Timing Optimization Model and PVT Tracked Scheme for STT-MRAM Voltage-Mode Sense

Yongliang Zhou, *Member, IEEE*, Xiao Lin, Zixuan Zhou, Yingxue Sun, Yiming Wei, Zhen Yang, Chengxing Dai, JingXue Zhong, Xiulong Wu, *Member, IEEE*, and Chunyu Peng, *Member, IEEE*

***Abstract*— The impact of process variations on the read operation of low-voltage STT-MRAM becomes severe, posing a challenge in determining the optimal sensing timing of the sense amplifier. This study investigates techniques for refining the timing scheme of sensing circuits in order to improve the sensing reliability of the STT-MRAM. The supply voltage V_{DD}, the Tunneling Magnetoresistance Ratio TMR, the low resistance state of bit-cell R_P, and the parasitic capacitance of bit-line C_{BL} are analyzed along with the voltage sense amplifier (VSA) involved in sensing yield. We develop a timing model through theoretical analysis to determine the optimal VSA enable signal (SAE). In addition, an innovative Process-Voltage-Temperature (PVT) tracking scheme is proposed that can track the optimal VSA enable signal (SAE) and suppress timing variations. Monte-Carol simulation in the 28nm CMOS and magnetic tunnel junction (MTJ) process confirms that the combined scheme significantly enhances the robustness of sensing operation. The proposed scheme improves yield by 20% to 35%, reduces power consumption by 43% to 63%, and reduces read access delay by 47% to 59% compared to conventional sensing schemes at 0.6V supply voltage.**

***Index Terms*— STT-MRAM, sensing yield optimization, analytical model, voltage-mode sense amplifier.**

I. Introduction

With the scaling down of the device dimensions and supply voltages under CMOS technology, static power consumption has increased significantly and becomes a bottleneck in the development of Static Random Access Memory (SRAM). Spin-transfer torque magnetic random access memory (STT-MRAM) has emerged as a promising alternative owing its unique attributes such as zero standby power consumption and greater scalability [1], [2], [3], [4], [5]. However, the unreliability of read operations under low voltage is a non-negligible issue of STT-MRAM, as process variations are exacerbated when power supply voltage is reduced. This poses challenges in the design of STT-MRAM sensing schemes, especially for high-data reliability and high-energy efficiency application scenarios [6], [7].

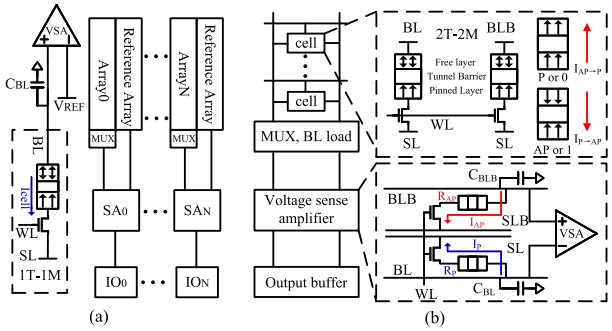

Fig. 1. STT-MRAM with (a)1T+1MTJ bit-cell (b)2T+2MTJ bit-cell.

A widely used STT-MRAM architecture is displayed in Fig.1, the structure of 1T+1MTJ bit-cell is shown in Fig.1(a) and 2T+2MTJ bit-cell is shown in Fig.1(b). The 1T+1MTJ bit-cell employs a traditional data unit reference comparison reading scheme, with a non-differential (single-ended) reading path. When reading data, an additional reference source is required. The advantage of the 1T+1MTJ is that its smaller unit area allows for higher array density. In the 2T+2MTJ memory, two MTJs exhibit opposite resistance states and their readout path is a differential (dual-ended) structure. The sampling of the dual-ended structure employs a local reference, enabling the sampling margin to be doubled in comparison to the 1T+1MTJ structure. However, this comes at the cost of increased area. In order to discuss the optimal sampling time more clearly, we adopt the 2T+2MTJ bit-cell for more intuitive signal analysis.

The 2T+2MTJ structure consists of two access transistors and two magnetic tunnel junctions (MTJ). The tunnel layer of MTJ is sandwiched between two ferromagnetic layers, one is called the pinned layer (PL), which has a fixed magnetization, and the other layer is the free layer (FL), with the two relative magnetization directions to represent the state of the MTJ. During the write operation, if the current flowing from the source line (SL) to the bit-line (BL) exceeds the threshold

Manuscript received 31 March 2024; revised 20 June 2024; accepted 6 July 2024. Date of publication 22 July 2024; date of current version 29 August 2024. This work was supported in part by Anhui Provincial Natural Science Foundation under Grant 2308085QF214, in part by the Natural Science Foundation of the Higher Education Institutions of Anhui Province under Grant 2023AH040011, and in part by the National Natural Science Foundation of China under Grant 62274001. This article was recommended by Associate Editor L. Fick. *(Corresponding author: Chunyu Peng.)*

Yongliang Zhou is with the School of Integrated Circuits, Anhui University, Hefei 230601, China, also with Anhui High-Performance Integrated Circuit Engineering Research Center, Hefei 230601, China, and also with Anhui Anxin Electronic Technology Company Ltd., Chizhou 247000, China.

Xiao Lin, Zixuan Zhou, Yingxue Sun, Yiming Wei, Zhen Yang, Chengxing Dai, JingXue Zhong, Xiulong Wu, and Chunyu Peng are with the School of Integrated Circuits, Anhui University, Hefei 230601, China, and also with Anhui High-Performance Integrated Circuit Engineering Research Center, Hefei 230601, China (e-mail: cyupeng@ahu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2024.3425935.

Digital Object Identifier 10.1109/TCSI.2024.3425935

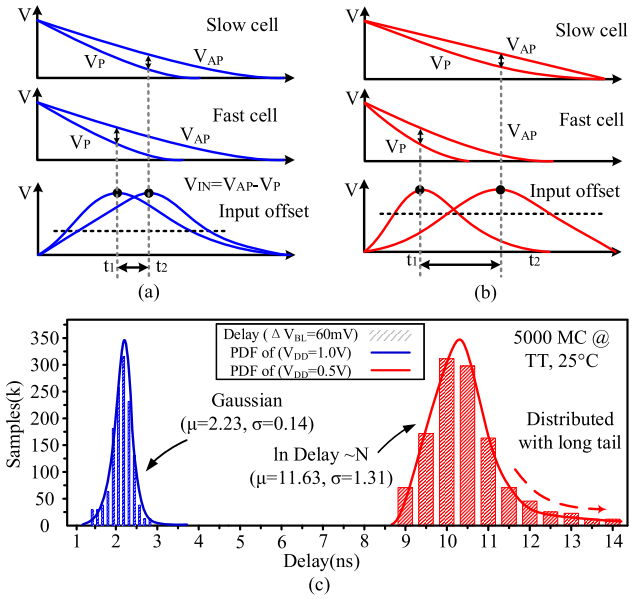

Fig. 2. The bit-lines swing during the read operation at (a) standard supply voltage and (b) low supply voltage. (c) The distribution of bit-lines delay under process variation when the ΔV_{BL} develops to $60mV$ at different supply voltages.

switch current (I_{th}) and flows from the source line (SL) to the bit-line (BL), the data stored in MTJ is written as '0'. Conversely, currents with opposite directions can be written as '1'.

VSA is implemented with the advantage of strong positive feedback and high input resistance. The voltage sensing process can be divided into three phases. The first phase is the pre-charged phase, where the bit-line load generates a direct current voltage by pre-charging the additional clamp transistors. In the discharge stage, current (I_{cell}) is drawn into the selected cell, discharging the capacitance associated with the bit-line (BL) and causing a voltage swing across the bit-line. The voltage swing in the RH state is smaller than that in the RL state. The complementary bit-line (BLB) will also undergo the same discharge process. Finally, in the sensing stage, the voltage sense amplifier (VSA) detects the voltage difference (V_{IN}) between BL and BLB.

Fig.2 shows the schematic diagram of the bit-lines swing and the input voltage of the sense amplifier during the STT-MRAM read operation. The mismatch in driving capability between memory cells and bit-lines results in the differentiation in discharge speeds of the cells, which subsequently defines the classification of cells as fast or slow. As shown in Fig.2 (a), at the standard supply voltage, the maximum value of the input voltage of the sense amplifier occurs at relatively concentrated times and the common activated timing of the sense amplifier can be delayed to take into account the slowest cell.

However, under the low voltage, as shown in Fig.2 (b), the worsening impact of random variations lead to an intensification of the mismatch of transistors and MTJ, and the occurrence time of the maximum V_{IN} for different cells shows a wider distribution. More specifically, Fig.2 (c) presents distribution of bit-lines delay when the ΔV_{BL} develops to 60mV at different supply voltages. The delay of bit-lines follows a Gaussian distribution at 1.0V. In contrast, at 0.5V supply voltage, the delay with a lognormal distribution due to a few simulated samples are above 14ns, forming a long tail for the distribution. Therefore the sensing margin of VSA is subject to degradation at low voltage with the applying of conventional sense schemes, resulting in an increasing sensing failure rate.

Previous research has focused on optimizing the variation tolerance of the sensing circuit and sense amplifier to improve the reliability of read operations for STT-MRAM. Novel sensing schemes introduced in [8] and [9] that provide wide voltage amplitudes for the sense amplifier before the sensing operation. References [7] and [10] have aimed to propose different sense amplifier designs to cancel the offset voltage. Our previous work have proposed MTJ-Based Loop Replica Bit-line technique [11], which improved performance degradation caused by timing in MRAM,and a Self-Timed Voltage-Mode sensing scheme [12], which achieved a significant increase in sensing output under low voltage.

As an alternative strategy, modeling the operation mechanisms can effectively guide the design of optimal embedded memories without the costs of area and power consumption. A previous work proposed an analytical hierarchical yield model [13] that optimized the design trade-offs of Dynamic Random Access Memory (DRAM) circuits. Additionally, Patel et al. [14] developed statistical models that represented the error characteristics of DRAM devices to help designers address scaling challenges. In another study, Singh et al. [15] conducted theoretical modeling to quantitatively predict the array behavior of SRAM, identifying designs that are more power-efficient and have higher yield. Nonetheless, there have been no attempts to model the timing characteristic of STT-MRAM during the read period to guide the design of low-voltage STT-MRAM with high stability.

This paper presents a methodology for optimizing the sensing timing for voltage sensing of STT-MRAM design to improve the sensing yield and minimize the negative impact of process variation. The proposed analytical model provide the guide to determine the active timing of the voltage sense amplifier (VSA) enable signal (SAE). Additionally, a sequential control technique is proposed for suppressing the timing variation of SAE to enhance the robustness of the reading operation. The primary contributions of this study are as follows:

- It is the first exploration of timing optimization methodology for the STT-MRAM voltage sensing scheme design, aiming to improve the sensing yield of read operations.
- A PVT tracking replica bit-line scheme is proposed based on the analytical model to suppress the variation of the SAE timing, enhancing the robustness of timing circuit.
- The results of the simulation demonstrate that the proposed analytical model and timing scheme achieve a significant improvement in performance and sensing yield for voltage sensing of STT-MRAM.

The remainder of this paper is organized as follows. Section II discusses the process of voltage sensing operation. Section III analyzes the main factors that impact the reliability

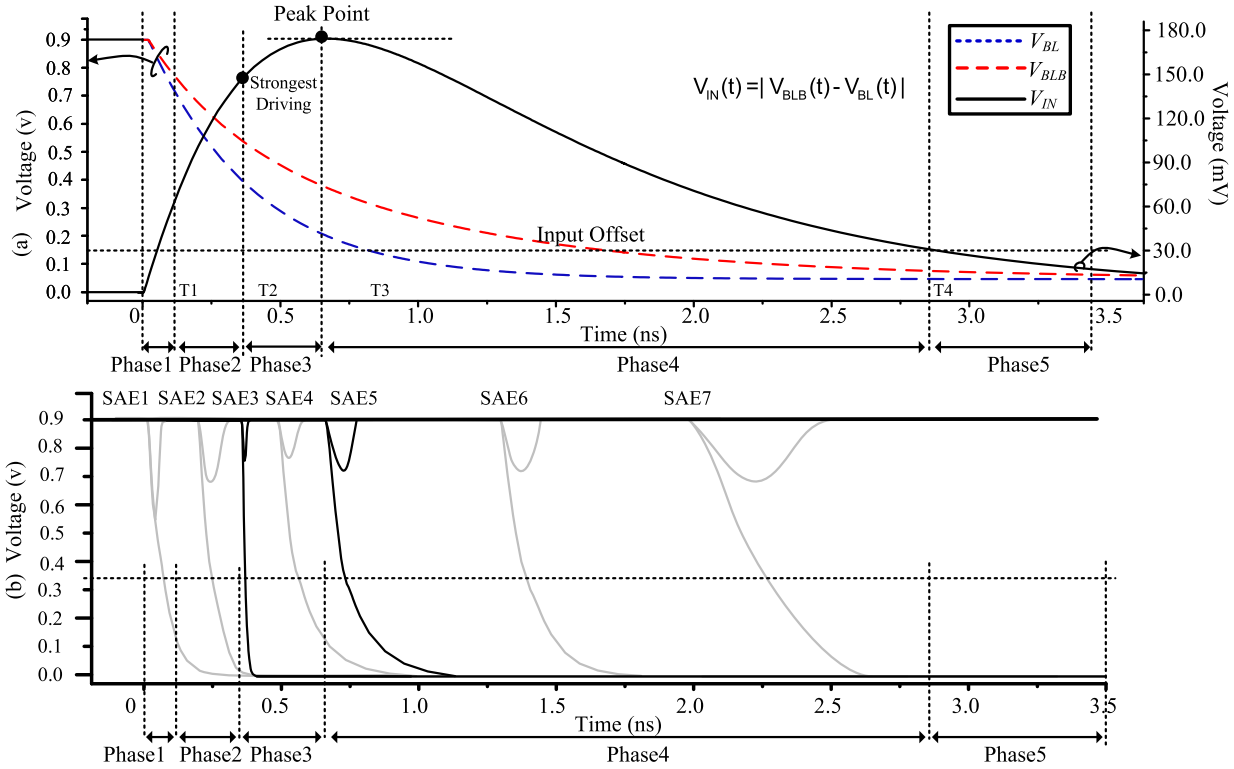

Fig. 3. (a) Simulated transient behavior of the input voltage difference in access period. (b) Simulated transient behavior of the latch-type sense amplifier at different enable timing.

of voltage sensing circuits. Section IV describes the proposed model and the expression of the optimal activated timing of the SAE. Section V presents the PVT tracking scheme for sequential control circuit. Section VI discusses the simulation result and comparisons, and finally, conclusions are presented in Section VII.

II. Temporal Behavior Of Sensing Operation

The transient process of the signal path defines the temporal behavior of voltage sensing operations in STT-MRAM. The process variation and parasitic parameters on the critical path of the read operation can result in instability in the read operation timing of voltage sensitive amplifiers. Before discussing the temporal behavior of a sequential circuit, the detailed analysis of the sensing process is presented in this section, which is based on the coupling latch-type voltage sense amplifier (CL-VSA).

Wicht et al. [16] proposed a quantitative yield analysis of CL-VSA based on the constant differential input, which suggests that the sensing yield increases with a higher input voltage difference. For the SRAM reading process, the input voltage difference gradually increases to the peak value and remains constant. Therefore, enabling the sensing operation later leads to a higher yield. However, the situation is entirely distinct for STT-MRAM [17], [18].

Fig.3 (a) depicts the transient behavior of the bit-lines. It is assumed that the pre-charged BL and BLB storing '0' and '1', respectively, are connected to the internal nodes of the cell. Upon reading access, the bit-line voltage V_{BL} and V_{BLB} are discharged through the corresponding access transistors. The differing resistance of the MTJ connected to the bit-lines results in varying discharging speeds of the bit-lines. The intersection time between the offset voltage and the input voltage during the rising period is represented by T1, while T2 indicates the point of time when the input voltage difference reaches the strongest driving point. The time of the peak input voltage is denoted as T3, and T4 represents the intersection time between the offset voltage and the input voltage during the falling of the input voltage.

Fig.3 (b) illustrates the flipping behavior of CL-VSA at different enable timings. In the application as the voltage-mode sense amplifier of MRAM, the input dc voltage can become nearly any value between ground and V_{DD}. The process of BL discharge can be divided into five phases. In Phase 1, the input voltage is lower than the offset voltage of the sense amplifier. When the amplifier is enabled during this phase, the sensing result mainly relies on the offset voltage. If the offset voltage and the input voltage swing to the same direction, the sensing result is correct; and if they swing to the different direction, the output result is incorrect.

In Phase 2, V_{IN} is higher than the input offset, and the outputs of SA are dictated by the input voltage. As the sensing time moves closer to the strongest driving time T2, the flipping of SA outputs becomes more robust. As shown in Fig. 3(b), the transient behavior of SAE2 is slower than that of SAE3.

Phase 3 is when the strongest driving time T2 transitions to the maximum V_{IN} time T3. The simulation results show that the flipping behavior enabled by SAE3 is stronger than that by SAE5. In other words, the strongest driving point of input is different from the maximum V_{IN} point. In Phase 4, the driving ability of the input weakens, and the flipping time of outputs gradually lengthens. Finally, it steps into

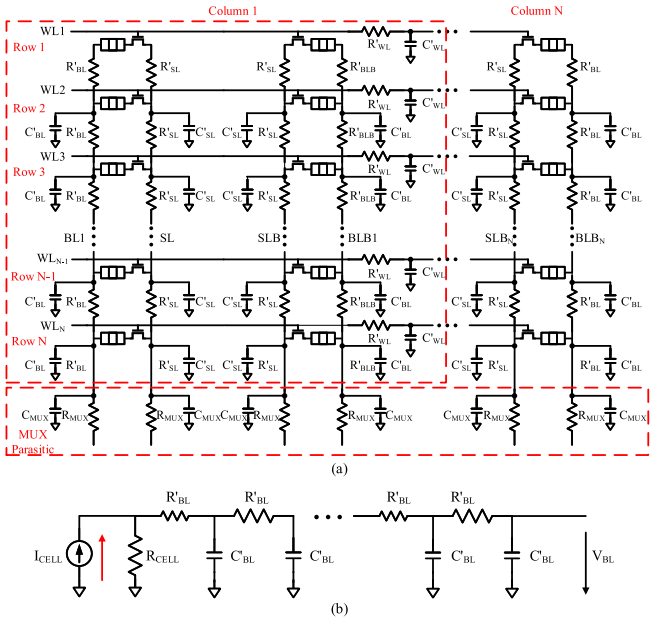

Fig. 4. (a) Equivalent circuit of array including memory cell, dis-tributed bit-line model and MUX. (b) Equivalent circuit of one-column bit-line.

Phase 5, where the input voltage is lower than the offset voltage.

Therefore, it is impossible to intuitively judge that the optimal sensing occurs at the input voltage peak. To rigorously analyze the optimum enable timing of sense amplifier in the MRAM read process, a combination of theoretical derivations and experimental observations is necessary.

III. Analysis of Timing and Yield of Sensing Process

For the voltage-mode sensing scheme, the instability of sensing process arises mainly from the variations in process and parasitic parameters along the critical path of the read operation. This section highlights significant parameters such as parasitics, supply voltage and MTJ device parameter on the read path. It also provides statistical analysis of the impact of these parametric variations on sensing yield and timing of STT-MRAM. The statistics are obtained from the Monte-Carol simulations under 28nm CMOS & MTJ technology. Notice that, applied to a certain number of samples, this effect is referred to as parametric yield Y:

$$Y = \frac{\text{number of correct bits}}{\text{number of readout bits}} \times 100\% \quad (1)$$

which represents the percentage of correct decisions during the sensing process.

A. Equivalent Circuit and Parasitic of STT-MRAM Array

The presence of fabrication nonidealities and parasitics along the critical path can lead to variations in the total capacitance and resistance received at the read circuitry. During the sensing process, the total capacitance and resistance affecting the VSAs is the serial sum of the capacitance and resistance of the sensing access path, as well as the capacitance and resistance of the device. The former comprises the access transistor,

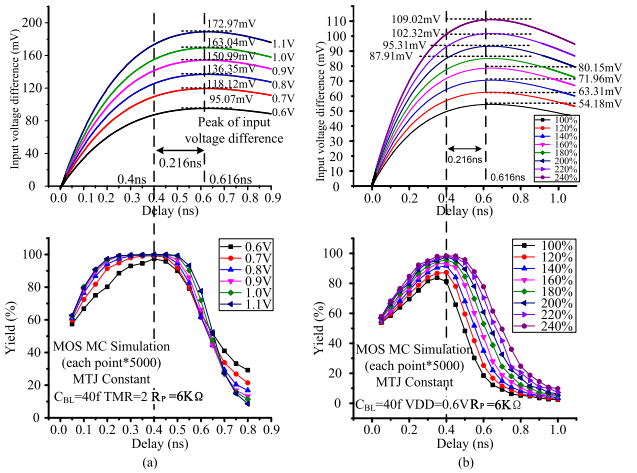

Fig. 5. Simulated yield of the sense amplifier versus supply voltage and TMR among 0.05ns step delay. The yield is defined as achievement whether the falling output reaches $V_{DD}/2$ in 15ns sensing period. (a) supply voltage. (b) TMR.

muxes, and any parasitics along the bit-line (BL) and source line (SL) [19], [20], [21] as shown in Fig.4 (a). As the structure of the 2T+2MTJ bit-cell is completely symmetrical, analyzing the discharging process of one side bit-line is sufficient.

Fig.4 (b) shows the equivalent circuit of one column bit-line, which is characterized by resistances and parasitic capacitances. As the bit-line voltage decreases, I_{cell} changes dynamically, making the input voltage delivered to the SA dependent on the parasitic path. Furthermore, fabrication gradients cause the optimal read conditions to differ for each cell. The capacitance and resistance loaded on the sensing path are determined by summing the parasitics of the bit-line, transistor, and MTJ device. In the remainder of the paper, we denote the effective capacitance and resistance as the sum of the parasitics on the critical path. The sensing delay can be calculated from a time constant τ, where τ=RC, R and C represent the total resistance and total parasitic capacitance on the critical path respectively.

B. Impact of Supply Voltage and TMR

Fig.5(a) illustrates the sensing yield as a function of sensing delay for supply voltage between 0.6V and 1.1V. The distribution of the sensing yield is shown in the bottom diagram of Fig.5(a), whereas the input voltage difference V_{IN} is plotted in the top diagram. The supply voltage has a clear impact on V_{IN} as demonstrated by the top diagram, indicating that V_{IN} increases as V_{DD} becomes larger. With an increase in V_{DD}, the overall sensing yield distribution improves, especially in the high voltage region where it is smooth and stable. On the other hand, in the low voltage region, the sensing yield distribution is steep.

Additionally, the experimental results as shown in Fig.5 (a) indicate that the time of peak V_{IN} (T_P) is almost constant, irrespective of V_{DD} changes. Nevertheless, V_{DD} does have a slight influence on the peak time of V_{IN}. Furthermore, for supply voltages ranging from 0.6V to 1.1V, the time of maximum sensing yield (T_Y) remains almost constant, and the delay between the maximum-yield time and the time of peak V_{IN} is unchanged.

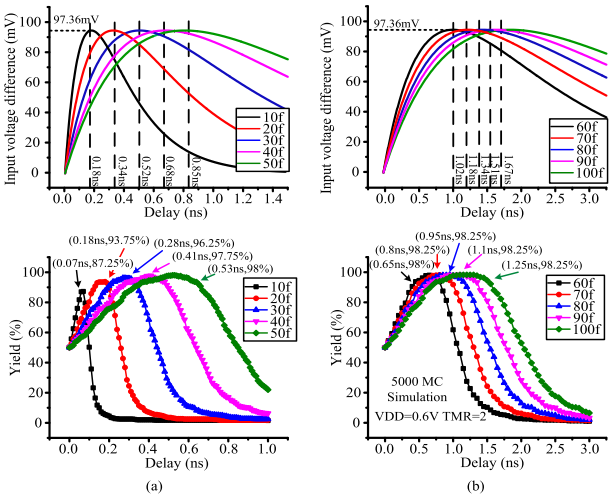

Fig. 6. Sensing yield according to Monte-Carlo simulation versus sensing delay for values of τ (a) $0.6 \times 10^{-10}S$~$3.0 \times 10^{-10}S$. (b) $3.6 \times 10^{-10}S$~$6.0 \times 10^{-10}S$.

Fig.5(b) depicts the sensing yield versus sensing delay for TMR ratio ranging from 100% to 240%. The sensing yield distribution is plotted in relation to V_{IN}, and it can be observed that both the distribution of sensing yield and the V_{IN} curve become higher with larger TMR ratio. It is evident from Fig.5(b) that TMR has a slight impact on the peak sensing yield, and within a certain range of TMR, the maximum-yield point remains nearly constant.

C. Impact of Resistance and Parasitic Capacitance

Fig.6 presents the sensing yield as a function of sensing delay for time constant(τ) ranging from $0.6 \times 10^{-10}S$ to $6.0 \times 10^{-10}S$. It has been verified that during the conduction of the access transistors, the on-resistance of the MOSFETs and the parasitic resistances along the path are significantly lower than the low resistance state of the MTJ (R_P). In the following text, R_P represents the total resistance on the path. The sensing yield distribution is plotted in Fig.6 in relation to the input voltage difference V_{IN}. The transfer characteristics of V_{IN} demonstrate the impact of time constant τ, which exhibits a unique effect on V_{IN} compared to V_{DD} and TMR, as shown in the top diagram of Fig.6.

The time of peak V_{IN} is delayed with an increase in τ, while the peak value remains nearly constant. The time of maximum sensing yield is postponed, and the delay between the maximum-yield time and the peak-V_{IN} time is linearly correlated to τ. Furthermore, the peak time of V_{IN} is mainly influenced by τ, and a larger τ leads to a smoother and more stable distribution of sensing yield, while a smaller (τ) results in a steep sensing yield distribution.

IV. Optimal Timing Description And Modeling

In Section III, the factors affecting the sensing yield distribution and input voltage difference of VSA are thoroughly discussed. It has been found that the parasitic capacitance on the sensing path is a crucial factor that significantly influences the sensing yield distribution. To determine the optimal activated timing of VSA accurately, this section develops a precise model based on statistical analysis and theoretical calculations. Fig.7 displays the simplified derivation process. Firstly, by analyzing the voltage swing of the bit-line during the discharge period, we establish the theoretical expression of the input voltage of the VSA versus time t. Then, the peak time T_P at which V_{IN} attains the maximum value is found by means of the obtained V_{IN} expression. By fitting MC simulations and theoretical analyses, we then present the linear correlation between T_P and T_Y associated with τ. Finally, we use this linear dependence to derive the equation for T_Y, which represents the optimal timing to enable the VSA.

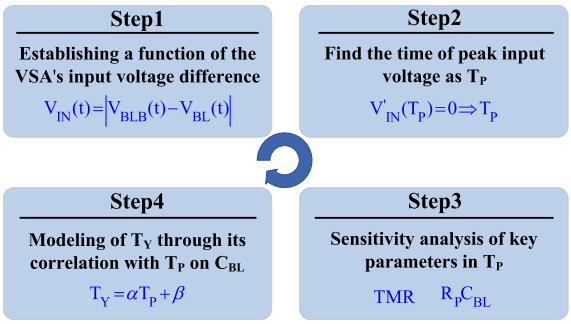

Fig. 7. Process of analyzing and modeling the optimal SAE timing.

A. Peak Time of Input Voltage

Initially, the peak value of the input voltage difference is discussed. The input voltage difference V_{IN}, which is defined as the difference between the bit-lines. Its behavior can be determined by analyzing the function of the bit-line transient behavior, and expressed by equation (2), which describes V_{IN} as a function of time t:

$$V_{IN}(t) = V_{BLB}(t) - V_{BL}(t) \tag{2}$$

It is assumed that the bit-cell connected to BL is in P state, while the bit-cell connected to BLB is in AP state. As explained in section III, the simplest approach is to calculate the average capacitance current during the charge down process to obtain the transfer process of BL. Following the word line activation, the voltage of BL and BLB at time t can be calculated as

$$V_{BL}(t) = V_{pre} - V_{pre} \times [1 - exp(-t/R_P C_{BL})] \tag{3}$$

$$V_{BLB}(t) = V_{pre} - V_{pre} \times [1 - exp(-t/R_{AP} C_{BLB})] \tag{4}$$

where V_{pre} represents the pre-charge voltage of the bit-line; R_P and R_{AP} indicate the total resistance of bit-cell, C_{BL} and C_{BLB} represent the line parasitic capacitances, and t represents the time variable.

In this analysis, we assume that the capacitance of two bit lines (C_{BL} and C_{BLB}) in the same column is the same, as

$$C_{BL} = C_{BLB} \tag{5}$$

To determine the time of peak voltage difference (T_P), we first need to evaluate the derivative of (2) as

$$V_{IN}{}'(t) = (V_{pre}/C_{BL}) \times [exp(-t/R_P C_{BL})/R_P - exp(-t/R_{AP} C_{BL})/R_{AP}] \tag{6}$$

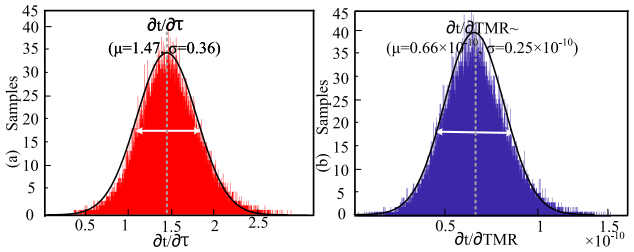

Fig. 8. Histogram (10k samples) of T_P partial derivative with respect to the different control variables. (a) Sensibility of T_P regarding τ. (b) Sensibility of T_P regarding TMR.

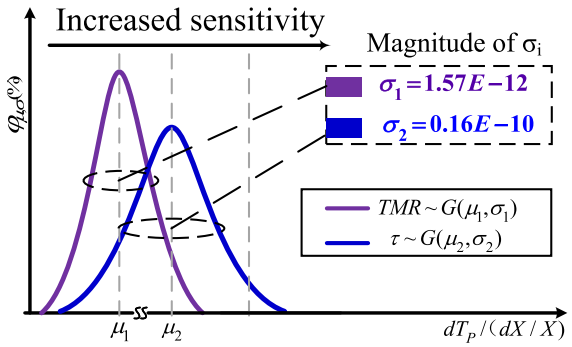

Fig. 9. Comparison of T_P sensitivity to key parameters.

As shown in Fig.3, the V_{IN} is a unimodal function, which means it is a single-peaked univariate function. We define the time of the peak input voltage as T_P, and $V_{IN}(t)$ increases for t $\leq T_P$ and decreases for t $> T_P$. Using the characteristics of a unimodal function, we set the derivative function (2) equal to 0 at T_P, which gives

$$V_{IN}{}'(T_P) = 0 \tag{7}$$

Solving for T_P obtains the desired expression:

$$\Rightarrow exp(-T_P/R_P C_{BL})/R_P = exp(-T_P/R_{AP} C_{BL})/R_{AP} \tag{8}$$

Taking the derivative of both sides with respect to e results in

$$\Rightarrow T_P[(1/R_P C_{BL}) - (1/R_{AP} C_{BL})] = \ln(R_{AP}/R_P) \tag{9}$$

Therefore, T_P can be calculated as

$$\Rightarrow T_P = \frac{R_{AP} R_P C_{BL} \times [\ln(R_{AP}) - \ln(R_P)]}{R_{AP} - R_P} \tag{10}$$

For the MTJ, the effective tunneling magnetoresistance ratio (TMR) is defined as

$$TMR = \frac{R_{AP} - R_P}{R_P} \times 100\% \tag{11}$$

The R_{AP} can be expressed as

$$\Rightarrow R_{AP} = (1 + TMR) R_P \tag{12}$$

By substituting Eq. (12) into Eq. (10), we arrive at the final equation:

$$\Rightarrow T_P = \frac{R_P C_{BL}(1 + TMR) \times \ln(1 + TMR)}{TMR} \tag{13}$$

B. Sensitivity Analysis

In previous analyses, mathematical models of circuit systems are typically characterized by multiple parameters [22], [23], [24]. These parameters represent the search space dimensions for optimization tasks, particularly in the case of size and shape optimization. For complex models, it can be difficult to estimate which parameters have a high influence on the results and which do not [25], [26], [27]. This is why sensitivity analysis is used to quantify parameter influences on model solutions [28], [29], [30].

To estimate the quantitative relation between input-voltage peak points and control variables, we perform a sensitivity analysis to analyze the variation of peak time when some control variables are changed under a given operating state. The partial derivative (PaD) is a widely used method to estimate sensitivity [31], [32]. In this method, the sensitivity of T_P can be expressed as the first-order partial derivative with respect to the control variables. Therefore, a profile of T_P variations for small changes of each control variable can be calculated by using this method.

Equation (13) explains the impact of various parameters. As expected, peak time (T_P) is directly proportional to the product of the bit-line parasitic capacitance (C_{BL}) and the low resistance of the MTJ (R_P), the tunnel magnetoresistance ratio (TMR). We denote the product of parasitic capacitance and resistance as the time constant τ. To obtain sensitivities between the peak point and crucial parameters, the derivatives of the peak point formula expressed by the factors TMR and τ must to be derived. The first-order derivative can be obtained by differentiating the Eq. (13) with respect to τ, which gives

$$\frac{\partial T_P}{\partial \tau} = \frac{\ln(TMR + 1) \times (1 + TMR)}{TMR} \tag{14}$$

The derivatives of Eq. (13) with respect to TMR can be expressed as

$$\frac{\partial T_P}{\partial TMR} = \frac{C_{BL} R_P \times [TMR - \ln(1 + TMR)]}{TMR^2} \tag{15}$$

To calculate the effect of the crucial parameters on the T_P, the data reported in recent papers were investigated to determine parameters ranges. The resistance value (R_P) of the low resistance state of MTJ reported in the literature fluctuates in the range of 2∼6 kΩ, [3], [5], [33], with a standard deviation (σ) of the variation distribution of R_P resistance value of 4%∼8% [34], [35] and TMR varies from 80∼200% [4], [36]. These values were substituted into the derivative equation (Eq. (14) - Eq. (15)) respectively.

Following the derivation in Eq. (14) and Eq. (15), the sensitivity of T_P with respect to τ and TMR is given by ∂T_P / $\partial \tau$ and ∂T_P / ∂TMR. For the parameter of R_P, we assume a normal distribution with a mean value of 6 kΩ and sigma value of 0.48 kΩ. The parameter of C_{BL} is set as a normal distribution with a mean of 40fF and a sigma of 10fF. As shown in Fig.8 (a), the calculation results the sensitivity of TP regarding τ are distributed between 0.5∼2.5. The parameter of TMR is set as a normal distribution with a mean of 150% and a sigma of 20%. Fig.8 (b) shows the sensitivity of T_P regarding TMR,

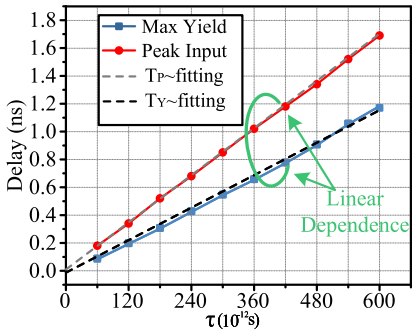

Fig. 10. Linearization of the terms C_{BL} and T_P and T_Y (obtained from simulation) by ($k_1\ \tau\ +\ b_1$) and ($k_2\ \tau\ +\ b_2$), respectively. Parameters are k_1 = 0.0027, k_2 = 0.0022, b_1 = 0.0167, b_2 = -0.101, TMR=150% and V_{DD} = 0.6V.

which is expressed as $\partial T_P\ /\ \partial TMR$. The results obtained indicate the determination of partial derivative with respect to each variable and the estimation of T_P sensitivity. Fig.9 shows that the partial derivatives have a Gaussian distribution G(μ_i,σ_i), and the sensitivity of T_P to TMR and τ gradually increases. The results have been normalized to facilitate a fair comparison when comparing sensitivities. It is evident that the sensitivity of the σ of $\partial T_P\ /\ \partial\tau$ is significantly higher than the σ of $\partial T_P\ /\ \partial TMR$. This implies that the impact of T_P, TMR can be disregarded when compared to the resistance and parasitic capacitance of BL. The theoretical derivation is in agreement with the simulation outcomes discussed in section III.

C. Timing Model of Strongest Sensing Point

According to the simulation results, we have derived distribution curves for peak voltage difference point T_P and peak sensing yield point T_Y as τ changes, as displayed in Fig.10. The linear correlation between T_P and T_Y concerning τ has been proven through theoretical derivation, permitting to express T_P and T_Y as linear equations:

$$T_P = k_1\tau + b_1 \tag{16}$$

$$T_Y = k_2\tau + b_2 \tag{17}$$

To determine the T_Y in terms of the T_P, we can express it as

$$\Rightarrow T_Y = \frac{k_2}{k_1}(T_P - b_1) + b_2 \tag{18}$$

This equation can be rearranged to

$$\Rightarrow T_Y = \frac{k_2}{k_1}T_P - (\frac{k_2}{k_1}b_1 - b_2) = \alpha T_P + \beta \tag{19}$$

By substituting Eq. (13) in Eq. (20), we get

$$\Rightarrow T_Y = \alpha \times \frac{\tau \times (1+TMR)\ln(1+TMR)}{TMR} + \beta \tag{20}$$

By utilizing this formula, the optimal timing for activating the voltage sense amplifiers whilst carrying out voltage sensing scheme can be established.

V. MTJ-Variation-Tolerant Replica Bit-Line

In the previous sections, it has been discovered that the transfer function described in Eq. (20) of VSA is determined by the behavior of BL discharge. The optimal enable timing of VSA is taken for the input voltage to reach its maximum driving point.

For suppressing the timing variation of SAE, a commonly used technique involves using a replica bit-line [11], [37], [38]. This technique makes use of replica cells and replica bit-lines. The replica cells are created using memory-cell transistors rather than logic transistors, which ensures that replicated bit-line delay is traced in relation to systematic transistor threshold voltage, supply voltage, and temperature variations of a normal bit-line delay. The replica cells are fabricated with memory-cell transistors, not with logic transistors. This is much better than a logic gate delay composed of an inverter chain.

Although conventional RBL techniques are usually proposed for SRAM, and the BL swing behaviors of reading '1' and '0' are the same, making it easy to track the discharging process of BL by conventional replica techniques, this is not the case with MRAM.

The conventional RBL techniques are generally proposed for SRAM, and the BL swing behaviors of reading '1' and '0' are the same, making it easy to track the discharging process of BL by conventional replica techniques [38], [39]. However, this is not the case with MRAM, since the swing process of BL at AP state and P state are different in MRAM, and the delay distributions of ΔV development are different either [11]. Fig.11 illustrates the delay distributions of BL discharging caused by process variations. The discharging delay distribution of AP state is larger than that of P due to asymmetrical STT efficiency [40]. Therefore, accurately controlling the timing in MRAM with traditional RBL techniques is difficult.

To overcome the problem, we propose a MTJ-Variation-Tolerant replica bit-line (MVT-RBL) technique that achieves both the suitable timing generation for SAE and tracks the different states of MTJ. MVT-RBL technology ensures that SAE always locks in its strongest sensing time, minimizing BER(Bit Error Rate) and achieving an increase in sensing yield. BER indicates that sensing errors occur when the sensing margin cannot overcome the mismatch and offset of the sensing circuit [9], [41], [42].This technique is especially useful for lower-voltage MRAM.

A. Principle of the MVT-RBL Technique

Fig.12 (a) shows the block diagram of proposed MVT-RBL. The 2T+2MTJ replica cell (RC) is comprised of two MTJs with complementary states, similar to a normal bit-cell. If the right MTJ is in parallel state and the left one is in antiparallel state, we designate this configuration as '0'. The replica bit-lines (RBL/RBLB) and replica cells are arranged to replicate the normal bit-line capacitance (C_{BL}). The replica column is divided into two stages with an inverter inserted between the left column and right column, and the logical threshold voltage of the inserted inverter is set to $V_{DD}/2$.

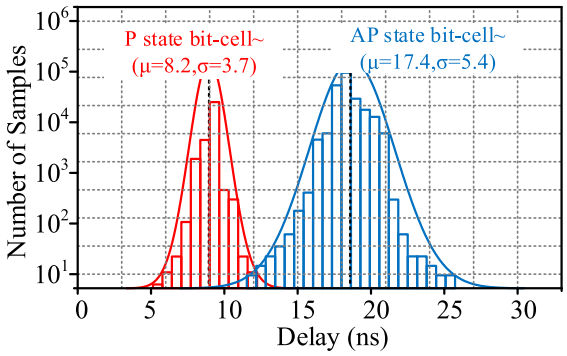

Fig. 11. Delay variability of BL discharging to $V_{DD}/2$ in P/AP state (10^8 runs Monte-Carlo simulation, V_{DD}=0.9 V). The mean value of AP/P state is 8.2ns/17.4ns and the sigma is 3.7ns/5.4ns.

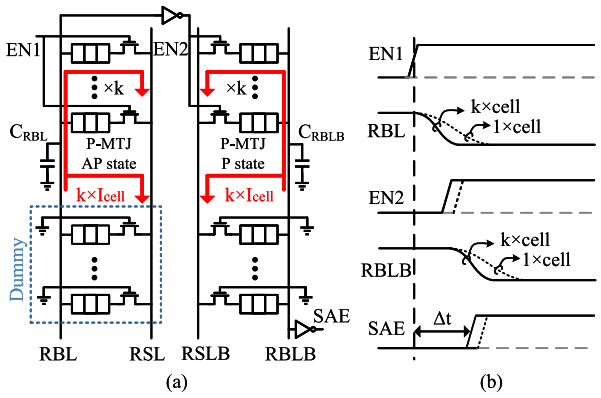

Fig. 12. (a) Block diagram of proposed MVT-RBL. (b) Operation-al waveforms of proposed timing replica circuit.

Fig.12 (b) shows operational waveforms of the proposed MVT-RBL. The sequential control of SAE is executed in following manner: first, the replica bit-line and the normal bit-lines are pre-charged to V_{DD}. Subsequently, the first-stage enable signal (EN1) and selected WL are activated. Then, the left-column replica cells draw current from the first-stage replica bit-line (RBL), developing a voltage drop. Once the RBL voltage level drops to $V_{DD}/2$, the second-stage enable signal (EN2) is initiated, and the corresponding replica bit-line (RBLB) releases its voltage.

The replica cell count of each column is k, the replica cell current I_{rep} flowing from the replica bit-line is k times I_{cell}. The value of k is selected for achieving precise trimming of the timing of SAE. In order to cover the different distributions of MTJ states, the timing of SAE is composed of two different type delays. To accommodate the different distributions of MTJ states, the timing of SAE consists of two categories of delays. One delay originates from the RBL discharging delay caused by the P-state replica cells, whereas the other delay is activated by the AP-state replica cells causing the RBLB discharging delay.

B. Tracking Of Strongest Sensing Point

To ensure that the SAE generated by the proposed MVT-RBL always locked at the strongest sensing time, we firstly establish the timing model for the SAE signal. The delay of SAE (T_{SAE}) is determined by the BL discharging

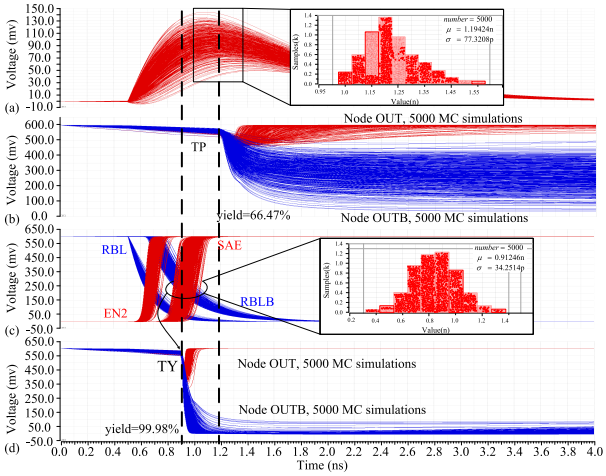

Fig. 13. (a) The 5000 times Monte-Carlo simulation results of timing variation of the V_{IN} in the SS corner, 25°C, V_{READ}=0.6 V and C_{BL}=41 fF.(b) Monte-Carlo transient behaviors of node OUT and OUTB when the VSA enabled at T_P. (c) Simulation waveforms of the proposed MVT-RBL and the SAE timing at the replica cell count k is 2 when V_{DD} is 0.6 V.(d) Monte-Carlo transient behaviors of node OUT and OUTB when the sampling time of VSA set at T_Y.

delay (T_{BL}) the BLB discharging delay (T_{BLB}), resulting in

$$T_{SAE} = T_{RBL} + T_{RBLB} \tag{21}$$

We assume the R_{AP} and R_P are the two-states resistances of replica cell, C_{RBL} /C_{RBLB} are the capacitances of the replica bit-lines, which are the same as C_{BL}. The threshold voltage of the inverter between the discharging stages is $V_{DD}/2$, and k represents the replica cell count of each column. The discharging delay of two-stage replica bit-line can be calculated by

$$T_{RBL} = \frac{1}{k} R_{AP} \times C_{RBL} \times \ln 2 \tag{22}$$

$$T_{RBLB} = \frac{1}{k} R_P \times C_{RBLB} \times \ln 2 \tag{23}$$

By substituting Eq. (23) into Eq. (24) in Eq. (22), we get

$$T_{SAE} = \frac{1}{k}(R_P + R_{AP}) \times C_{BL} \times \ln 2 \tag{24}$$

To calculate the k value for SAE enabling to be enabled at the optimal timing, the following formula is obtained

$$T_{SAE} = T_Y \tag{25}$$

Then substitute Eq. (21) into Eq. (26), and as discussed in Eq. (21) β can be approximated by 0, k obtained as

$$\Rightarrow k \approx \frac{TMR(2+TMR)ln2}{\alpha(1+TMR)ln(1+TMR)} (\beta \approx 0) \tag{26}$$

$\beta \approx 0$ was obtained by fitting a large amount of experimental data, as shown in Fig.10, where the intersection point of the fitted curve with the y-axis is close to 0. Equation (26) elucidates the impact of TMR and α, where TMR is an inherent parameter of the device, and α is associated with the array design, including factors such as array size and bit-cell configuration. In summary, the value of k is an internal circuit parameter. Once the value of k is determined, the enabled time of SA is always locked at the strongest driving point.

TABLE I
PHYSICAL PARAMETERS OF STT-MTJ COMPACT MODEL

Parameters	Description	Default Value
ΔH_0	Activation energy	0.8eV
Γ	Field acceleration parameter	1.7 cm/MV
β	Shape parameter	1.5
k_B	Boltzmann constant	$8.625 \cdot 10^{-5}$ eV/K
T_0	Ambient temperature	300 K
Variable	Description	Default Value
t_{OX}	Thickness of oxide barrier	0.85nm
TMR(0)	TMR ration with 0 stress voltage	150%
Area	MTJ surface	$40\text{nm} \cdot 40\text{nm} \cdot \pi/4$
t_{sl}	Thickness of free layer	1.3nm
V_{sl}	Volume of free barrier	Area $\cdot\, t_{sl}$

C. Estimate of Tracking Capability

To evaluate the tracking capability of the proposed scheme for the optimal sensing point, we conducted Monte-Carlo simulations under the condition of 0.6V supply voltage, SS corner, 25°C.

The simulation results demonstrate that under the influence of CMOS & MTJ PVT variation, the peak time of input voltage difference conforms to a normal distribution with a mean value of 1.194ns and a sigma value of 77.32ps as shown in Fig.13 (a). Moreover, Fig.13 (b) displays the MC simulation waveforms of node OUT and OUTB with the VSA enabled at T_P. The sensing yield is defined as the whether the OUTB can be pulled down to $V_{DD}/4$ at 2ns, was found to be 66.47%. Additionally, Fig.13 (c) shows the Monte-Carlo simulation results of the proposed MVT-RBL, the standard deviation of the SAE timing is 34.25 ps and the mean value is 0.91 ns. Compared to the MTJ-LRB technology proposed in [11], the standard deviation of SAE has been reduced by 77.4%. The TMR is set to 150% and k can be calculated by Eq. (26) to be approximately 2. Lastly, Fig.13 (d) shows the simulation waveforms of VSA output node under the control of the proposed MTJ-RBL. The sensing yield reaches 99.98%, which is significantly higher than that of the T_P sensing condition.

VI. SIMULATION RESULTS AND COMPARISONS

A. Simulation Setup

As previously stated, the enhancement of STT-MRAM yield concentrates on optimizing the sensing timing and the sequential control scheme for the SAE signal. The optimal active timing of SAE is corrected forward to T_Y based on Eq.(20), and the MVT-RBL technique is utilized to generate the SAE signal, rather than the CMOS logic chain. In this section, we compare the proposed schemes with conventional methods to demonstrate the performance advantage. The post-simulations are conducted based on a 256 × 256 STT-MRAM array with the same peripheral circuitry to ensure a fair comparison. It was designed using 28nm CMOS & MTJ technology following cell-based design flow, involving RTL (register-transfer level) verilog coding, logic synthesis, automatic placement-routing, and pre-/post-layout simulation. All simulation results in this section are obtained from 5000 MC simulations under tt corner and 25°C.

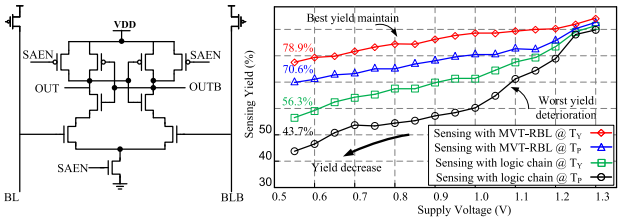

Fig. 14. Comparison of sensing yield for CL-VSA with different sensing techniques at wide voltages.

Table I summarizes all fundamental MTJ parameters used in our analysis, which were derived from the model in [43], and the process fluctuation parameters are set according to the simulation setup in [44]. Monte-Carlo simulations were used to analyze process variations using statistical models for both CMOS and MTJ devices. The MTJ area and the thickness of free layer were assumed to be Gaussian distributed as customary. The variability of the MTJ resistance equals to 5% according to the experimental data in [45].

B. Sensing Yield Improvement

By using the MTJ-RBL technique shown in Fig.12 and Eq.(26), the value of k is calculated to be approximately 2, which means that the replica cell count of each column is 2. Fig. 14 shows the circuit diagram of CL-VSA and the yield comparison of the 64kb STT-MRAM with CL-VSA is shown for sensing operation across various supply voltages under the condition of k=2. The control variable method is used to compare the four cases, where the proposed optimal timing of the SAE and control path are employed separately and jointly to elucidate yield optimization compared to counterparts. Sensing yield is defined based on Eq. (1) with a voltage simulation step size of 50mv. The yield remains above 90% when STT-MRAM operates above 1.2V, but the sensing yield shows different degradation in all four cases as the supply voltage decreases due to process variation. The sensing scheme with logic chain at TP shows the most substantial degradation at 0.55V, reduced to 43.7%. However, using the methods of sensing with logic chain at T_Y and MVT-RBL at T_Y for the voltage sensing scheme improves the yield of STT-MRAM by 12.6% and 26.9%, respectively. By employing the combined MVT-RBL technique and the optimal SAE timing (T_Y), the sensing yield improvement of 35.2% is obtained, maintaining 78.9% at 0.55V supply voltage. The sensing scheme at the strongest driving point improves the input voltage of SA while maintaining a large voltage difference, resulting in better yield than the scheme sampled at the peak point. It has been demonstrated that the combination of MVT-RBL technology and optimal SAE timing (T_Y) can also improve SA yield under other process corners.

To demonstrate the availability of the proposed optimization scheme for various voltage sense amplifiers, we select the following four representative VSAs commonly used in memory for simulation. We apply these VSAs in the same sensing circuit as the CL-VSA with the same condition and compare the sensing yield of STT-MRAM in four cases. Furthermore, we compared the advantages of the combined scheme employed by the voltage sensing circuitry in terms of

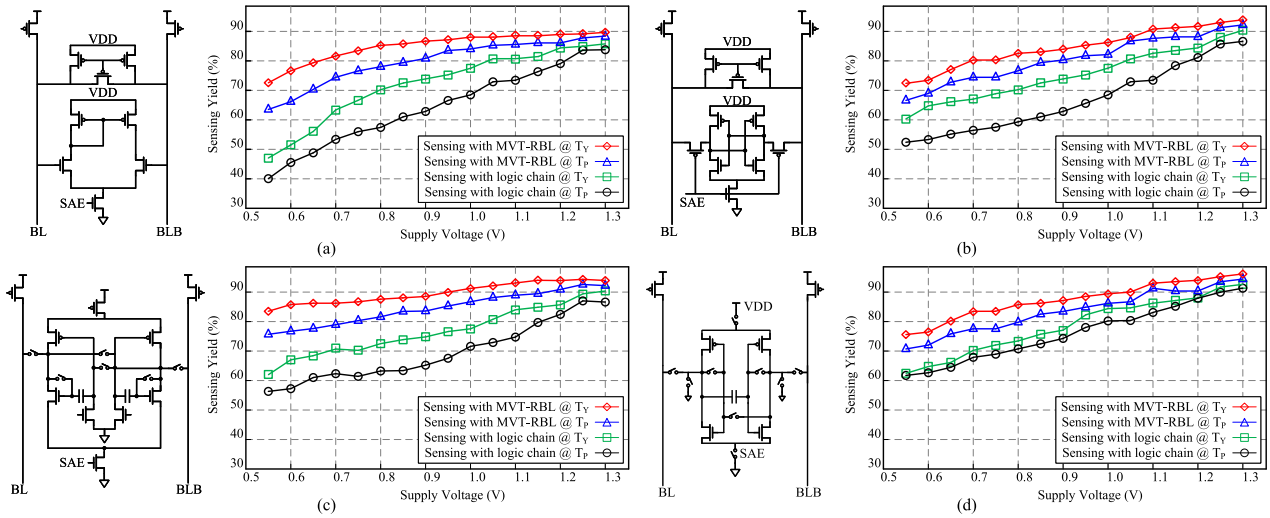

Fig. 15. Comparison of sensing yield for various voltage-type sense amplifiers with different sensing techniques at wide voltages. (a) current mirror voltage sense amplifier (CM-VSA). (b) decoupling latch-type voltage sense amplifier (DL-VSA). (c) dual-capacitor offset cancellation latch-type voltage sense amplifier (DCOC-VSA).(d) single-capacitor offset cancellation latch-type voltage sense amplifier (SCOC-VSA).

energy consumption and access delay with the corresponding conventional technique.

1) Fig.15(a) depicts the current mirror voltage sense amplifier (CM-VSA), which amplifies the voltage difference in bit-lines using a high open-loop gain current mirror. However, the use of wide transistors to reduce offset voltage and process variations results in high power consumption [46], [47].
2) Fig.15(b) shows the decoupling latch-type voltage sense amplifier (DL-VSA), a conventional latch-type sense amplifier that operates swiftly due to the positive feedback provided by the two inverters. Its power consumption is also low. Still, problems can occur during the decision phase as a result of shared input and output nodes [16].
3) Fig.15(c) displays the dual-capacitor offset cancellation latch-type voltage sense amplifier (DCOC-VSA). This compensates for mismatched transistor threshold voltages, effectively reducing the input offset voltage of the comparator. However, due to the longer compensation period required for offset cancellation, the DCOC-VSA has limited operational speed [48].
4) Fig.15(d) shows the single-capacitor offset cancellation latch-type voltage sense amplifier (SCOC-VSA), which employs a single capacitor for offset cancellation. This method improves sensing margin, while also reducing the area overhead compared to the DCOC-VSA. However, the access time of SCOC-VSA remains slow [7].

As shown in Fig.15, consistent with CL-VSA, the sensing yield of MRAM with the aforementioned four VSAs decreases as the voltage decreases. CM-VSA and DL-VSA are preferable choices for low-power and high-speed MRAM. However, if not offset-compensated for process variation, a significant decrease in supply voltage can lead to a notable drop in MRAM sensing efficiency. With the implementation of the SAE sequence control circuit using logical chains, a reduction in sensing yield of 40% and 51% was observed in CM-VSA and DL-VSA, respectively. In contrast, the proposed combined

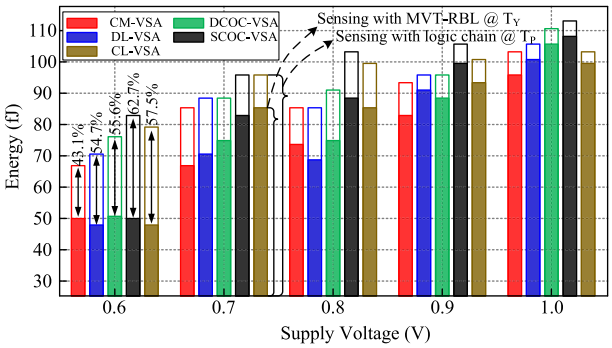

Fig. 16. Comparison of energy consumption for various voltage-type sense amplifiers with different sensing techniques at 0.6∼1.0V.

scheme resulted in improvements of 31% and 20% in sensing yield at 0.55V, respectively, with both remaining at 71%.

Though compensation of the capacitors for DCOC-VSA and SCOC-VSA neutralizes the offset of VSA when the conventional scheme (sensing with logic chain at T_P) is used for the sensing circuit, the enhancements in yield are not significant compared to the aforementioned VSAs at low voltages. For instance, at 0.55V supply voltage, the sensing yield of MRAM is a mere 58% and 61%, respectively. On the other hand, when the proposed optimization scheme is applied to the sensing circuit, the sensing yield of MRAM remains at 83% and 77%, achieving 25% and 15% improvement, respectively. As discussed above, the process variation on the critical path of the sensing circuit may result in the unreliability of the conventional sensing sequence scheme. At low voltages, merely compensating the offset of VSA isn't effective in improving the sensing yield. However, paying further attention to optimize SAE timing can significantly enhance the robustness of the sense circuit.

C. Energy and Access Timing Improvement

Fig.16 illustrates a comparison of energy consumption between STT-MRAM embedded with various VSA configurations at wide voltage ranges using two combination schemes.

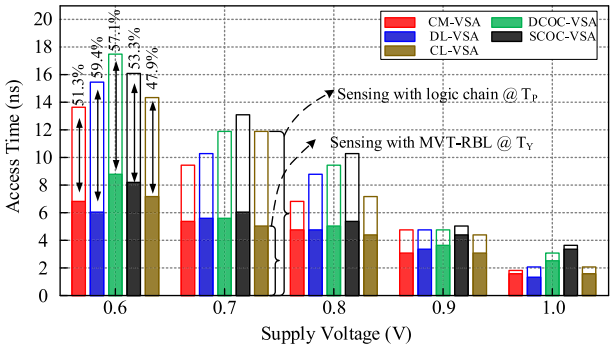

Fig. 17. Comparison of read latency for various voltage-type sense amplifiers with different sensing techniques at 0.6∼1.0V.

As mentioned earlier, reducing the supply voltage results in lower energy consumption. However, the conventional method is not effective due to the extra energy consumption caused by increased discharge time of the bit-lines when SAE is asserted at T_P, the discharge time of the bit-lines is increased, resulting in extra energy consumption. In contrast, the proposed combination scheme optimizes the enable timing for VSAs, resulting in a significant reduction in STT-MRAM energy consumption during the voltage sensing.In other words, due to the optimal timing strategy, the enabling time of SA is shortened, and the power consumption is further reduced. At a supply voltage of 0.6V, the combined adoption of optimal timing model and MVT-RBL technique enables reductions in energy consumption of 43.1%, 54.7%, 55.6%, 62.7% and 57.7%, respectively, compared to the conventional method, for STT-MRAM embedded with CM-VSA, DL-VSA, DCOC-VSA, SCOC-VSA and CL-VSA.

In addition to energy consumption and reliability benefits, the proposed combination scheme also improves read latency. Fig.17 shows that the access time of the 64Kb STT-MRAM embedded with the described above five VSAs at wide voltages applying two combination schemes. The combination T_Y and MVT-RBL technique scheme offers a substantial delay improvement over the conventional method. The optimal timing of SAE reduces the extra development time of V_{BL}, and at the strongest driving point T_Y, a faster flip is obtained by VSAs. The MVT-RBL technique suppresses timing variation of SAE, reducing read latency. For example, at a 0.6V supply voltage, the access time for STT-MRAM embedded with CM-VSA, DL-VSA, DCOC-VSA, SCOC-VSA and CL-VSA reduced by 51.3%, 59.4%, 57.1%, 53.3% and 47.9%, respectively.

These simulation results suggest that the proposed scheme applied to the voltage sensing circuit can achieve an optimal trade-off between reliability, energy consumption, and sensing speed. CL-VSA, CM-VSA, and DL-VSA are characterized by low energy consumption and high operation speed. However, the offset between the inverter pair deteriorates at low supply voltage, limiting reliability. Nevertheless, the proposed design method minimizes the impact of yield degradation, achieving a yield of more than 70% even when the supply voltage reduces to below 0.6V. On the other hand, DCOC-VSA and SCOC-VSA improve sensing margin by compensating for offset using a capacitor. However, the higher energy consumption with slower operational speed cannot be ignored. Nevertheless, implementing the proposed joint scheme moderates access time and energy consumption penalty. The access delay within 10ns and energy consumption of 80fJ are acceptable at 0.6V supply voltage.

VII. Conclusion

The present study develops yield optimized schemes for voltage sensing circuitry in STT-MRAM. By analyzing the timing characteristics of STT-MRAM sensing process, we propose a yield optimized model along with a PVT variation tracking plan. The post-MC simulations under 28nm CMOS&MTJ technology have demonstrated the proposed schemes can achieve a significant improvement in sensing yield, access latency, and a reduction in energy consumption for STT-MRAM read operation. Overall, the research has considerable implications for enhancing the reliability and performance of voltage sensing of STT-MRAM.

References

[1] H. Jeong et al., "Offset-compensated cross-coupled PFET bit-line conditioning and selective negative bit-line write assist for high-density low-power SRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 4, pp. 1062–1070, Apr. 2015.

[2] A. Salahvarzi, A. M. H. Monazzah, M. Fazeli, and K. Skadron, "NOSTalgy: Near-optimum run-time STT-MRAM quality-energy knob management for approximate computing applications," *IEEE Trans. Comput.*, vol. 70, no. 3, pp. 414–427, Mar. 2021, doi: 10.1109/TC.2020.2989243.

[3] M. Talebi, A. Salahvarzi, A. M. H. Monazzah, K. Skadron, and M. Fazeli, "ROCKY: A robust hybrid on-chip memory kit for the processors with STT-MRAM cache technology," *IEEE Trans. Comput.*, vol. 70, no. 12, pp. 2198–2210, Dec. 2021, doi: 10.1109/TC.2020.3040152.

[4] E. Cheshmikhani, H. Farbeh, S. Miremadi, and H. Asadi, "TA-LRW: A replacement policy for error rate reduction in STT-MRAM caches," *IEEE Trans. Comput.*, vol. 68, no. 3, pp. 455–470, Mar. 2019, doi: 10.1109/TC.2018.2875439.

[5] Y.-C. Shih et al., "Logic process compatible 40-nm 16-mb, embedded perpendicular-MRAM with hybrid-resistance reference, sub-μA sensing resolution, and 17.5-nS read access time," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1029–1038, Apr. 2019, doi: 10.1109/JSSC.2018.2889106.

[6] M. Natsui et al., "A 47.14-μW 200-MHz MOS/MTJ-hybrid nonvolatile microcontroller unit embedding STT-MRAM and FPGA for IoT applications," *IEEE J. Solid-State Circuits*, vol. 54, no. 11, pp. 2991–3004, Nov. 2019, doi: 10.1109/JSSC.2019.2930910.

[7] Q. Dong et al., "A 1-Mb 28-nm 1T1MTJ STT-MRAM with single-cap offset-cancelled sense amplifier and in situ self-write-termination," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 231–239, Jan. 2019, doi: 10.1109/JSSC.2018.2872584.

[8] W. Kang, T. Pang, Y. Zhang, D. Ravelosona, and W. Zhao, "Dynamic reference sensing scheme for deeply scaled STT-MRAM," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2015, pp. 1–4, doi: 10.1109/IMW.2015.7150282.

[9] Q. Trinh, S. Ruocco, and M. Alioto, "Dynamic reference voltage sensing scheme for read margin improvement in STT-MRAMs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 4, pp. 1269–1278, Apr. 2018.

[10] B. Song, T. Na, J. Kim, J. P. Kim, S. H. Kang, and S. Jung, "Latch offset cancellation sense amplifier for deep submicrometer STT-RAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 7, pp. 1776–1784, Jul. 2015, doi: 10.1109/TCSI.2015.2427931.

[11] Y. Zhou, H. Cai, B. Liu, W. Zhao, and J. Yang, "MTJ-LRB: Proposal of MTJ-based loop replica bitline as MRAM device-circuit interaction for PVT-robust sensing," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 12, pp. 3352–3356, Dec. 2020.

[12] Y. Zhou et al., "A self-timed voltage-mode sensing scheme with successive sensing and checking for STT-MRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 5, pp. 1602–1614, May 2020.

[13] Y. Li, H. Schneider, F. Schnabel, R. Thewes, and D. Schmitt-Landsiedel, "DRAM yield analysis and optimization by a statistical design approach," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 12, pp. 2906–2918, Dec. 2011, doi: 10.1109/TCSI.2011.2157741.

[14] M. Patel, J. S. Kim, H. Hassan, and O. Mutlu, "Understanding and modeling on-die error correction in modern DRAM: An experimental study using real devices," in *Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2019, pp. 13–25, doi: 10.1109/DSN.2019.00017.

[15] A. K. Singh, K. He, C. Caramanis, and M. Orshansky, "Modeling and optimization techniques for yield-aware SRAM post-silicon tuning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 8, pp. 1159–1167, Aug. 2014, doi: 10.1109/TCAD.2014.2317571.

[16] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE J. Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, Jul. 2004, doi: 10.1109/JSSC.2004.829399.

[17] Q. Dong et al., "A 1Mb 28 nm STT-MRAM with 2.8 ns read access time at 1.2V VDD using single-cap offset-cancelled sense amplifier and in-situ self-write-termination," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 480–482, doi: 10.1109/ISSCC.2018.8310393.

[18] T.-H. Yang, K.-X. Li, Y.-N. Chiang, W.-Y. Lin, H.-T. Lin, and M.-F. Chang, "A 28 nm 32Kb embedded 2T2MTJ STT-MRAM MACRO with 1.3ns read-access time for fast and reliable read applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 482–484, doi: 10.1109/ISSCC.2018.8310394.

[19] H. Lim, S. Lee, and H. Shin, "Switching time and stability evaluation for writing operation of STT-MRAM crossbar array," *IEEE Trans. Electron Devices*, vol. 63, no. 10, pp. 3914–3921, Oct. 2016, doi: 10.1109/TED.2016.2597195.

[20] T. Na, J. Kim, J. P. Kim, S. H. Kang, and S.-O. Jung, "Reference-scheme study and novel reference scheme for deep submicrometer STT-RAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 12, pp. 3376–3385, Dec. 2014, doi: 10.1109/TCSI.2014.2327337.

[21] A. Lee, R. Jagannathan, D. Wu, and K. L. Wang, "A 2-D calibration scheme for resistive nonvolatile memories," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 6, pp. 1371–1377, Jun. 2020, doi: 10.1109/tvlsi.2020.2975589.

[22] X. Yu and N. M. Neihart, "Design and characterization of symmetric multi-tap transformers," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2012, pp. 954–957, doi: 10.1109/ISCAS.2012.6272203.

[23] T. Levi, N. Lewis, J. Tomas, and S. Renaud, "Application of IP-based analog platforms in the design of neuromimetic integrated circuits," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 11, pp. 1629–1641, Nov. 2012, doi: 10.1109/TCAD.2012.2204992.

[24] Z. Liu, S. X.-D. Tan, H. Wang, S. Swarup, and A. Gupta, "Compact nonlinear thermal modeling of packaged integrated systems," in *Proc. 18th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Jan. 2013, pp. 157–162, doi: 10.1109/ASPDAC.2013.6509589.

[25] A. Chintaluri, H. Naeimi, S. Natarajan, and A. Raychowdhury, "Analysis of defects and variations in embedded spin transfer torque (STT) MRAM arrays," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 3, pp. 319–329, Sep. 2016, doi: 10.1109/JETCAS.2016.2547779.

[26] S. M. Nair, R. Bishnoi, and M. B. Tahoori, "A comprehensive framework for parametric failure modeling and yield analysis of STT-MRAM," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 7, pp. 1697–1710, Jul. 2019, doi: 10.1109/TVLSI.2019.2904197.

[27] K. Kim and C. Yoo, "Variation-tolerant sensing circuit for spin-transfer torque MRAM," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 62, no. 12, pp. 1134–1138, Dec. 2015, doi: 10.1109/TCSII.2015.2468971.

[28] S. M. Nair, R. Bishnoi, M. S. Golanbari, F. Oboril, F. Hameed, and M. B. Tahoori, "VAET-STT: Variation aware STT-MRAM analysis and design space exploration tool," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 7, pp. 1396–1407, Jul. 2018, doi: 10.1109/TCAD.2017.2760861.

[29] S. M. Nair, R. Bishnoi, and M. B. Tahoori, "Parametric failure modeling and yield analysis for STT-MRAM," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2018, pp. 265–268, doi: 10.23919/DATE.2018.8342016.

[30] R. Dorrance, F. Ren, Y. Toriyama, A. A. Hafez, C. K. Yang, and D. Markovic, "Scalability and design-space analysis of a 1T-1MTJ memory cell for STT-RAMs," *IEEE Trans. Electron Devices*, vol. 59, no. 4, pp. 878–887, Apr. 2012, doi: 10.1109/TED.2011.2182053.

[31] A. Karamoozian, H. Jiang, C. A. Tan, L. Wang, and Y. Wang, "An integrated approach for instability analysis of lattice brake system using contact pressure sensitivity," *IEEE Access*, vol. 8, pp. 19948–19969, 2020, doi: 10.1109/ACCESS.2020.2964337.

[32] W. Wen, Y. Zhang, Y. Chen, Y. Wang, and Y. Xie, "PS3-RAM: A fast portable and scalable statistical STT-RAM reliability/energy analysis method," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 11, pp. 1644–1656, Nov. 2014, doi: 10.1109/TCAD.2014.2351581.

[33] W. Kang, L. Chang, Z. Wang, W. Lv, G. Sun, and W. Zhao, "Pseudo-differential sensing framework for STT-MRAM: A cross-layer perspective," *IEEE Trans. Comput.*, vol. 66, no. 3, pp. 531–544, Mar. 2017, doi: 10.1109/TC.2016.2601330.

[34] Y. Zhao et al., "An STT-MRAM based in memory architecture for low power integral computing," *IEEE Trans. Comput.*, vol. 68, no. 4, pp. 617–623, Apr. 2019, doi: 10.1109/TC.2018.2879502.

[35] R. Zand and R. F. DeMara, "MRAM-enhanced low power reconfigurable fabric with multi-level variation tolerance," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 12, pp. 4662–4672, Dec. 2019, doi: 10.1109/TCSI.2019.2932379.

[36] K. Lee et al., "22-nm FD-SOI embedded MRAM technology for low-power automotive-grade-l MCU applications," in *IEDM Tech. Dig.*, Dec. 2018, p. 27, doi: 10.1109/IEDM.2018.8614566.

[37] Y. Niki et al., "A digitized replica bitline delay technique for random-variation-tolerant timing generation of SRAM sense amplifiers," *IEEE J. Solid-State Circuits*, vol. 46, no. 11, pp. 2545–2551, Nov. 2011, doi: 10.1109/JSSC.2011.2164294.

[38] J. Wu, J. Zhu, Y. Xia, and N. Bai, "A multiple-stage parallel replica-bitline delay addition technique for reducing timing variation of SRAM sense amplifiers," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 61, no. 4, pp. 264–268, Apr. 2014, doi: 10.1109/TCSII.2014.2304893.

[39] Z. Lin et al., "A pipeline replica bitline technique for suppressing timing variation of SRAM sense amplifiers in a 28-nm CMOS process," *IEEE J. Solid-State Circuits*, vol. 52, no. 3, pp. 669–677, Mar. 2017, doi: 10.1109/JSSC.2016.2634701.

[40] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori, "Asynchronous asymmetrical write termination (AAWT) for a low power STT-MRAM," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2014, pp. 1–6, doi: 10.7873/DATE.2014.193.

[41] J. Kim, Y. Jang, T. Kim, and J. Park, "A dual-domain dynamic reference sensing for reliable read operation in SOT-MRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 69, no. 5, pp. 2049–2059, May 2022.

[42] D. Zhang et al., "Reliability-enhanced separated pre-charge sensing amplifier for hybrid CMOS/MTJ logic circuits," *IEEE Trans. Magn.*, vol. 53, no. 9, pp. 1–5, Sep. 2017.

[43] *Spinmodel Library*. Accessed: Oct. 1, 2015. [Online]. Available: http://www.spinlib.com/STT_PMA_MTJ.html

[44] S. Wang, H. Lee, F. Ebrahimi, P. K. Amiri, K. L. Wang, and P. Gupta, "Comparative evaluation of spin-transfer-torque and magnetoelectric random access memory," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 2, pp. 134–145, Jun. 2016, doi: 10.1109/JETCAS.2016.2547681.

[45] J. P. Kim et al., "A 45 nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2011, pp. 296–297.

[46] T. Shakir and M. Sachdev, "A body-bias based current sense amplifier for high-speed low-power embedded SRAMs," in *Proc. 27th IEEE Int. Syst.-Chip Conf. (SOCC)*, Sep. 2014, pp. 444–448, doi: 10.1109/SOCC.2014.6948970.

[47] H. Nambu et al., "A 1.8-ns access, 550-MHz, 4.5-mb CMOS SRAM," *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1650–1658, Nov. 1998, doi: 10.1109/4.726553.

[48] J. Javanifard et al., "A 45 nm self-aligned-contact process 1Gb NOR flash with 5MB/s program speed," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 424–624, doi: 10.1109/ISSCC.2008.4523238.

Yongliang Zhou (Member, IEEE) received the Ph.D. degree from the School of Microelectronics, Southeast University, Nanjing, China. His current research interests include signal processing, analog IC design, and high-performance memory technology.

Xiao Lin received the bachelor's degree in electronic science and technology from Hunan University, Changsha, China, in 2022. He is currently pursuing the master's degree in integrated circuit engineering with the School of Integrated Circuits, Anhui University. His current research interests include digital circuit design and emerging nonvolatile memory.

Zixuan Zhou is currently pursuing the master's degree in circuits and systems with the School of Integrated Circuits, Anhui University, Hefei, China. Her current research interests include ultra-low power emerging memory design and computing in memory.

Yingxue Sun is currently pursuing the master's degree with the School of Integrated Circuits, Anhui University, Hefei, China. Her current research interests include memory performance improvement.

Yiming Wei is currently pursuing the master's degree in microelectronics and solid-state electronics with the School of Integrated Circuits, Anhui University, Hefei, China. He is also working on CIM-based neural network designs and non-volatile memory and system designs.

Zhen Yang received the bachelor's degree from Anhui University, Hefei, China, in 2021, where he is currently pursuing the master's degree in microelectronics and solid-state electronics with the School of Integrated Circuits. His current research interests include low-power circuit design, computing-in-memory, and non-volatile memory design.

Chengxing Dai received the bachelor's degree in microelectronics science and engineering from Anhui University, Hefei, China, in 2023, where he is currently pursuing the master's degree in integrated circuit engineering with the School of Integrated Circuits. His research interests include circuit design and memory radiation hardening.

JingXue Zhong is currently pursuing the master's degree with the School of Integrated Circuits, Anhui University, Hefei, China. She is committed to the detection of circuit timing under 77k.

Xiulong Wu (Member, IEEE) received the B.S. degree in computer science from the University of Science and Technology of China (USTC), Hefei, China, in 2001, and the M.S. and Ph.D. degrees in electronic engineering from Anhui University, Hefei, in 2005 and 2008, respectively.

From 2013 to 2014, he was a Visiting Scholar with the Engineering Department, The University of Texas at Dallas, Richardson, TX, USA. He is currently a Professor with Anhui University. He has published about 60 articles and holds more than ten Chinese patents. His research interests include high-performance static random access memory and mixed-signal ICs.

Chunyu Peng (Member, IEEE) received the B.S. degree in communications engineering and the M.S. degree in circuits and systems from Anhui University, Hefei, China, in 2010 and 2013, respectively, and the Ph.D. degree in microelectronics and solid-state electronics.

He is currently an Associate Professor of microelectronics and solid-state electronics with Anhui University. His research interests include signal processing, analog IC design, and high-performance, and reliable memory technology.