensive operations with/within the memory structure, the computation efficiency of the CIM accelerator is also improved. Among the variety of CIM designs, the CIM design with NVM technologies has become a good candidate.

Fig. 1 provides the concept the NVM-based CIM architecture. Typically, both the operand rows and results are transferred using the narrow interface, as shown in Fig. 1 (a). By contrast, the NVM-based CIM architecture employs the modified sense amplifier to calculate the results without data movement on the interface, as demonstrated in Fig. 1 (b). In addition, the NVM-based CIM architecture also can be performed by the modified write driver. In this paper, we discuss the write-computing data operations based on the Spin-Orbit-Torque magnetic random access memory (SOT-MRAM), to obtain an ultra-fast and energy-efficient CIM architecture for neuromorphic computing.

Fig. 2 indicates the bit-cell of MRAM. Generally, the core element of the MRAM is the magnetic tunnel junction (MTJ) with a tunnel barrier (TB) sandwiched between a ferromagnetic pinned layer (PL) and a free layer (FL). The parallel (P) or anti-parallel (AP) magnetization alignments between the PL and FL decides the value “0” and “1”, respectively. Compared to the spin-transfer torque MTJ (STT-MTJ) as shown in Fig. 2 (a), a heavy metal is added and contacted to the FL in the SOT-MTJ as shown in Fig. 2 (b). Only a small and short-pulse current flows the HM to change the state of the FL solving the incubation delay and large current problems of STT-MTJ [3]–[7]. In this paper, SOT-MTJ with anti-ferromagnet/ferromagnet bilayer structure is employed to develop the CIM architecture [8]. The SOT-MTJ is suitable for the CIM architecture thanks to the ultra-fast and energy-efficient write operation.

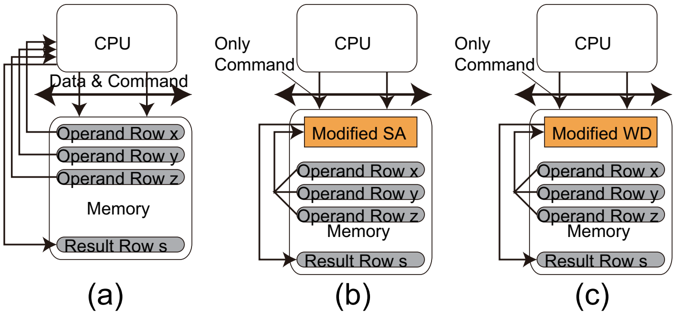


Fig. 1. The concept of computing-in-memory with the NVM technology. (a) The conventional load-store-computing structure with CPU and main memory. (b) The novel read-computing data CIM architecture, where the computation is performed by the modified sense amplifier (SA). (c) The novel write-computing data CIM architecture, where the computation is implemented using the modified write driver (WD).

II. Write-Computing Data Operation

Fig. 3 (a) indicates the array structure of the computing element, which includes function control unit, row/column address control unit, sense amplifier, output driver, modified write driver, and the memory array developed by the SOT-MRAM bit-cell. In the write driver, we add more control transistors compared to the conventional write driver (i.e., one transistor). Fig. 3 (b) gives an example of the write driver with three control signals $\overline{A}$, $\overline{B}$, and $\overline{C}$. $\overline{C}$ is the functional control signal for controlling the bit-wise OR and AND operations. $\overline{A}$ and $\overline{B}$ are the input signals. A reset/load signal is used to initialize the state of the SOT-MTJ and control the direction of the write current. Both BL and SL are connected to the different write drivers for different bit-wise operations. The function of the memory structure is controlled by the function

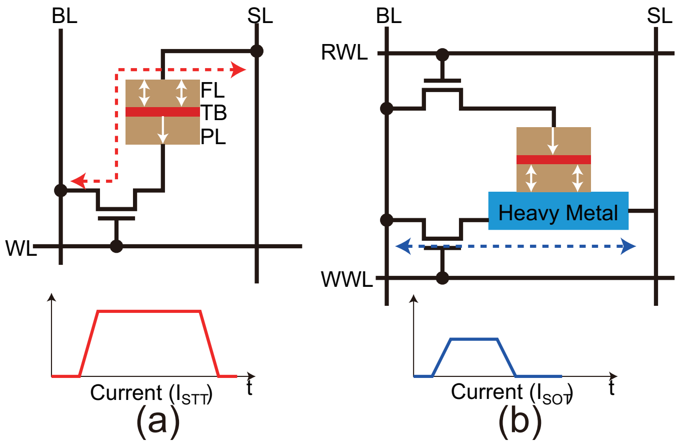
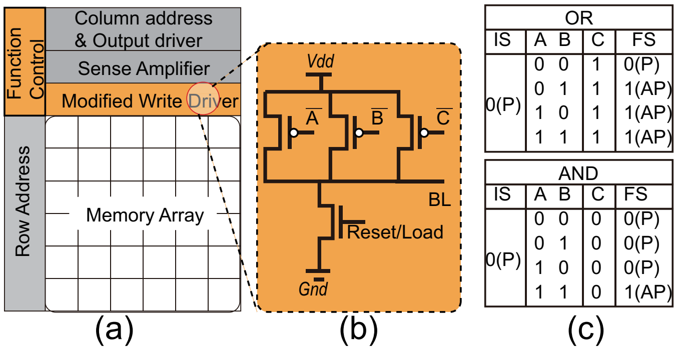


Fig. 2. The bit-cells of the STT-MRAM and SOT-MRAM. (a) Bit-cell of STT-MRAM with 1T1MTJ. (b) Bit-Cell of SOT-MRAM with 2T1MTJ.



OR				
IS	A	B	$\overline{C}$	FS
0(P)	0	0	1	0(P)
	0	1	1	1(AP)
	1	0	1	1(AP)
	1	1	1	1(AP)

AND				
IS	A	B	$\overline{C}$	FS
0(P)	0	0	0	0(P)
	0	1	0	0(P)
	1	0	0	0(P)
	1	1	0	1(AP)

Fig. 3. The concept of write-computing data scheme. (a) A memory array with a modified write drvier. (b) The modified write driver with several control transistors (three in the figure). (c) The truth tables for the OR and AND bit-wise operations with the SOT-MTJ.

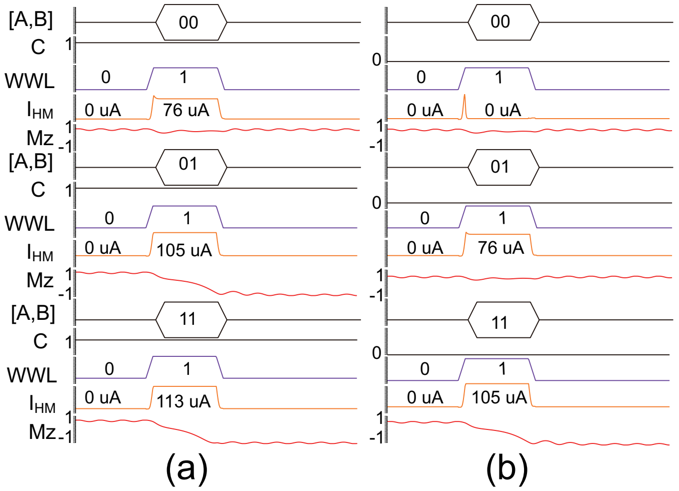


Fig. 4. The validation and evaluation results of the bit-wise OR and AND operations by using the modified write driver. (a) Results of OR operation. (b) Results of AND operation.

control unit. The function control unit can be configured before using the memory structure. We can support normal memory access, read-computing (to modify the sense amplifier), and write-computing modes. In the computing mode, most bit-wise operations can be supported such as OR/NOR, AND/NAND, XOR/XNOR, and minority/majority operations.

Fig. 3 (c) indicates the truth tables of bit-wise OR and AND operations. The initial state (IS) of the SOT-MTJ for these two operations is P state representing binary value "0". The control signal $\overline{C}$ is configured to "1" or "0" to control the OR or AND operation, respectively. The mechanism of the write-computing interpreted as follows: more activated control transistors provides more sufficient current to write the SOT-MTJ.

III. Experiment and Results

In the device level, we employ the SOT-MTJ model and parameters provided by [5]. In the circuit level, we develop the write driver and the bit-cell under CMOS 28 nm technology using Cadence SPICE simulation tool. We set the three NMOS transistors in the same size. The critical switching of the SOT-MTJ is $\sim 90\mu A$. Fig. 4 demonstrates the validation and evalution results. Only one transistor is activated, the current added on the heave metal is $76\mu A$. The magnetization of the SOT-MTJ holds the initial state (P) as the final state (FS). The magnetization can be changed as more than one transistors are activated. These results match the truth table shown in the Fig.3 (c). Also, we observe that the write-computing operation is ultra-fast, which only requires 0.5 ns for both bit-wise operations. The average energy dissipation is $46.5fJ$ and $30.5fJ$ for bit-wise OR and bit-wise AND operations, respectively. These bit-wise operations can be used to accelerate the binary neural networks. The write-computing operations are combined with read-computing operation to developed streamed patterns to improve the parallelism of the CIM architecture [9].

IV. Conclusion

The parallelism of the NVM-based CIM architecture could be further improved by using both the read-out and write-in operations. This paper proposed the write-computing data operations through modifying the write driver of the memory structure. The hybrid CMOS/MTJ simulation results demonstrated the ultr-fast and energy-efficent results.

Acknowledgement

The work was supported by the National Natural Science Foundation of China (Grant No. 61704005, 61627813).

References

[1] S. Angizi et al., in *2019 DATE*. IEEE, 2019, pp. 378–383.
[2] L. Chang et al., in *2019 DATE*. IEEE, 2019, pp. 384–389.
[3] Z.Wang, et al., *IEEE EDL*, vol. 39, no. 3, pp. 343–346, 2018.
[4] M. Wang, et al., *Nature communications*, vol. 9, no. 1, p. 671, 2018.
[5] L. Chang et al., in *2017, ICCAD*. IEEE, 2017, pp. 245–252.
[6] M. Wang, et al., *Nature electronics*, vol. 1, no. 11, p. 582, 2018.
[7] Z. Wang, et al., *IEEE EDL*, vol. 40, no. 5, pp. 726–729, 2019.
[8] S. Fukami, et al, *Nature Materials*, vol. 15, no. 5, p. 535, 2016.
[9] L. Chang et al., *IEEE TVLSI*, 2019.