Speeding-up Successive Read Operations of STT-MRAM via Read Path Alternation for Delay Symmetry

Taehwan Kim and Jongsun park
School of Electrical Engineering
Korea University

***Abstract*— Recent research on data-intensive computing systems has demonstrated that system throughput and latency are critically dependent on memory read bandwidth, highlighting the need for fast memory read operations. Although spin-transfer torque magnetic random-access memory (STT-MRAM) has emerged as a promising alternative to CMOS-based embedded memories, STT-MRAM continues to face challenges related to read speed and energy efficiency. This paper introduces a novel read scheme that enhances read speed and energy in successive read operations by alternating read paths between data and reference cells. This approach effectively mitigates worst-case read scenarios by balancing the read voltage swings. HSPICE simulations using 28nm CMOS technology show a 31.5% improvement in read speed and 48.8% reduction in energy consumption compared to the previous approach. SCALE-Sim system simulations also demonstrate that applying the proposed read scheme to STT-MRAM embedded memories in AI accelerators shows a significant reduction in memory energy for CNN inference tasks compared to the SRAM embedded memory.**

Keywords—STT-MRAM, Energy Efficient, High-Speed Read, Delay Imbalance

I. Introduction

With ever increasing demand for the applications that need large datasets, such as artificial intelligence (AI) and image rendering, circuit designers are facing several design challenges regarding the size and bandwidths of embedded memories within processors. Also, recent studies have shown that the successive read operation constitutes a predominant portion of embedded memory operations. To address the memory footprint issues, spin-transfer torque magnetic random-access memory (STT-MRAM) is a promising alternative. However, the *delay asymmetry* in the read speed, where the read time differs depending on the stored data, brings a disadvantage in overall read speed [1].

In this paper, we propose a high-speed and energy efficient STT-MRAM read scheme that uses alternating read paths to achieve read timing symmetry during successive read operations. By alternating the read paths between data and reference cell, the reference voltage from previous read step is leveraged as the initial voltage for the data path in the following step. This scheme achieves the same benefits as the equalization [2] and mid-point precharge [3] methods, which are proposed to alleviate the *delay asymmetry,* without additional time for pre-charge or equalization.

II. Proposed Work

A. Motivations and the main concept of the proposed work

To address the issue of read delay asymmetry, both the data-reference equalizing scheme [2] and the mid-point precharge scheme [3] involve initializing the data path voltage to a near-reference voltage. However, these techniques need additional equalization time or an external voltage source. Drawing on the fact that the initial voltage is already developed in the reference path during the previous step, the proposed read scheme leverages the pre-developed reference voltage as initial read voltage by alternating the data and reference paths between successive operations. Thus, identical benefits can be achieved in the proposed read scheme while eliminating the overheads. Given the frequent requirement for successive and predictable read patterns in digital signal processing, the proposed read scheme is well-suited for application-specific embedded memories.

B. Detailed circuit schematic & operation flow

Fig. 1 (a) illustrates the configuration of the STT-MRAM macro with the proposed read scheme. Fig. 1 (b) shows the operation flow of the proposed read scheme. First, the S1 signal is set to 0, activating the right switch and the outer NMOS transistors, resulting in a configuration identical to the conventional read scheme [1]. Then, the read voltage (V_{Data}) is developed at the left node (V_L), while in the right read path, whereas the reference voltage is developed at the right node (V_R). In the subsequent second read operation (the right figure of Fig. 1 (b)), S1 toggles to 1, activating the left switch and the inner NMOS transistors, causing the data and reference paths to swap. As a result, the left read path connects to the reference cell, while the right read path connects to the data cell. At this stage, the reference voltage developed in the previous operation is leveraged as the initial voltage of V_R (denoted as V_{Ref} to V_{Data}) the same read voltage swing is achieved regardless of whether the data cell is in the P or AP state. The same benefits are continuously leveraged across multiple successive read operation cycles. Although the proposed scheme provides advantages, its effectiveness is fully realized when memory cells in the same column are accessed successively.

Fig. 2 illustrates the waveform of successive read operations using the proposed read scheme compared to the data-reference equalization scheme [2]. As shown in the upper

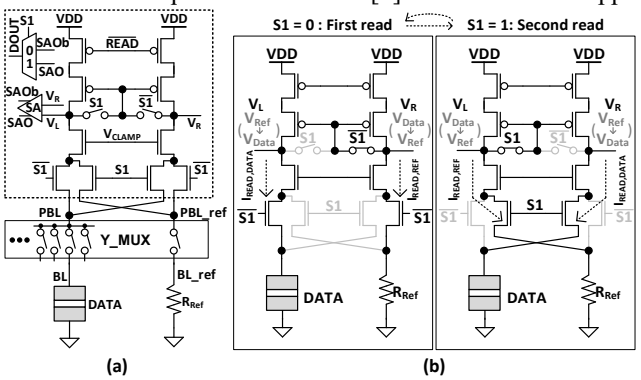

Fig. 1. (a) The schematic of the proposed alternating read-reference paths read scheme (b) Operation flow of proposed read scheme.

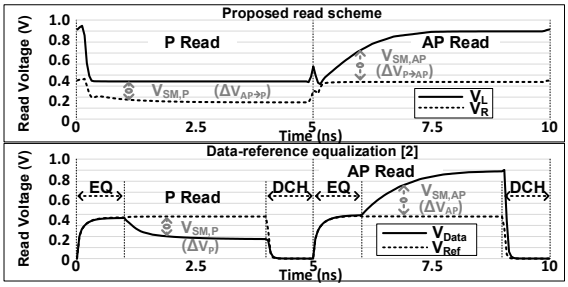

Fig. 2. Waveform of the proposed and the data-ref. equalization [2] scheme in successive read operation.

978-3-9826741-0-0/DATE25/© 2025 EDAA

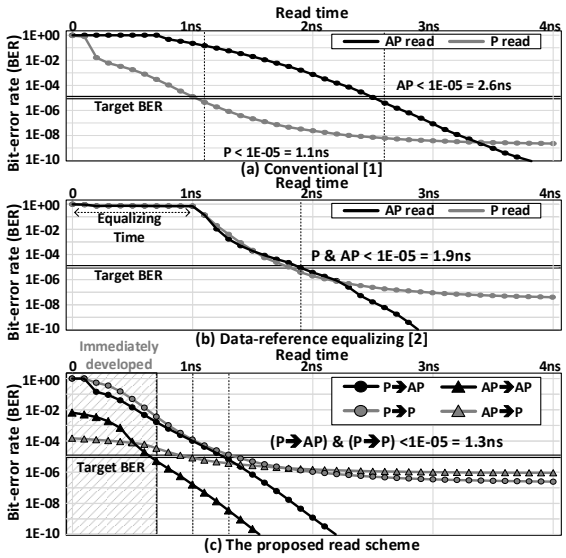

Fig. 3. Comparison of BER in terms of read time on 512x512 array between (a) conventional current-mode [1] (b) data-reference equalizing [2] (c) proposed read scheme.

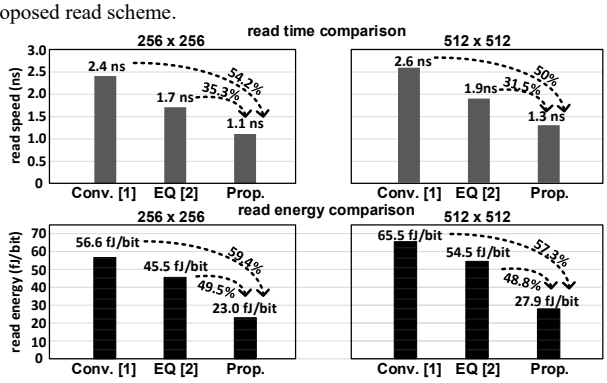

Fig. 4. Comparison of read speed and energy between read schemes.

figure of Fig. 5, in successive read operations, the read voltage in the proposed scheme develops quickly after the operation begins, with only a brief stabilization period required after switching the read paths.

III. Numerical Results

Fig. 3 illustrates the BER comparison of read time between various read schemes in 512x512 STT-MRAM macro, simulated with 2K HSPICE Monte-Carlo simulations with 28nm CMOS technology on 0.9V supply voltage. For the STT device during the Monte-Carlo simulation, a 5% resistance variation at 1-sigma [4] has been assumed. The read BER is calculated using the equations proposed in [4]. While conventional works [1], [2] show slow read time due to read time asymmetry or equalizing time, the proposed read scheme shown in Fig. 3 (c), shows almost immediate read voltage development after the read operation begins, thanks to the initialization at the reference voltage level. So, it achieves balanced read speeds across all cases. Fig. 4 compares read energy and speed across different macro sizes for various read schemes. As shown in the figure, the proposed read scheme demonstrates improved read energy efficiency along with faster read speeds for all array sizes.

To evaluate the impact of the proposed read scheme on the embedded memory of AI processors, NVSim [5] is used to obtain macro-level energy consumption of STT-MRAM and SRAM. For STT-MRAM, sub-array read latency and energy are brought from HSPICE circuit simulations, and the interconnect energy is sourced from NVSim. Utilizing the results obtained from NVSim, memory energy consumption

TABLE I. SCALE-Sim Parameters

Clock frequency: 1GHz	
SCALE-Sim Parameters	**Spec**
Ifmap,filter buffer size	256KB (STT), 128KB (SRAM)
Processing Element (PE) array size	14 (width) x 12 (height)
Dataflow	Output stationary
Network	MobileNet
Memory Parameters	**Spec**
SRAM Read Energy (128KB)	24.3 fJ/bit
SRAM Write Energy (128KB)	10.3 fJ/bit
SRAM Leakage Power (128KB)	9.73 mW
STT-MRAM Read Energy (256KB)	Prop: 63.8 fJ/bit, EQ: 90.4 fJ/bit
STT-MRAM Write Energy (256KB)	0.153 pJ/bit
STT-MRAM Leakage Power (256KB)	2.924 mW

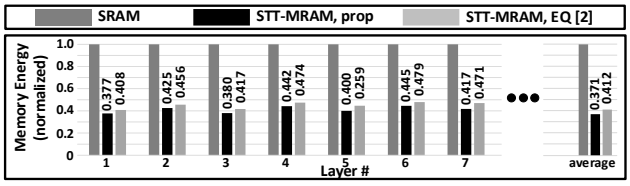

Fig. 5. Layer-wise analysis of embedded memory energy consumption in AI processor using SCALE-Sim [6].

in the AI accelerator is investigated using SCALE-Sim [6]. The detailed configuration parameters for the SCALE-Sim simulations are presented in Table I. Fig. 5 illustrates the layer-wise analysis of embedded memory energy consumption under iso-area conditions when processing MobileNet [7]. Thanks to the reduced leakage power and significant energy reduction in the successive read operation, the proposed read scheme offers 62.9% and 10.1% average energy savings compared to the SRAM and data-reference equalization [2] (STT-MRAM, EQ) read scheme respectively.

IV. Conclusion

This paper introduces a high-speed read scheme for STT-MRAM that addresses the asymmetry issue by alternating read paths and recycling reference voltages to facilitate read voltage development during successive read operations. These findings highlight the potential of applying the STT-MRAM with the high-speed read scheme to enhance both performance and embedded memory energy efficiency in AI accelerator applications.

V. Acknowledgment

This research was supported in part by the National Research Foundation of Korea (NRF) grant (2022M3H4A1A 04096339 and RS-2024-00345481), and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant (RS-2023-00229028) funded by the Ministry of Science and ICT (Information and Communications Technology) (MSIT).

References

[1] D. Halupka *et al.*, "Negative-resistance read and write schemes for STT-MRAM in 0.13µm CMOS," *2010 IEEE International Solid-State Circuits Conference - (ISSCC).*

[2] J. P. Kim *et al.*, "A 45nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance", *2011 Symposium on VLSI Circuits - Digest of Technical Papers.*

[3] Y.-C. Chiu *et al.*, "A 22-nm 1-Mb 1024-b Read Data-Protected STT-MRAM Macro With Near-Memory Shift-and-Rotate Functionality and 42.6-GB/s Read Bandwidth for Security-Aware Mobile Device," *IEEE J. Solid-State Circuits*, *2022.*

[4] J. Kim *et al.,* "A Dual-Domain Dynamic Reference Sensing for Reliable Read Operation in SOT-MRAM," *IEEE Trans. Circuits Syst. Regul. Pap.*, May 2022.

[5] Xiangyu Dong *et al.,* "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, July 2012.

[6] A. Samajdar *et al.,* "A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim," *2020 IEEE ISPASS.*

[7] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv,* Apr. 16, 2017.