An Efficient On-Chip Reference Search and Optimization Algorithms for Variation-Tolerant STT-MRAM Read

Kiho Chung
Department of Semiconductor and Display Engineering
Sungkyunkwan University
Suwon, Republic of Korea
Memory Division, Samsung Electronics
Hwaseong, Republic of Korea
happykeeho@g.skku.edu

Youjin Choi
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
cyjin0227@g.skku.edu

Donguk Seo
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
westduk@skku.edu

Yoonmyung Lee
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
Yoonmyung@skku.edu

***Abstract*—A novel reference search algorithm is proposed in this paper to significantly reduce the reference search time of embedded spin transfer torque magnetic random access memory (STT-MRAM). Unlike conventional methods that sequentially search reference levels with linearly increasing references, the proposed Dual Read Reference Search (DRRS) algorithm requires only two array read operations. By analyzing the statistical characteristics of the read data using a customized function, the optimal reference level can be quickly determined in a few steps. Consequently, the number of read operations required for a reference search is reduced, providing a substantial improvement in the reference search time. The DRRS algorithm can be operated on-chip, and its effectiveness was confirmed through simulations. The optimization speed was improved by 85% compared to the conventional methods. Additionally, an Triple Read Reference Search (TRRS) algorithm is proposed to decrease the variation occurring across different cell arrays and to enhance optimization accuracy. STT-MRAM is composed of numerous cell arrays, where the cell distributions in each array exhibit different characteristics. The TRRS algorithm enhances optimization accuracy for variations occurring in each array, achieving over a 2x increase in accuracy compared to the DRRS algorithm. Furthermore, Simultaneous Reference Search for P and AP (SRS) algorithm that significantly reduces the search time by simultaneously optimizing Parallel (P) and Anti-parallel state (AP) reference cells is also proposed. Lastly, regarding cell degradation after power-up, we enable prompt re-optimization through revolutionary time-saving algorithms (DRRS, TRRS and SRS). This allows for rapid re-optimization in the event of errors caused by cell degradation and ensures regular optimization to maintain maximum read margin even before errors occur, thereby enhancing reliability.**

I. Introduction

Spin transfer torque magnetic random access memory (STT-MRAM) is gaining attention as a high-speed, high-density non-volatile memory. A typical STT-MRAM macro, shown in Figure 1(a), consists of small slices containing cell arrays and sense amplifiers. Each cell stores data in either P (parallel, data 0) or AP (anti-parallel, data 1) states, and the sense amplifier reads the data by comparing currents. However, small resistance differences between P and AP cells, affected by process and temperature variations, make reading challenging, requiring a reference that can track these changes.

Variations in sense amplifiers and references across cells degrade read margins. Cell location and process-voltage-temperature (PVT) variations also impact cell characteristics, such as threshold voltage and current. The conventional approach uses a single reference for all cell arrays, which reduces sensing accuracy and increases read errors.

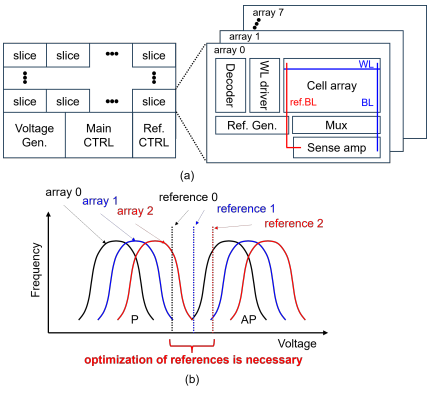

Fig. 1: (a) Block diagram of STT-MRAM (b) Adjusting the reference for each array to increase the read margin

Therefore, adjusting the reference for each array, as shown in Figure 1(b), compensates for 'within-die' and 'chip-to-chip' variations, improving read margins and reducing errors. Tradi-

978-3-9826741-0-0/DATE25/© 2025 EDAA

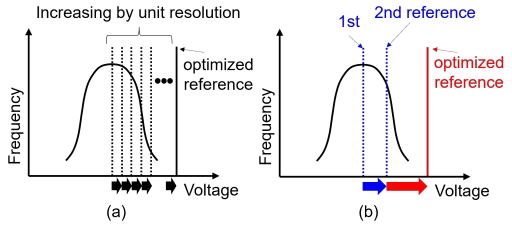

Fig. 2: Concept for DRRS (a) Conventional (sequential linear search) (b) DRRS

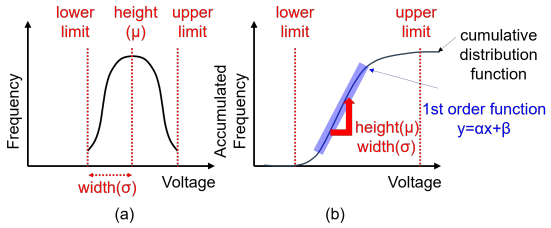

Fig. 3: (a) Four components for compensation (b) Calculation function for compensation

tional methods sequentially adjust the reference, which is time-consuming, especially as STT-MRAM density increases [3].

In this paper, we propose several algorithms to dramatically reduce the essential reference optimization time, thereby compensating for 'within-die variation'. Additionally, the proposed algorithms can be executed on-chip. The conventional reference is fixed by trimming at fab out. And only one optimization was possible after power-up. However, the proposed algorithms allow for the selection and use of the required algorithm at any time , immediately after power up or in other situations.

II. On-chip DRRS Algorithm for Optimized Reference Search

We will first introduce the Dual Read Reference Search (DRRS) algorithm that uses the minimum number of reads to optimize the reference. This is the most fundamental algorithm among those proposed in this paper, and it provides the fastest optimization. Reference optimization refers to adjusting the reference level to the midpoint between P and AP cells to maximize the read margin. To do this, a process of determining pass and fail is required, which typically involves performing a large number of read operations. However, in this paper, we present a method that achieves optimization using only two reads. This method is designed to be executed on-chip to allow compensation of time-varying variations including aging.

A. Concept and Compensation Function for DRRS Algorithm

The key idea to reduce the number of read operations is to optimize the reference level based on two consecutive reads. This process predicts the optimal reference level using a specialized function derived from the cell array's read data. Figure 2 (a) shows the conventional method at VLSI 2023 [3], which sequentially increases the reference level by a fixed unit resolution, necessitating many reads. In contrast, Figure 2 (b) depicts the proposed method, where the 1st read uses the initial reference, and the 2nd read employs a reference with a predefined offset added to the first, requiring only two reads in total. The data from these reads are analyzed to quickly adjust to the optimal reference level. Most MRAM systems read multiple bits (e.g., 32 bits) simultaneously, and the DRRS algorithm needs a substantial amount of read data for statistical analysis. For the applied cell model, 128 bits are read at once, though if this is not feasible, four consecutive 32-bit reads can be performed. The read operation determines the data written in the cell by comparing the voltage levels of the bit lines (BL) connected to the main and reference cells, distinguishing between P and AP data. With a constant current flowing through the BL, a single reference cell appears as a straight line, while multiple main cells create a distribution reflecting MTJ resistance variations. The x-axis indicates voltage based on MTJ resistance variation, and the y-axis represents the number of cells at each voltage.

The function for reference optimization employs statistical analysis to predict the cell resistance/conductance distribution. As illustrated in Figure 3 (a), the distribution of STT-MRAM cells is characterized by four components: lower limit, upper limit, distribution width, and distribution height. These components form the basis for a cumulative distribution function, as shown in Figure 3 (b). The slope in the linear region of this function can be represented as a first-order function. In this paper, we integrate first-order functions with digital logic gates to achieve discrete inputs and outputs. To maintain sufficient linearity, we include an internal register that allows users to adjust the variables (α, β).

$$y = \alpha x + \beta$$

This function calculates the optimal reference level for each cell array in response to process variations by predicting the slope of the cumulative distribution function using read data from two reads. It enables rapid adjustment to the optimal reference level, minimizing read errors. This operation occurs after power-up, with the algorithm and function executed on-chip for convenience.

In this paper, we utilized a commercial MRAM model at 28nm technology and employed ten first-order functions by subdividing the linear region to enhance the accuracy of linear approximation. The type and quantity of functions used vary depending on the cell distribution, and as the functions become more complex, the accuracy of the approximation increases.

B. Flow Chart, State Machine and Block Diagram for DRRS Algorithm

Figure 4 shows the flow chart of the proposed DRRS algorithm, which enables rapid reference optimization with just two reads. In Figure 4(a), the conventional method at VLSI 2023 [3] involves incrementally adjusting the reference level by unit resolution after each read until the data passes. This approach can require many reads if the initial reference level is far from the optimized value.

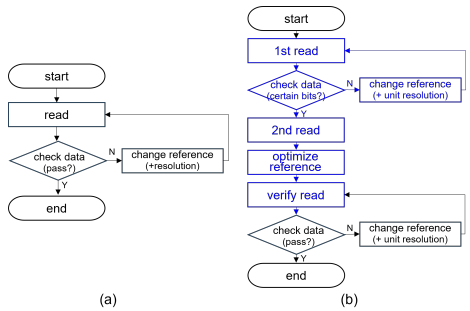

Fig. 4: Flow chart for algorithm (a) Conventional (sequential linear search) (b) DRRS

The DRRS algorithm, as depicted in Figure 4(b), first performs a read, followed by a second read where the reference level is adjusted by a pre-determined delta. To avoid outliers, the first read counts the number of passing cells, and the second read proceeds only if this count exceeds a certain threshold. The results of the two reads are then XORed to identify cells within a specific region of the distribution. This count is used to adjust the reference level to an optimized value without needing further iterations. The optimal reference is thus determined in just two reads. Finally, after the adjustment, a re-read is conducted for verification. If all cells pass, the process ends. If errors remain, a safety mechanism allows fine adjustment of the reference level by unit resolution.

Figure 5 illustrates the state machine for the reference controller circuit used in reference level optimization. In Figure 5 (a), the conventional method at VLSI 2023 [3] repeatedly performs reads until finding the optimal reference level. In contrast, Figure 5 (b) shows a more efficient approach that reduces the number of reads needed by addressing various risk scenarios proactively. The DRRS algorithm can proceed through three cases: Case 1 occurs when all cells pass after the 1st read, indicating that no optimization is needed, and the algorithm terminates. Case 2 represents situations where outliers in the cell distribution lead to incorrect predictions during the 1st read. The algorithm continues only if a predefined minimum number of bits is present in the read results to

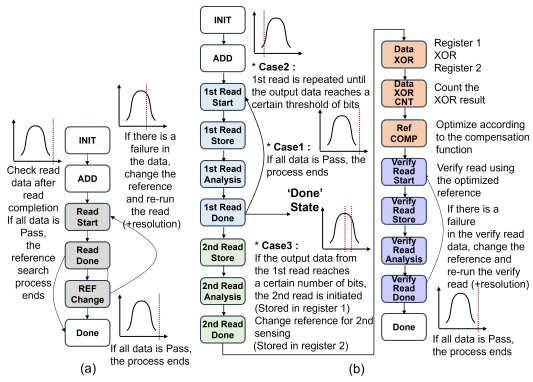

Fig. 5: State machine (a) Conventional (sequential linear search) (b) DRRS

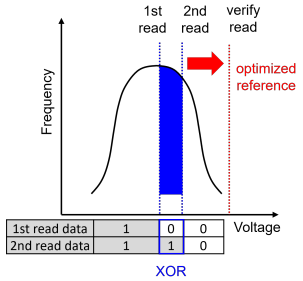

Fig. 6: Computing the number of cells between 1st and 2nd

avoid misprediction. Case 3 arises when the 1st read results in failures, prompting the algorithm to store the read data and perform a 2nd read. During this read, the reference level is adjusted to a predetermined value (1st Reference Level + Delta), and the same read operation is repeated. The number of cells between the two reads is then calculated based on the data from both reads.

The blue area in Figure 6 can be calculated by XOR operation between the 1st read data and the 2nd read data. The calculated result is applied to the compensation function to adjust the reference level to the optimal position. After adjusting the reference level, a verification read is performed, and the process is terminated.

Figure 7 presents a block diagram for operating the proposed DRRS algorithm, featuring additional functionalities in the reference controller module (Ref. CTRL). When the algorithm starts, a read start signal is sent from the reference controller to the read controller to perform the read operation, which is repeated twice. The read data is returned to the reference controller, where two registers store the results of the 1st and 2nd reads. The stored data is processed through an XOR operation in the XOR logic, and the results are counted in the Counter Logic. Calculation logic uses a 1st order function circuit to determine the optimal reference. The optimized reference data is then stored in registers for each array and sent to the sense amplifiers for improved reads. All circuits within the reference controller module are synthesized using commercial 28nm technology, allowing users to adjust compensation accuracy based on cell variation. This compensation accuracy has a trade-off with the area of the reference controller module circuit.

C. Compensated Reference Level for Sensing

The optimized data from the reference controller is applied to the sense amplifier circuit. The sense amplifier used in the read operation employs a conventional current-sensing approach [1], where a constant voltage is applied to the BL and the sense amplifier detects the current to differentiate between P and AP

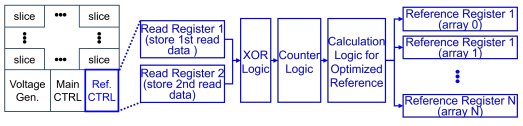

Fig. 7: Block diagram of the reference controller

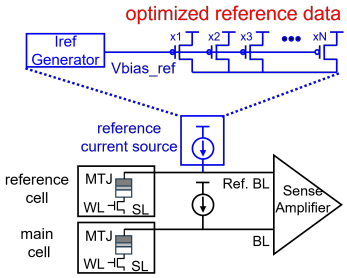

Fig. 8: The reference data is applied to the reference bit line.

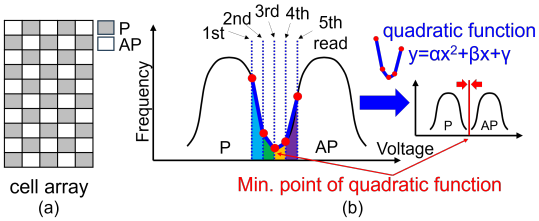

Fig. 10: SRS algorithm (a) Written randomly in the cell array (b) Simultaneously optimizing P and AP

data. To mitigate margin decreases caused by PVT variation, the reference BL connected to the reference cell is compared with the main BL connected to the main cell [8].

In this paper, a circuit is added to the reference cell's BL, as shown in Figure 8, to ensure that optimized data is reflected during sensing operations. Based on the optimized reference data values, the current strength option of the Iref generator is adjusted, altering the current strength. This optimized current is then supplied to both the reference BL and the sense amplifier, allowing the effects of the optimized reference to be realized during sensing.

III. TRRS Algorithm to Improve Cell Variation of Multiple Shapes (Improve Accuracy)

This paper introduces the Triple Read Reference Search (TRRS) algorithm, which employs three reads to enhance optimization accuracy for various cell variation shapes in the cell array. As shown in Figure 9(a), three reads (1st, 2nd, and 3rd) allow for identifying two regions in the cell distribution (blue and red areas). The number of cells in these areas is counted using the XOR operation detailed in Figure 6. The increased data from the TRRS algorithm facilitates more precise analysis compared to the two-read DRRS algorithm, enabling us to determine whether the reference level is positioned to the left or right of the distribution center. Additionally, using the three intersecting points with the distribution, we can calculate a more accurate gradient. The TRRS algorithm allows for more precise calculation of the first-order function, accommodating multiple gradients in the optimization process. As illustrated in Figure 9 (b), significant variations in distribution positions across cell arrays result in differing gradients. A single first-order function applied to multiple gradients may yield inaccurate optimization results.

While more reads provide better analysis data and improve reference level predictions, they also increase optimization time. Thus, the choice of algorithm should balance the required accuracy with the trade-off between optimization time and accuracy.

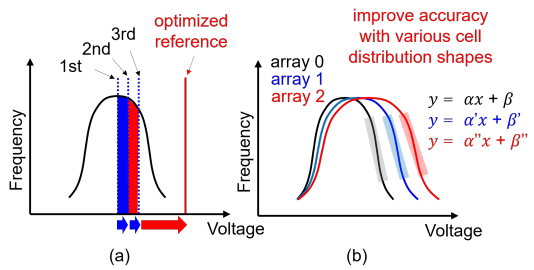

Fig. 9: TRRS algorithm (a) Increasing the number of reads to accommodate various shapes of arrays (b) Applying multiple functions when the slope varies across arrays.

IV. SRS Algorithm to Optimize Different Types of P and AP at the Same Time (Reduce Search Time)

Figure 10 presents an innovative Simultaneous Reference Search for A and AP (SRS) algorithm that enhances search speed by optimizing both P and AP cells simultaneously. In conventional methods, reference optimization requires two separate processes: first, P data is written to all cells for optimization, followed by rewriting the cells as AP for a second optimization. This approach necessitates two reference cells, one for P and one for AP.

In contrast, the SRS algorithm simplifies this by requiring only one optimization process for both cell types. The process involves two main steps: First, all cells are randomly written with a 50% P and 50% AP ratio, as shown in Figure 10 (a). Next, five reads (1st/2nd/3rd/4th/5th read) are conducted to gather data for four areas (blue, green, yellow, purple) and identify five distribution points, as depicted in Figure 10 (b). Using XOR operations, the area data is derived from the five reads. A quadratic function is then formulated to predict the minimum point, with the midpoint serving as the optimized reference. This quadratic function can be implemented with the same digital logic as a linear function. Moreover, since the SRS algorithm requires only one optimization process, it also needs just one reference cell, resulting in reduced area requirements.

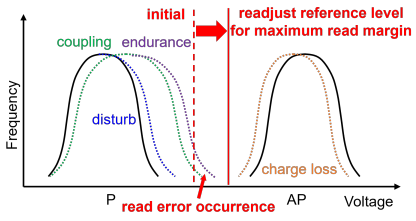

Fig. 11: Cell distribution degradation factors

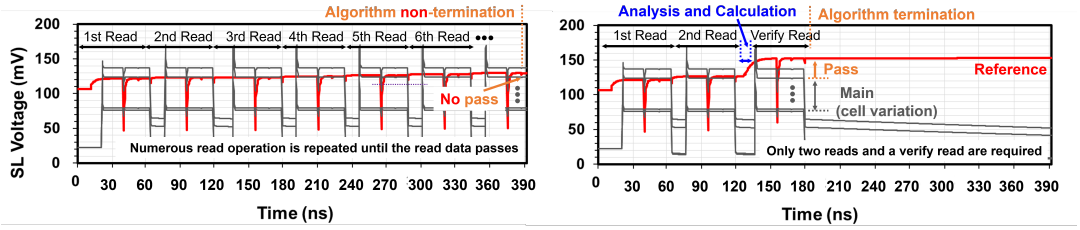

Fig. 12: The simulation results (a) Conventional (sequential linear search) (b) DRRS

V. Enhancing Reliability Using the Algorithms

In scenarios where cell distribution deteriorates after power-up, re-performing reference optimization can enhance the read margin. The proposed algorithms address various external degradation factors occurring post power-up, with each chip performing reference optimization immediately after activation. However, read errors may arise from these degradation factors. For example, the endurance failure rate increases with write pulse time after 1M write cycles [6], and external magnetic fields can elevate the bit error rate [2] [4]. Read disturb errors exceed 1 ppm after 100K read cycles [5], and errors can increase following 260°C reflow soldering [7]. Degradation factors such as cell coupling, disturb, endurance, retention, and charge loss can emerge after power-up, as illustrated in Figure 11, which predicts error scenarios due to external influences. If cell distribution changes, the read margin decreases, but re-performing reference optimization can help restore it.

Conventional algorithms require significant time for reference search, making it difficult to re-optimize after power-up. In contrast, the proposed algorithm focuses on reducing reference search time, enabling re-optimization after power-up. After a read error, reference optimization can be swiftly re-performed, such as executing DRRS or TRRS to improve outcomes. Additionally, regularly operating DRRS to maintain references at the midpoint of P and AP distributions can prevent errors before they occur. By utilizing these algorithms, it's possible to proactively prevent errors and quickly fix any that arise, strengthening the reliability of STT-MRAM and enabling rapid responses to degradation.

VI. Evaluation Results

The proposed algorithms were tested using 100 Monte Carlo simulations on a 128x128 array with a 1T-1MTJ STT-MRAM structure, based on the variation model of commercial MRAM in 28nm technology. The reference controller operated at 1Gbps with a voltage of 1.0V, under process variations of TT/FF/SS and temperatures of -40/25/125°C. Figure 12 illustrates the simulation results, showing that the DRRS algorithm rapidly optimizes the reference level compared to conventional methods at VLSI 2023 [3]. In the conventional approach, as shown in Figure 12 (a), numerous read operations are performed to gradually increase the reference level until a read pass is achieved. The "Main Level" reflects the source line voltage of the main cell, which varies among the 128 MTJ cells. The blue "Reference Level" must be set higher than the main cell's voltage to account for this variation. In contrast, the proposed method, illustrated in Figure 12 (b), requires only two reads to analyze and calculate the optimal reference level. After the 1st and 2nd reads, the calculated value is used to adjust to the optimal reference level, followed by a verify read to reassess it. Thus, only three reads are needed, significantly reducing the number of reads for reference searching compared to traditional methods. The simulation was conducted using a 28nm FD-SOI (STT-MRAM) cell, configured in a 128x128 array, for reference search read. The operating temperature was set to 25°C, and the operating voltage was 1.0V.

Figure 13 presents the simulation results for the iterative optimization of reference levels while randomly changing the positions of cells in the cell array. At TT and 25°C conditions, 100 Monte Carlo simulations were conducted on a 128x128 array, performing DRRS algorithm. The horizontal axis represents the number of reads used for reference level optimization, while the vertical axis indicates the probability of the number of reads used for optimization depending on the position of the cell array. When applying the DRRS algorithm (red) compared to the conventional method (black), there is an average improvement of 85% (from 20.42 to 3.06) in the number of reads required for reference level optimization. This substantial reduction suggests a significant decrease in reference search time. Additionally, the standard deviation is improved by 94% (from 4.26 to 0.28), indicating a more uniform improvement in reference search time across cell arrays. The conventional method was simulated based on VLSI 2023 algorithm [3].

Figure 14 illustrates the robustness of the DRRS algorithm under varying PVT conditions. Across nine simulation scenarios (TT/FF/SS at -40/25/125°C), 100 Monte Carlo simulations were performed, demonstrating that the average number of reads remained around three when utilizing the DRRS algorithm. The algorithm optimizes the reference with two initial reads, followed by a verification read. If the linear

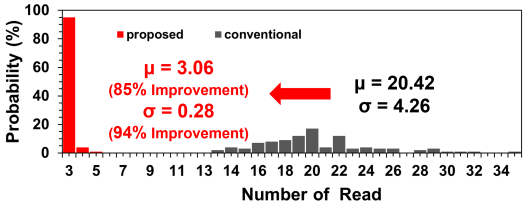

Fig. 13: Improvement in reference search time from DRRS

TABLE I: Comparison between proposed and conventional algorithm

		This work		Conventional Algorithm1 (linear search) [3]	Conventional Algorithm2 (linear search) [1]
		DRRS Algorithm	TRRS Algorithm		
Technology [nm]		28	28	28	28
Cell type		1T MTJ	1T MTJ	1T MTJ	1T MTJ
Sensing scheme		Current	Current	Current	Current
Reference search algorithm		DRRS	TRRS	Sequentially linear	Sequentially linear
On-chip reference search		O	O	O	X
Available points for reference search	After fab-out	O	O	O	O
	After power-up	O	O	O	X
	At any time	O	O	X	X
Average number of reads (reference search time)		3.06	4.02	20.42	20.42 + Data out
Standard deviation of the number of reads		0.28	0.14	4.26	4.26

approximation for reference optimization is successful, the process concludes with three reads; otherwise, additional reads are performed. As shown in Figure 14, the average number of reads remains consistent under all conditions: 3.17 at -40°C and 3.09 at 125°C (TT), and 3.07 (FF) and 3.06 (SS) at 25°C.

The TRRS algorithm, which analyzes results from three reads to optimize the reference, demonstrates enhanced accuracy compared to the DRRS algorithm, which uses two reads. Under TT and 25°C conditions, 100 Monte Carlo simulations were performed on a 128x128 array using both algorithms. Figure 15 illustrates the results of the TRRS algorithm in comparison to DRRS. In Figure 15 (a), the average number of reads for the TRRS algorithm increases by 31% (from 3.06 to 4.02), resulting in longer optimization times. However, Figure 15 (b) shows that the standard deviation improves by 50% (from 0.28 to 0.14). Since speed optimization and accuracy are trade-offs, it is important to consider the distribution characteristics of the cell array and apply the most suitable algorithm accordingly.

Table 1 compares the search time and accuracy of the DRRS and TRRS algorithms against conventional algorithms. Conventional algorithm 1 was presented at VLSI 2023 [3], and conventional algorithm 2 was introduced at ISSCC 2018 [1], both employ a sequential linear search algorithm for reference searching. These algorithms were simulated under identical conditions using the 28nm FD-SOI (STT-MRAM) cell and array architecture described in this paper. In contrast to conventional algorithms, both DRRS and TRRS algorithms implement a statistical intelligence search concept that minimizes the number of reads. The DRRS algorithm demonstrates a remarkably faster reference search time, while the TRRS algorithm achieves significant enhancements in accuracy. Moreover, both DRRS and TRRS can be re-executed on-chip at any time after fab-out or power-up, allowing for the improvement of read margin reduction caused by reliability degradation.

VII. Conclusion

Cell variation depends on array position, so individually adjusting reference levels can enhance sensing margins and reduce read errors. However, previous methods required significant time for reference level adjustment after power-up.

This paper proposes on-chip algorithms to optimize the time needed for reference level adjustment and improve usability. The DRRS algorithm performs minimal reads to analyze cell data and predict distribution, allowing for immediate adjustment of the reference level to the optimal position based on predictions. This approach reduces optimization time by 85% compared to conventional methods and improves array-to-array variation by 94%, enhancing uniformity.

Additionally, the TRRS algorithm (utilizing three reads) enhances accuracy across varied cell distributions, with further improvements expected as read counts increase. The trade-off between read count and search time is acknowledged.

A novel SRS algorithm is also introduced, which simultaneously optimizes P and AP reference cells, drastically reducing search time and the number of required reference cells, leading to area improvements.

Lastly, methods for applying reliability improvements during operation are proposed. Reference-level optimization is performed only once immediately after power-up. However, with the significant reduction in optimization time, it is now possible to execute multiple optimizations for various variations anytime after power-up. Therefore, applying these approaches to various cell degradation scenarios after power-up is expected to help improve reliability degradation during operation.

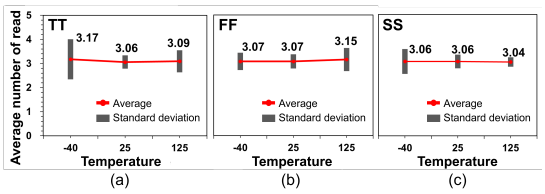

Fig. 14: Simulation results with PVT variation applied in the DRRS (a) TT (b) FF (c) SS

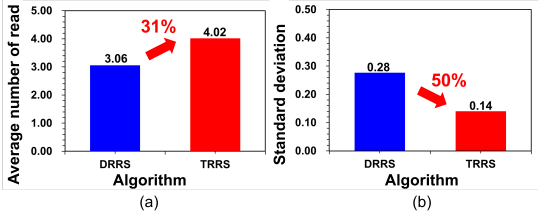

Fig. 15: The simulation results of accuracy improvement by the TRRS (a) Average number of reads (b) Standard deviation

REFERENCES

[1] DONG, Q., WANG, Z., LIM, J., ZHANG, Y., SHIH, Y.-C., CHIH, Y.-D., CHANG, J., BLAAUW, D., AND SYLVESTER, D. A 1mb 28nm stt-mram with 2.8 ns read access time at 1.2 v vdd using single-cap offset-cancelled sense amplifier and in-situ self-write-termination. In *2018 IEEE International Solid-State Circuits Conference-(ISSCC)* (2018), IEEE, pp. 480–482.

[2] HAN, S., LEE, J., SUH, K., NAM, K., JEONG, D., OH, S., HWANG, S., JI, Y., LEE, K., SONG, Y., ET AL. Reliability of stt-mram for various embedded applications. In *2021 IEEE International Reliability Physics Symposium (IRPS)* (2021), IEEE, pp. 1–5.

[3] KANG, G., SHIN, H., JUNG, H., LEE, S., CHOI, J., BAEK, S., JUNG, H., KIM, D., HWANG, S., HAN, S., ET AL. A 14nm 128mb embedded mram macro achieved the best figure-of-merit with 80mhz read operation and 18.1 mb/mm^2 implementation at 0.64 v. In *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (2023), IEEE, pp. 1–2.

[4] LEE, T. Y., YAMANE, K., HAU, L. Y., CHAO, R., CHUNG, N. L., NAIK, V. B., SIVABALAN, K., KWON, J., LIM, J. H., NEO, W.-P., ET AL. Magnetic immunity guideline for embedded mram reliability to realize mass production. In *2020 IEEE International Reliability Physics Symposium (IRPS)* (2020), IEEE, pp. 1–4.

[5] NAIK, V., YAMANE, K., LEE, T., KWON, J., CHAO, R., LIM, J., CHUNG, N., BEHIN-AEIN, B., HAU, L., ZENG, D., ET AL. Jedec-qualified highly reliable 22nm fd-soi embedded mram for low-power industrial-grade, and extended performance towards automotive-grade-1 applications. In *2020 IEEE International Electron Devices Meeting (IEDM)* (2020), IEEE, pp. 11–3.

[6] NAIK, V. B., LIM, J., YAMANE, K., ZENG, D., YANG, H., THIYAGARAJAH, N., KWON, J., CHUNG, N., CHAO, R., LING, T., ET AL. Superior endurance performance of 22-nm embedded mram technology. In *2019 IEEE International Reliability Physics Symposium (IRPS)* (2019), IEEE, pp. 1–4.

[7] SHIH, M.-C., WANG, C.-Y., LEE, Y.-H., WANG, W., THOMAS, L., LIU, H., ZHU, J., LEE, Y.-J., JAN, G., WANG, Y.-J., ET AL. Reliability study of perpendicular stt-mram as emerging embedded memory qualified for reflow soldering at 260° c. In *2016 IEEE Symposium on VLSI Technology* (2016), IEEE, pp. 1–2.

[8] TRINH, Q.-K., RUOCCO, S., AND ALIOTO, M. Dynamic reference voltage sensing scheme for read margin improvement in stt-mrams. *IEEE Transactions on Circuits and Systems I: Regular Papers 65*, 4 (2017), 1269–1278.