A High-speed and High-reliable Fully Digital STT-MRAM Based Computing-in-Memory for Binary Neural Network

Yongcheng Wang[1], Tao Li[2], Li Zhang[2], and Tetsuo Endoh[1,2], *Fellow, IEEE*

[1]School of Engineering, Tohoku University, Sendai, 980-8577, Japan, wang.yongcheng.p7@dc.tohoku.ac.jp
[2]Center for Innovative Integrated Systems, Sendai, 980-0845, Japan, tetsuo.endoh.b8@tohoku.ac.jp

This work proposes a novel STT-MRAM based Computing-in-Memory (CiM) cell architecture for binary neural network that addresses the reliability limitations inherent in prior analog STT-MRAM CiM designs, which suffer from low resistance and limited tunnel magnetoresistance (TMR) in magnetic tunnel junctions (MTJs). The proposed CiM technique introduces key enhancements: (1) accelerated computation via a differential-type memory configuration; (2) improved tolerance to TMR variation through a fully digital scheme that decouples storage and computation (3) reduced power consumption owing to the high-speed computing that makes longer no power standby durations. With SPICE simulations, it is shown that the novel CiM brings over a 50% reduction in computation latency, enhanced robustness to TMR fluctuations, and a 26.2% decrease in computation power consumption at a supply voltage of 1.2 V.

***Index Terms*— Computing-in-memory (CiM), spin-transfer torque magnetic random access memory (STT-MRAM), high-speed and high-reliable, low-power, binary neural network (BNN).**

I. Introduction

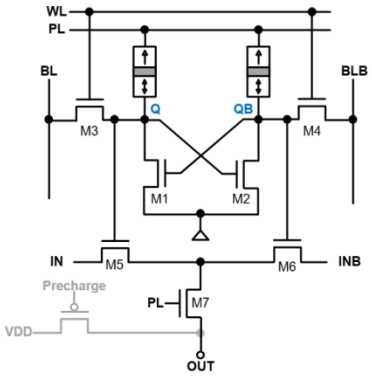
Spin-torque-transfer magnetic random access memory (STT-MRAM) has advantages of low power, high endurance, etc., that considered as an ideal candidate for computing-in-memory (CiM). Existing STT-MRAM based CiM architectures predominantly employ analog multiply-accumulate operations using the intrinsic characteristics of magnetic tunnel junctions (MTJs) [1]-[3]. However, these analog designs are challenged in terms of reliability and readout latency due to the low TMR of MTJs [2]. In contrast, a digital CiM design introduced in [4] utilizes a counter-based accumulation mechanism, improving computational accuracy by avoiding current-sum or capacitor-sum methods. However, the method proposed in [4] still employs an analog computing approach for the multiplication operation, which suffers from the same issue reported in [2].

To overcome these constraints, the proposed work introduces a fully digital STT-MRAM based CiM cell for binary neural network (BNN) applications. This architecture achieves complete digitalization in both multiplication and accumulation, thereby significantly enhancing computation reliability, latency and thus a power reduction.

II. Proposed STT-MRAM based digital CiM architecture

The proposed CiM cell structure, shown in Fig. 1, comprises seven NMOS transistors and two MTJs. The upper sub-circuit adopts a 4-transistor-2-MTJ differential-type STT-MRAM configuration [5], while the lower portion incorporates an XNOR logic circuit tailored for BNN operations. Transistors M5 and M6 implement the XNOR logic, and M7 functions as an output switch. During idle states, the output node (OUT) is precharged to V_{DD} using an external and shared switch.

For BNN processing, weight data are written into the MTJs via BL/BLB lines. During standby periods, the power line (PL) is power gated to minimize leakage. Upon activation, the PL restores stored weight data from the MTJs to the internal nodes Q and QB, which govern M5 and M6. The activation inputs IN and INB subsequently determine the state of OUT, which either remains high or discharges, representing logical '1' or '0', respectively.



Q/QB	IN/INB	OUT
0/1	0/1	1
0/1	1/0	0
1/0	0/1	0
1/0	1/0	1

Fig. 1. Proposed CiM cell structure and the XNOR operand table.

III. Evaluation and benchmarking

The proposed CiM cell was designed using a 55 nm CMOS technology. For comparative analysis, a reference circuit from [4] was replicated using the same CMOS model parameters. An MTJ model characterized by a nominal TMR of 122% and an ambient temperature of 25 ℃. All simulations were conducted using the NS-SPICE simulator. To emulate parasitic effects in memory arrays, a 1 pF capacitor was added to the proposed cell and the reproduced cell.

Simulation results at 1 V supply voltage are shown in Fig. 2. When computing a logic '1', the output voltage remains at V_{DD}, whereas for a logic '0', it discharges from a V_{DD} to 0, and thus for a logic '0' computation, a threshold voltage is needed to be defined. In [4], the logic '1' is defined as when the output voltage is 150mV larger than the reference voltage. Following the definition in [4], a logic '0' is interpreted as an output voltage below 250 mV (with a MOSFET threshold voltage V_{th} of 0.4 V), with the primary delay metric determined by the discharge behavior. According to this, the latency and reliability of this work and [4] are mostly influenced by the logic '0' and logic '1' computation respectively.

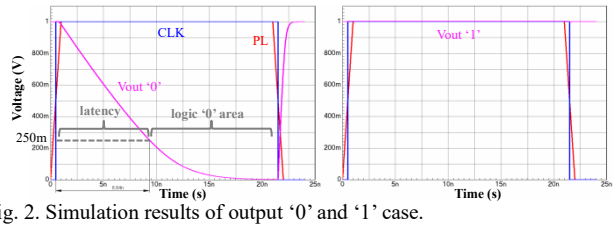


Fig. 2. Simulation results of output '0' and '1' case.

Figure 3 compares the XNOR computation delay of the proposed design and the reference method in [4] based on simulation results. The delay is measured by the time taken for the output voltage to reach the defined threshold, as described earlier. As the supply voltage varies from 1V to 1.2V, the proposed design consistently achieves shorter computation delays compared to the reference, with reductions exceeding 50% across all tested voltages. This demonstrates the high computation speed of the proposed CiM cell.

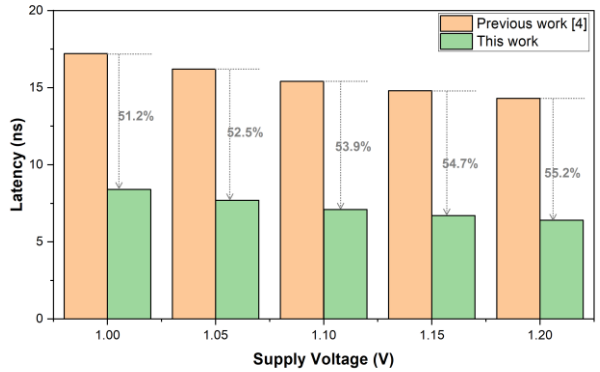


Fig. 3. Computation latency of this work and reference method under different supply voltages.

Figure 4 compares the output voltage variations under different TMR values for the proposed cell and the reference method in [4]. It shows that the proposed cell maintains a nearly constant output voltage as TMR varies from 50% to 500%, while the reference method has a voltage drop by 0.42V when TMR decreases from 500% to 50%. This observation indicates that the proposed structure exhibits enhanced stability against TMR variations. This robustness originates from the fact that the MTJ resistance does not directly affect the XNOR computation process, thereby significantly reducing the circuit's sensitivity to MTJ parameter fluctuations and enhancing its compatibility with a wide range of MTJ devices.

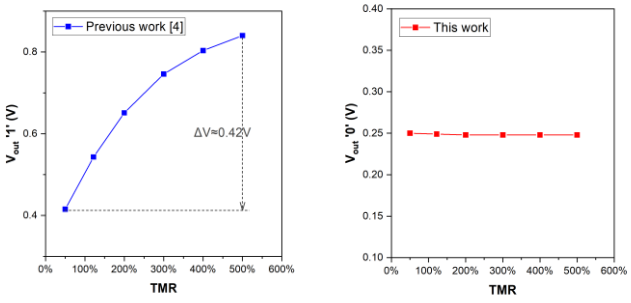


Fig. 4. Output voltages of this work and reference method under different TMR.

Figure 5 compares the computation power consumption of the proposed design and the reference method under supply voltages ranging from 1 V to 1.2 V. It is important to note that although the two designs exhibit different computation delays, the comparison uses a fixed switching duration for both. In the proposed design, the power line (PL) is turned off immediately after computation to eliminate static power consumption.

The comparison is based on the average power consumption across both logic '1' and logic '0' output cases. The results show that the proposed design consistently consumes less power than the reference. At the nominal supply voltage of 1.2V for 55 nm process technology, the proposed design achieves a 26.2% reduction in power consumption compared to the reference method.

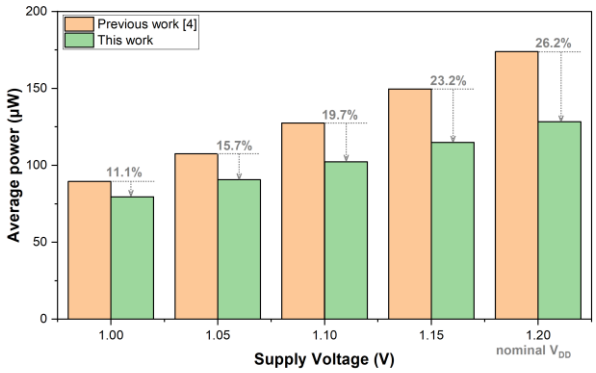


Fig. 5. Power consumption of this work and reference method under different supply voltages.

IV. CONCLUSIONS

A novel STT-MRAM based fully digital CiM architecture has been proposed. Through a state-of-the-art comparison with SPICE simulation, the proposed and designed CiM demonstrates over 50% reduction in computation latency, enhanced tolerance to TMR variation, and 26.2% power reduction at a 1.2V supply voltage. These advantages underscore its suitability for energy-efficient and high-speed and high-reliability BNN accelerators.

V. ACKNOWLEDGES

This work was supported by MEXT Initiative to Establish Next-generation Novel Integrated Circuits Centers (X-NICS) Grant Number JPJ011438 and JST SPRING, Grant Number JPMJSP2114.

REFERENCES

[1] Jung, S., Lee, H., Myung, S. et al. A crossbar array of magnetoresistive memory devices for in-memory computing. Nature 601, 211–216 (2022).

[2] H. Cai et al., "Proposal of Analog In-Memory Computing With Magnified Tunnel Magnetoresistance Ratio and Universal STT-MRAM Cell," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 4, pp. 1519-1531, April 2022.

[3] T. -N. Pham, Q. -K. Trinh, I. -J. Chang and M. Alioto, "STT-BNN: A Novel STT-MRAM In-Memory Computing Macro for Binary Neural Networks," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 12, no. 2, pp. 569-579, June 2022.

[4] T. Na, "Ternary Output Binary Neural Network With Zero-Skipping for MRAM-Based Digital In-Memory Computing," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 7, pp. 2655-2659, July 2023.

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