A Pipeline Replica Bitline Technique for Suppressing Timing Variation of SRAM Sense Amplifiers in a 28-nm CMOS Process

Zhiting Lin, *Senior Member, IEEE*, Xiulong Wu, Zhi Li, Lijun Guan,
Chunyu Peng, Changyong Liu, and Junning Chen

***Abstract*— With advances in semiconductor technology, the threshold voltage variation has worsened, which has a great impact on the speed and stability of static random access memory (SRAM). This paper proposes a pipeline replica bitline (RBL) delay technique designed to reduce the timing variation of SRAM sense amplifiers. This design takes full advantage of all cells in the RBL as replica cells (RCs). A tunable pipeline structure is applied to control the discharge of groups of RCs. The structure is designed based on theoretical analysis and fabricated using an SMIC 28-nm CMOS process. The measurement results show that the delay variation can be reduced by approximately 43% and 32% compared with the conventional RBL and multistage RBL, respectively. Furthermore, with slight tuning of the normal 28-nm foundry process, four wafers were obtained under extreme conditions to comprehensively test the proposed technique. The results show that the proposed technique is more stable than other techniques in any extreme condition.**

***Index Terms*— Pipeline, replica bitline (RBL), sense amplifier enable (SAE), static random access memory (SRAM), timing variation.**

I. Introduction

IN RECENT years, mass application of static random access memory (SRAM) has occurred, and therefore, it is important to find a structure that can achieve higher speed and lower power dissipation [1]–[5]. Small cells can save area but require a longer time to discharge the large bitline (BL) capacitance. To reduce power and the cell area, sense amplifiers (SAs) are used to amplify small voltage differences between pairs of BLs. When the small voltage difference is built, whether or not the SA enable (SAE) signal is asserted properly becomes highly significant for high-speed and low-power SRAMs. If the SAE is asserted too early, the SA cannot amplify the small voltage difference correctly. In contrast, the overhead of the access time and power consumption is increased [6], [7].

To control the timing of the SAE, an inverter chain was previously used in SRAM. However, advances in semiconductor fabrication processes have led to the increasing variation of the threshold voltage (V_{th}) of transistors [8], [9]. The delay of an inverter chain and the time needed to discharge a BL do not match. Therefore, a conventional replica BL (CONV) technique was proposed in [10] and [11]. However, with the reduction in the supply voltage (VDD) and the scaling of CMOS technology, the random variation of the transistor V_{th} creates more serious consequences [12]–[14]. The V_{th} variation cannot be well tracked by the CONV any more [7], [15]. Consequently, the SRAM access time deteriorates, particularly for lower supply voltage [7], [16]–[18]. To overcome this problem, a replica bitline (RBL) with multistage (MRB) was proposed in [9] and a digitized RBL delay technique was proposed in [19] that used a new timing multiplier circuit (TMC) to generate suitable SAE timing. Subsequently, an area-efficient dual RBL delay technique was proposed in [20] and a four-transistor (4T) dual delay technique was proposed in [21].

In this paper, we propose a pipeline RBL (PRB) technique designed to achieve suitable SAE timing generation with a smaller area cost and smaller variation. The main contributions of this paper are as described as follows.

1) A pipeline control structure is proposed to control the replica cells (RCs). The PRB technique improves utilization and decreases the area because it takes full advantage of all dummy cells (DCs).
2) A detailed theoretical analysis of this scheme is presented to explain how the PRB decreases the variation and area overhead.
3) A tunable pipeline RBL is proposed that can create a relatively large range of SAE timing. The tunable delay concept can be easily integrated into the structure of the pipeline RBL.
4) With slight tuning of the normal 28-nm process, four wafers were obtained under extreme conditions to comprehensively test the proposed technique. The statistical analysis shows that the standard deviation of the SAE timing can be decreased to a lower level.

Manuscript received June 27, 2016; revised September 1, 2016 and November 2, 2016; accepted November 24, 2016. Date of publication December 26, 2016; date of current version March 3, 2017. This paper was approved by Associate Editor Vivek De. This work was supported in part by the National Natural Science Foundation of China under Grant 61474001 and Grant 61574001 and in part by the Natural Science Fund for Colleges and Universities in Anhui Province under Grant KJ2013A006.

Z. Lin, X. Wu, L. Guan, C. Peng, C. Liu, and J. Chen are with the School of Electronics and Information Engineering, Anhui University, Hefei 230601, China (e-mail: xiulong@ahu.edu.cn).

Z. Li is with Semiconductor Manufacturing International Corporation, Shanghai 201210, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2016.2634701

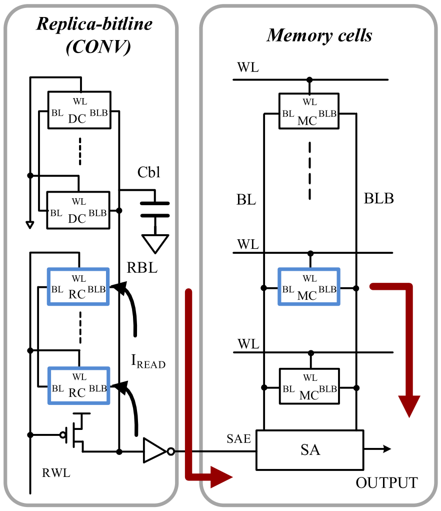


Fig. 1. Block diagram of an SRAM array with a CONV.

The remainder of this paper is organized as follows. Section II describes selected existing RBL delay techniques, calibrated delay chain techniques, and a subset of our previous studies. Section III presents the PRB technique in detail. The simulation and test results based on the SMIC 28-nm CMOS process are provided in Section IV. These results show that the PRB technique can better reduce the timing variation. Section V concludes this paper.

II. Existing Replica Bitline Technologies

Fig. 1 shows the block diagram of an SRAM memory cell (MC) array with a CONV. There are two paths in the read operation. A path is used in discharge of the BLs (BL&BLB) by the MCs, and a control path is used to generate SAE with a replica word line (WL) signal [9]. An extra BL, RCs, and DCs are used to replicate the normal BL capacitance (*Cbl*) and cell current (I_{READ}) [10], [19]. In the CONV, the number of RCs cannot be set arbitrarily because of the target SAE timing, which results in a large timing variation [7], [20].

The MRB technique was introduced in [9] to reduce the timing variation of the SAE. In this scheme, the RBL is divided into M stages with inverters inserted between any two successive stages [7]. The total standard deviation is the square root of the sum of the variances of each stage if the random variation of the delay of each stage is independent. Compared with the CONV, the standard deviation (σ) of the MRB delay is suppressed to $1/\sqrt{M}$ [9].

A digitized BL delay replica (DBDR) technique was proposed in [19] and is composed of the conventional timing replica circuit and a TMC. The TMC is used to obtain the target SAE timing. The number of RCs is K multiplied by that in the CONV, so the timing is $1/K$ of the CONVs. The TMC makes the SAE delay to be K multiplied by the RBL delay and thus suitable SAE timing can be generated. Compared with the CONV, DBDR generates the SAE timing with a small standard deviation, which is divided by $1/\sqrt{K}$ [19]. Unfortunately, with the increase in the number of the RCs, the quantization noise of the TMC increases. Moreover, the delay timing variation of the logic gates used in the TMC even becomes larger than that of the CONV in low-voltage operation [21].

To decrease the area overhead, we proposed a cascade control RBL delay (CCRBD) technique in [6]. Both RBLs are utilized and one is cascade controlled by another with an inverter. During the standby phase, both RBLs are precharged to the supply voltage, then the proposed scheme uses the left RBL (LRBL) to discharge. When the LRBL is sufficiently low to overturn the inverter, the right RBL (RRBL) begins to discharge. The standard deviation of CCRBD is theoretically reduced to $0.5\times$ that of the CONV, but the results are based on an ideal simulation environment without silicon proven.

Compared with the CONV, [16] uses four-fold RCs and makes the best use of both replica columns. The LRBL is discharged by $2j$ RCs and the RRBL is discharged through different $2j$ RCs, where j is the number of RCs in the CONV. This technique is a preliminary attempt to use more DCs as RCs. The variation is also suppressed to 1/2 compared with the CONV, but it uses two different sets of RCs, and thus it is more reliable than CCRBD. To use additional sets of RCs, we proposed a pipeline control structure in [22]. However, this approach requires a certain number of control modules that lead to area overhead. The results are also only based on an ideal simulation environment.

A 4T dual RBL delay technique was proposed in [21] to further reduce the area cost. This strategy uses a 4T cell to replace the conventional 6-transistor (6T) cell and adds another RBL. Every RC or DC has two paths with which to connect to the dual parallel RBLs. Due to the additional parallel RBL, the variation of the 4T dual RBL delay technique is suppressed to $1/\sqrt{2}$. Nevertheless, because of the doubled capacitance of the RBL, extra time is required in the precharging stage [16], [23].

The SAE timing can also be obtained using a well-calibrated delay chain. Usually, the calibrated delay chains induce extra test cost. Viveka and Amrutur [15] proposed a delay tuning algorithm based on random sampling that can significantly reduce the tester time and cost. Furthermore, they extended the operating range of the memory using tunable delay lines for timing generation in [24]. Lai and Huang [25] added an extra reconfigurable delay line on the control path. The area of the entire SRAM macro using their scheme is comparable to that of the SRAM macro that uses the traditional timing tracking.

III. Pipeline Replica Bitline

This paper proposes a technique known as PRB that can take full advantage of all cells in the RBL as RCs and makes them discharge in turn with the help of a pipeline structure.

A. Architecture of the PRB

Fig. 2 shows the structure of the proposed strategy, which consists of two control modules, named Control Modules 1 and 2, two columns of pipeline RCs, i.e., the left pipeline and right pipeline RCs, and a switch control module. Control Modules 1 and 2 alternately send signals WL1B and WLB to the switch control module, and they affect each other via the signals *xb* and *zb*. Each column of the pipeline RCs consists of *x* cells, which is set to 16 in Figs. 2 and 3. Every

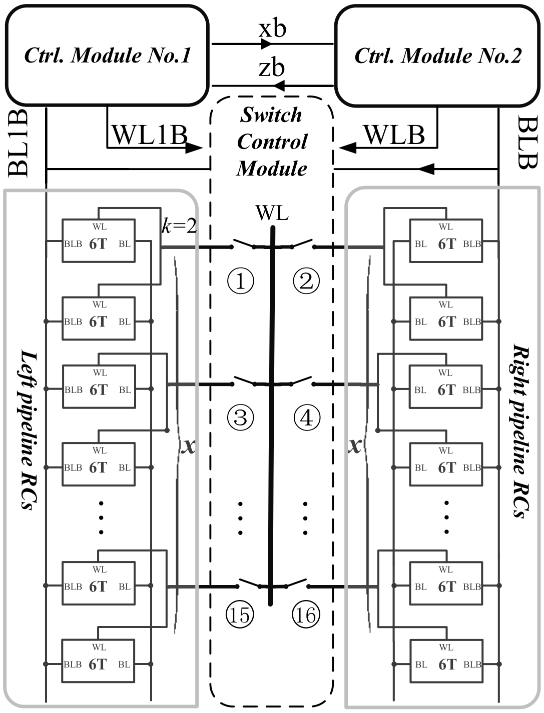


Fig. 2. Structure of the PRB.

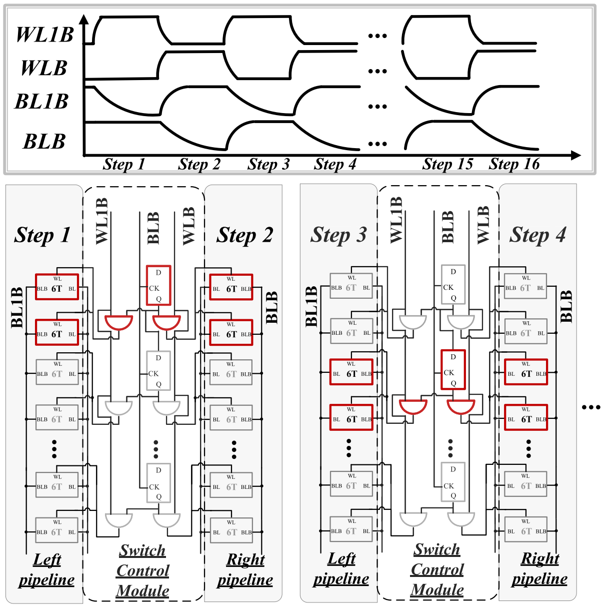


Fig. 3. Diagram of the procedure of the PRB.

k (for the sake of simplicity, $k = 2$ in Fig. 2) adjoining RCs are controlled by the same WL signal such that each column of the pipeline RCs can be divided into x/k groups. In other words, there are $2x/k$ groups in total. The switch control module receives WL1B and WLB from Control Modules 1 and 2, respectively, and subsequently chooses the proper group from the left pipeline RCs or the right pipeline RCs to discharge the BL1B or BLB. As a result, the PRB discharges $2x/k$ times.

B. Procedure of the PRB

The procedure of the PRB is detailed in this section.

1) Step 1: Control Module 1 sends the WL1B signal to the switch control module. The output of the top flip-flop in Fig. 3 is initialized to 1 and the other seven flip-flops are initialized to 0 (for simplicity, the input of the shift register of the switch control module is not shown in Fig. 3). The output of the left highlighted AND gate is high because of the high WL1B and the high output of the flip-flop. Thus, the switch control module chooses the first group of RCs of the left pipeline to discharge the BL1B. When the voltage of BL1B is sufficiently low to trigger Control Module 2, the PRB enters step 2 and Control Module 2 begins operation.

2) Step 2: Control Module 2 sends the WLB signal to the switch control module. The switch control module chooses the first group of RCs of the right pipeline to discharge the BLB, as shown in Fig. 3. At the same time, Control Module 1 charges the BL1B. The low voltage of BLB triggers Control Module 1 and controls the shift register to shift. The PRB enters step 3.

3) Step 3: The WL1B signal of Control Module 1 stimulates the switch control module to choose the second group of RCs of the left pipeline to discharge the BL1B. At the same time, the BLB is charged by Control Module 2. Subsequently, Control Module 2 is triggered by the low voltage of BL1B. The PRB enters step 4.

4) Step 4: The WLB signal of Control Module 2 is sent to the switch control module. The second group of RCs of the right pipeline is chosen by the switch control module to discharge the BLB. Control Module 1 charges the BL1B. The low voltage of BLB triggers Control Module 1 and shifts the register again. The PRB enters the next step.

The remaining steps are performed in the same manner. Finally, the last groups of RCs in the left/right pipeline alternately discharge the BL1B and BLB, and the PRB procedure is complete. Note that this structure takes full advantage of all cells as RCs, which decreases the circuit area.

Let N be the number of cells in the CONV, j be the number of RCs in the CONV, t_i be the discharging time of the i^{th} step, and t be the original discharging time. The total discharging time of the PRB should be equal to that of the CONV as follows:

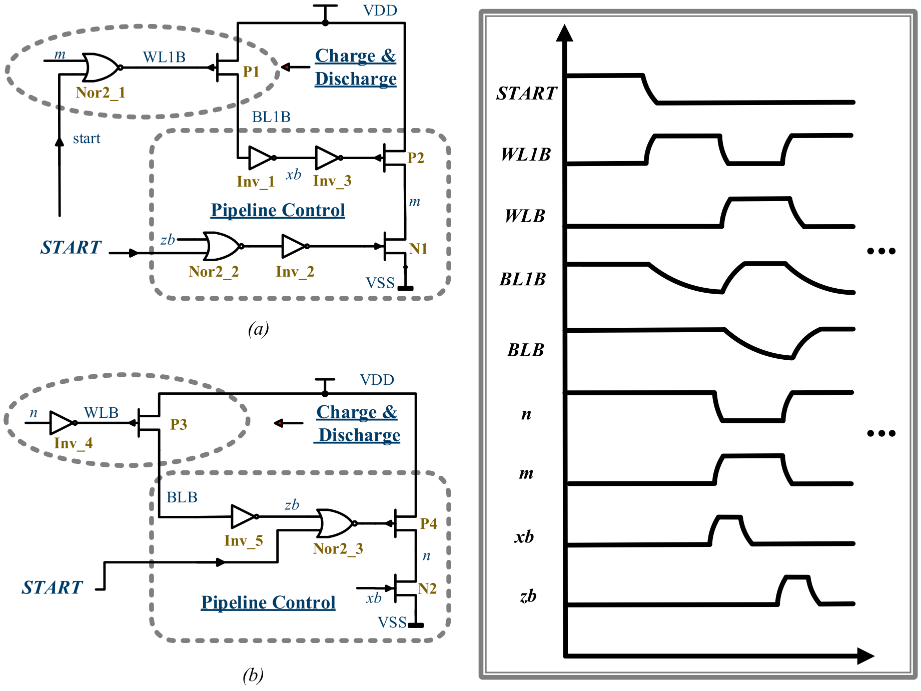
$$\sum_{\frac{2x}{k}} t_i = t \Rightarrow \left(\frac{2x}{k}\right) \times \left(\frac{x}{N}\frac{j}{k}t\right) = t \Rightarrow x = \sqrt{\frac{Nk^2}{2j}} \tag{1}$$

where $t_i = (x/N)(j/k)t$ denotes the discharging time of one step of the PRB. Assuming that the BLs are discharged by independent groups, the PRB delay variation is decreased to

$$\frac{\sigma_{\text{PRB}}}{\sigma_{\text{CONV}}} = \frac{x}{N}\frac{j}{k}\sqrt{\frac{j}{k}}\sqrt{\frac{2x}{k}} = \frac{j}{k}\sqrt{\frac{x}{N}} \tag{2}$$

compared with the conventional scheme. x/N is obtained with the use of fewer cells. $j/k(j/k)^{1/2}$ is caused by increasing the number of RCs in one discharging step. $(2x/k)^{1/2}$ is due to the accumulation of the variances of each discharging step. The number of cells used in the PRB is decreased to

$$\frac{\text{Cell}_{\text{PRB}}}{\text{Cell}_{\text{CONV}}} = \frac{2x}{N}. \tag{3}$$



(a)

(b)

Fig. 4. (a), (b) Schematics of Control Modules 1 and 2.

To test the performance of the PRB, we set $x = 16$, $N = 128$, and $j/k = 1/2$ in the test chip.

C. PRB Control Module

Fig. 4(a) shows the schematic of Control Module 1. The schematic can be divided into two submodules according to their functions. They are Charging & Discharging and Pipeline Control submodules.

1) The Charging & Discharging submodule is active in charging and discharging of the RBL and consists of a NOR and a pMOS.
2) The Pipeline Control submodule generates the signals for pipeline function and is composed of a NOR, three inverters, an nMOS, and a pMOS in Control Module 1.

Fig. 4(b) shows the schematic of Control Module 2. The schematic can also be divided into two submodules.

1) The Charging & Discharging submodule consists of an inverter and a pMOS.
2) The Pipeline Control submodule in Control Module 2 is composed of a NOR, an inverter, an nMOS, and a pMOS.

At the beginning, the start signal is high. In Control Module 1, the signal BL1B is at a high level and the signals WL1B *m*, and *xb* are low. In Control Module 2, the signals BLB and *n* are high, and the signals WLB and *zb* are low.

When the start signal changes to a low level, the signal WL1B becomes high and thus pMOS P1 is OFF. The RCs discharge BL1B. When the voltage of BL1B is sufficiently low, the signal *xb* (which is the output of the inverter Inv_1) becomes high. The signal *xb* is connected to the gate of N2 in Control Module 2, such that the signal *n* is discharged by the ground (VSS) through N2. When the signal *n* is discharged to a low level, WLB changes to a high level, P3 is OFF and the RCs discharge BLB to a low level.

When *xb* is at high level, it conducts P2 through Inv_3, signal *m* is charged by VDD to a high level and WL1B changes to a low level. Therefore, BL1B returns to a high level, which leads to a low *xb*.

When the level of BLB is sufficiently low to change *zb* to a high level, P4 is conducted through NOR2_3 and the signal *n* becomes high again. The signal *zb* is connected to one of the inputs of NOR2_2 in Control Module 1, such that the output of NOR2_2 is low. N1 is conducted, and *m* is discharged by VSS to a low level. At the same time, the signal WL1B changes to a high level again because of the low level of *m*. On the other side, the signal *n* is high, which causes the signal WLB to become low. Thus, P3 is ON and charges BLB, which leads to a low *zb*.

Control Modules 1 and 2 affect each other. The signals in both modules change one after another and return to the original state in a cycle-by-cycle manner.

D. Tunable Control Module

A tunable delay concept can be easily integrated into the structure of the PRB. Two columns of the PRB are divided into $2x/k$ groups, and every two groups are controlled by a flip-flop in the switch control model. Therefore, if we change the initial value of the flip-flops, we can control the number of groups involved in the discharge process. The total discharging time is affected by the initial values of the flip-flops. The initial value 8’b10000000 corresponds to the longest delay configuration, which activates the SA at the latest time. When the initial value of the flip-flop chain is changed from 8’b10000000 to 8’b00001000, the total discharging time is decreased to half of the original value. If a further decrease in the total discharging time is required, the initial values can be set with

TABLE I
COMPARISON OF THE DELAY VARIATION IN DIFFERENT VOLTAGES

	CONV		MRB		4T Dual		CCRBD		PRB	
VDD	σ/μ	%	σ/μ	%	σ/μ	%	σ/μ	%	σ/μ	%
0.60V	0.3625	—	0.2213	38.95	0.2422	33.19	0.1599	55.89	0.0447	87.67
0.75V	0.1763	—	0.0919	47.87	0.1246	29.33	0.0798	54.73	0.0250	85.82
0.90V	0.1098	—	0.0579	47.27	0.0807	26.50	0.0530	51.73	0.0170	84.52
1.05V	0.0834	—	0.0439	47.36	0.0611	26.74	0.0454	45.56	0.0132	84.17
1.20V	0.0682	—	0.0358	47.50	0.0500	26.69	0.0343	49.71	0.0109	84.01

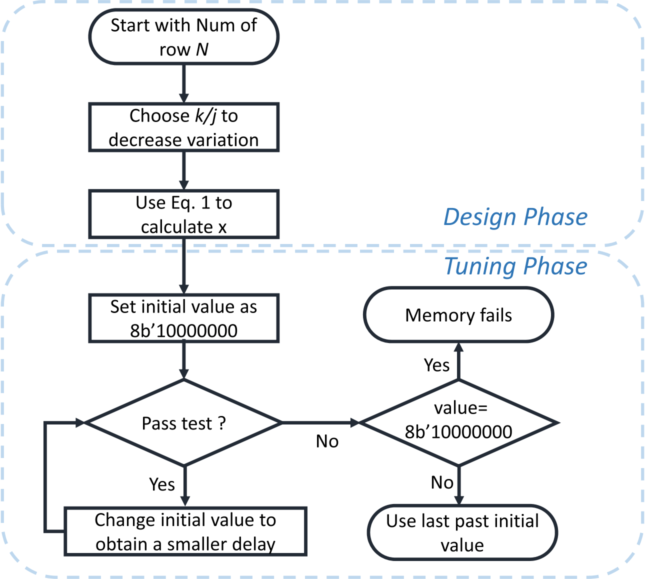

Start with Num of row *N*
Choose *k/j* to decrease variation
Use Eq. 1 to calculate x
Design Phase
Tuning Phase
Set initial value as 8b'10000000
Pass test ?
Yes
Change initial value to obtain a smaller delay
No
value= 8b'10000000
Yes
Memory fails
No
Use last past initial value

Fig. 5. Flow diagram of design and tuning phases of tunable PRB.

more than one "1," such as 8'b11000000 and 8'b01100000. In other words, the pipeline structure allows tuning to produce a relatively large range of SAE timing. The tunable PRB can further improve the performance of the SRAM.

A flow diagram, including the design phase and tuning phase, is presented in Fig. 5. Once the number of rows N is determined, k/j is selected to decrease variation and x can be obtained using (1). The tuning phase begins with an initial value that causes maximum delay. The initial value is changed according to the test results. Note that random-sampling-based tuning can be applied for larger memories [15], [24].

IV. SIMULATION AND TESTING

The proposed PRB technique is evaluated in the SMIC 28-nm CMOS process. The RC number in each stage of the MRB is equal to that of the CONV. The N of the CONV, MRB, 4T dual, or CCRBD is set to 128. We set $x = 16$ and $j/k = 1/2$ in the PRB. Fig. 6 shows the scatter diagram of the Monte Carlo simulation results of the CONV, MRB, 4T dual, CCRBD, and PRB schemes. The supply voltage is 1.05 V for all techniques. To clearly demonstrate the differences, the mean values (μ) of the SAE timing are normalized to 1 in Fig. 6. We find that the PRB performance is much better than that of the others.

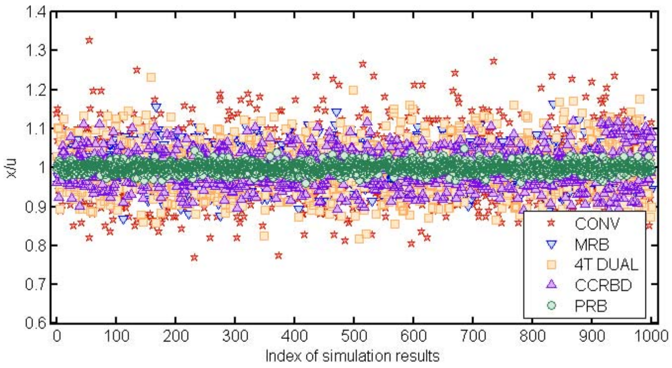


Fig. 6. Scatter diagram of the Monte Carlo simulation results.

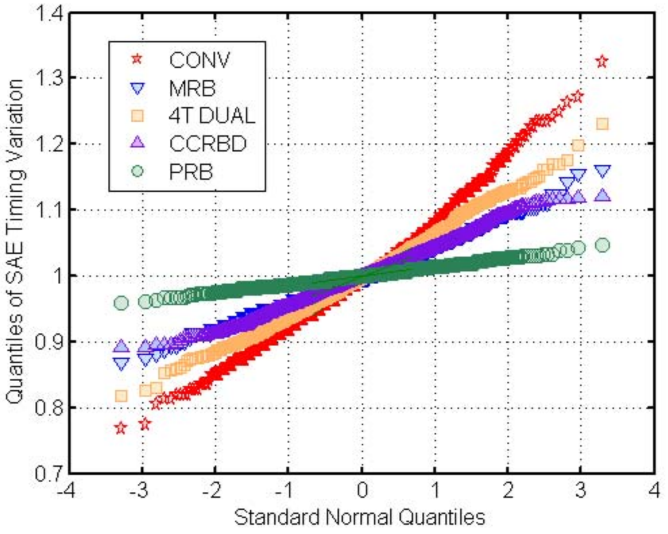


Fig. 7. Comparison of different designs in the Q–Q plot.

Table I summarizes the Monte Carlo simulation results for different supply voltages. The delay variation is measured by the coefficient of variation σ/μ, which is a standardized measure of dispersion of a frequency distribution. μ and σ are the mean value and the standard deviation of the SAE timing, respectively. When VDD is decreased to 0.6 V, the σ/μ of the PRB is 0.0447, which is much less than the CONV whose σ/μ

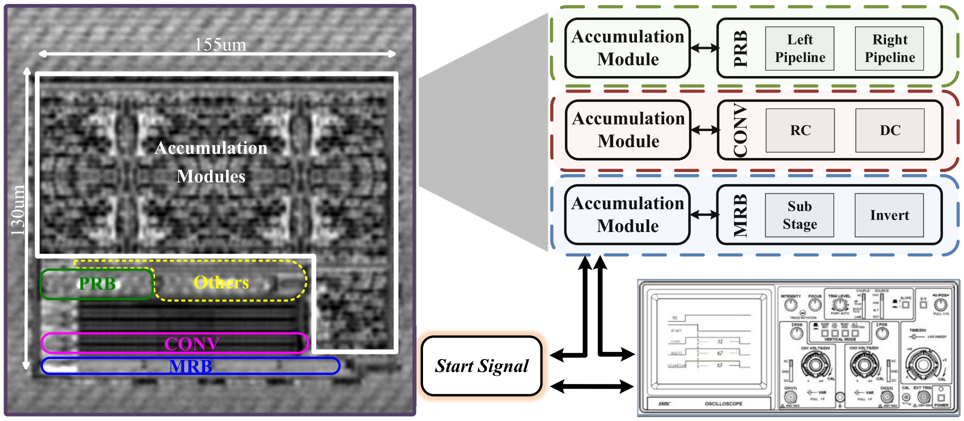


Fig. 8. Measurement setup, blocks, and area in the implemented chip.

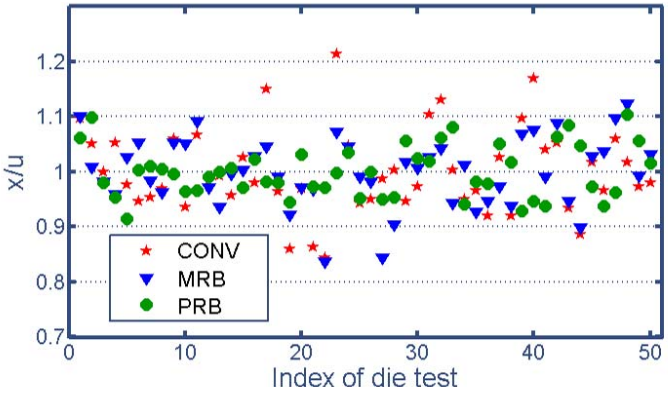


Fig. 9. Die test results in TTC.

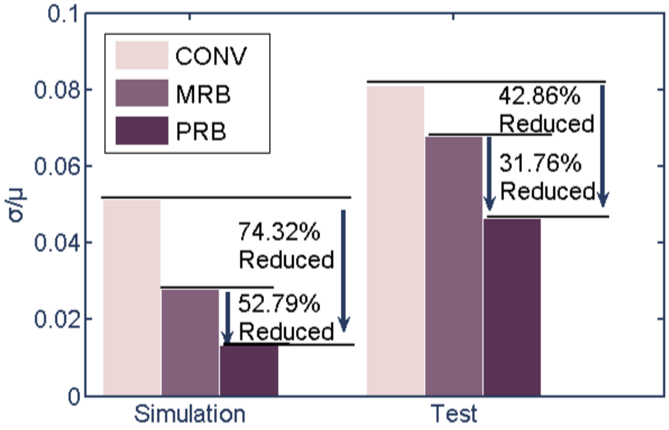


Fig. 10. Comparison of different designs in simulation and die test (TTC).

is 0.3625, the MRB whose σ/μ is 0.2213, the 4T Dual whose σ/μ is 0.2422, and the CCRBD whose σ/μ is 0.1599. The column indicated by "percentage" indicates the improvement relative to the CONV. When VDD decreases, the PRB displays better and more stable performance than the other techniques.

The area overhead of the CONV is $(4N+1)v+(2N+1)w$, that of the MRB is $(4N+M)v+(2N+2M)w$, that of the 4T dual is $(4N+1)v+2w$, that of the CCRBD is $(4N+2)v+(2N+4)w$, and that of the PRB is $(38x/k+13)v+(27x/k+17)w$, where v is the area of nMOS, w is the area of pMOS, M is the number of stages, and x/k is the number of discharging groups in an RC column. If $N=128$, $x/k=8$, and $M=4$, then the CONV requires $513v+258w$, the 4T dual requires $513v+2w$, the CCRBD requires $514v+260w$, the MRB requires $516v+264w$, and the PRB requires $317v+233w$. Compared with the CONV, the area of the PRB in our test chip is decreased by 23.5%. The RBL active period is short, so the overall power overhead is relatively small. The overall power of the MRB is almost the same as that of the CONV. Compared with the CONV, the power overhead of the 4T dual is 0.2%, that of the CCRBD is 0.3%, and that of the PRB is 1.2%.

To measure the tail of the distribution, the Q–Q plot is shown in Fig. 7, where the x-axis represents the standard normal quantiles and the y-axis represents the quantiles of SAE timing variation. The plot shows that the PRB has smaller variation up to 3 sigma. When the SAE timing margin of each RBL technique is set to the same value, i.e., three times the standard deviation of the PRB, the Pfails of other techniques are much higher. The Pfail of the CONV is 32.1%, that of the MBR is 17.7%, that of the 4T dual is 26.5%, and that of the CCRBD is 19.6%. Additionally, the Pfail of the PRB is 0.1%.

To demonstrate the efficacy of the proposed PRB design, a test chip was fabricated using the SMIC 28-nm process. The chip contains circuits of CONV, MRB, and PRB. There are 128 cells in the CONV. The traditional 6T bit cell was used due to its versatile balance of area, performance, and power. The MRB is divided into four stages, and every stage contains 32 cells. The PRB uses 16 cells in each column, which are grouped by every two adjoining cells. Fig. 8 shows the measurement setup, various blocks, and area in the implemented chip. To measure the small variation in the SAE timing, we used large parasitic capacitances and implemented accumulation modules in the chip that can accumulate thousands of discharging times. Take the CONV technique as an example, we implemented two CONV modules with an accumulation module. These two CONV modules are connected by a control module similar to the control module proposed in this paper. Therefore, these two CONV modules are able to discharge alternately. The accumulated time can be measured using an oscilloscope. In addition, the SMIC high-K metal gate 28-nm process was tuned slightly by the foundry

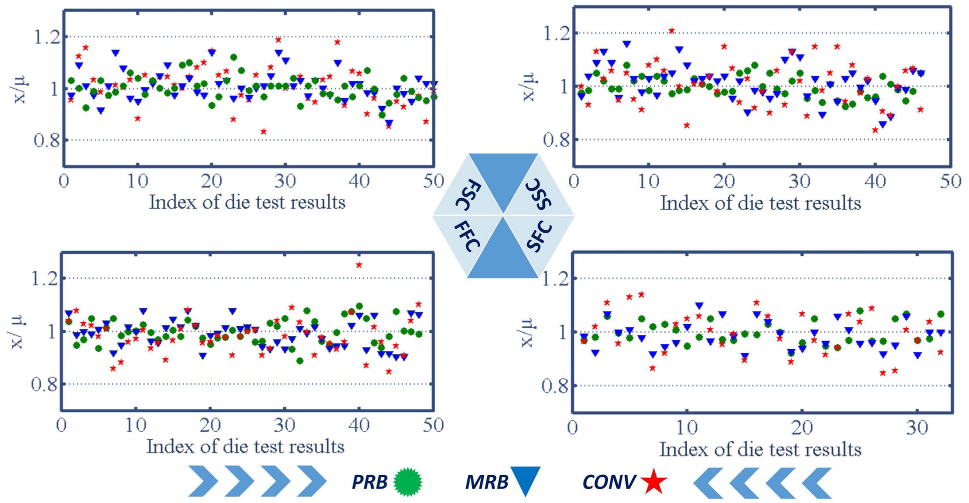


Fig. 11. Die test results in extreme conditions (FFC, FSC, SFC, and SSC).

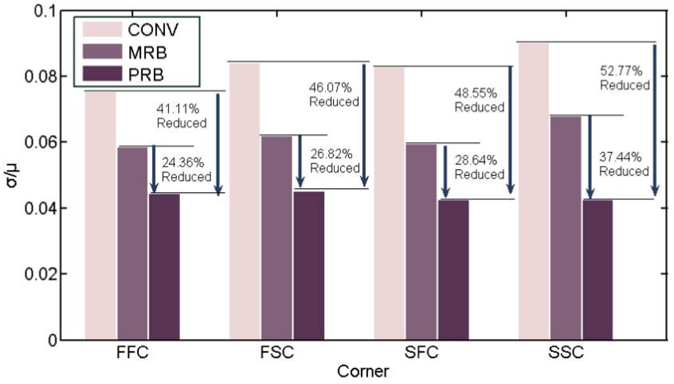


Fig. 12. Comparison of different techniques in different extreme conditions (die test).

to comprehensively test the PRB. In addition to the one wafer obtained in normal 28-nm process, referred to as "TTC," four other wafers, i.e., "FFC," "FSC," "SFC," and "SSC," were obtained by the changing doping density, oxide thickness, and other parameters.

Fig. 9 shows the scatter diagram of die test results in TTC. The mean values of different technologies are normalized to 1 to show the variation clearly. This result shows that the PRB is more stable than the CONV and MRB.

Fig. 10 presents the histogram of the simulation results and die test results in TTC. The delay variation (measured by σ/μ) of the CONV is 0.0812, that of the MRB is 0.0680, and that of the PRB is 0.0464 in the die test. The delay variation of the PRB can be reduced by 42.86% compared with the CONV and 31.76% compared with the MRB. According to Fig. 10, we conclude that when the designs are fabricated, the PRB displays a better performance than the CONV and MRB in suppressing the variation of the SAE timing, which is attributable to the random V_{th} and other process variations.

It appears that the die test results do not exactly match the simulation results. The MRB can only reduce 16.26% of the delay variation compared with the CONV, whereas it should suppress 50% of the delay variation according to the independent assumption. The PRB can reduce 42.86% in the test chips and 74.32% in the simulation. The reason for this observation is that the equations and theoretical results were obtained in an ideal environment, and in the above simulations, the simulation tool assumed that all transistors were independent of each other. In reality, a certain amount of correlation exists between them, which degrades the results. The mismatch of a certain parameter P of two identical MOS transistors of width W and length L can be modeled as [26]

$$\langle (P_a - P_b) \rangle^2 = \frac{\alpha^2}{WL} + \beta^2 D_{ab}^2, \ a \neq b \tag{4}$$

where α and β are two technology-dependent coefficients calculated from measurement data and D_{ab} is the distance between devices a and b. The mismatch is related to the distance. However, the theoretical improvement of most of the existing techniques was deduced assuming that all factors were independent and was not validated by the silicon implementation. Although the results of the die test are not as good as those of the theoretical results or the simulation results, the PRB still shows a much better performance than the other designs, because it uses all the cells in the RBL as RCs to discharge the RBL.

Fig. 11 shows the scatter diagrams of the measured results including CONV, MRB, and PRB in different conditions. When the process was changed to extreme conditions, the PRB still worked better than the MRB and CONV in suppressing the timing variation. Fig. 12 summarizes the results of the different techniques in different extreme conditions. In different conditions, the PRB still performed better. Taking the SSC as an example, the delay variation of the PRB can be reduced by 52.77% compared with the CONV and 37.44% compared with the MRB. Although the demonstration of this idea involves data from small size, all the techniques are compared under similar conditions. The PRB shows a better performance than other designs. Given that the small size has limitations, we make full use of all the RCs in the discharge to capture the variability. Table II compares this paper with other RBL techniques, where M is the number of stages and K is the multiple of the number of RCs. Most of the results from other techniques are based on an ideal simulation environment. Our

TABLE II
COMPARISON OF THIS STUDY WITH OTHER RBL DESIGNS

	Technology	Improvement	Validation
CONV RBL 1998 [9]	0.35μm	-	Test Result
MRB 2009 [8]	40nm	$\frac{1}{\sqrt{M}}$	Test Result
DBDR 2011 [20]	40nm	$\frac{1}{\sqrt{K}}$	Simulation Result
Dual-RBL 2014 [21]	65nm	$\frac{1}{\sqrt{2}}$	Simulation Result
4T Dual 2014 [22]	65nm	$\frac{1}{\sqrt{2}}$	Simulation Result
MPRDA 2014 [14]	65nm	$\frac{1}{K\sqrt{M}}$	Simulation Result
CCRBD 2015 [6]	65nm	$\frac{1}{2}$	Simulation Result
4-Fold 2015 [16]	65nm	$\frac{1}{2}$	Simulation Result
MDRBD 2015 [24]	65nm	$\frac{1}{\sqrt{2M}}$	Simulation Result
This work	28nm	$\frac{j}{k}\sqrt{\frac{x}{N}}$ (Eq.2)	Test Result

PRB technique can achieve a suitable SAE timing generation with smaller variation and area cost.

V. CONCLUSION

This paper proposes a tunable PRB technique designed to achieve lower SAE timing variation. Theoretical analysis of the structure is provided to indicate how the structure can save area and suppress the delay variation. Compared with the CONV, the time variation of the PRB is reduced by approximately 43% in TTC. The area is decreased by 23.5%. The overall power overhead is 1.2%. Although many partially correlated factors affect the performance of the PRB, which leads to differences between the measured results and simulation results, the PRB still performs better than the CONV and MRB in suppressing time variation.

REFERENCES

[1] M.-F. Chang, C.-F. Chen, T.-H. Chang, C.-C. Shuai, Y.-Y. Wang, and H. Yamauchi, "A 28nm 256Kb 6T-SRAM with 280mV improvement in VMIN using a dual-split-control assist scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 314–315.

[2] I. J. Chang, D. Mohapatra, and K. Roy, "A priority-based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 101–112, Feb. 2011.

[3] T. Fukuda *et al.*, "A 7ns-access-time 25μW/MHz 128kb SRAM for low-power fast wake-up MCU in 65nm CMOS with 27fA/b retention current," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 236–237.

[4] M.-F. Chang *et al.*, "A sub-0.3 V area-efficient L-shaped 7T SRAM with read bitline swing expansion schemes based on boosted read-bitline, asymmetric-V_{TH} read-port, and offset cell VDD biasing techniques," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2558–2569, Oct. 2013.

[5] H. Attarzadeh and M. Sharifkhani, "An auto-calibrated, dual-mode SRAM macro using a hybrid offset-cancelled sense amplifier," *Microelectron. J.*, vol. 45, no. 6, pp. 781–792, Jun. 2014.

[6] C. Peng *et al.*, "A novel cascade control replica-bitline delay technique for reducing timing process-variation of SRAM sense amplifier," *IEICE Electron. Exp.*, vol. 12, no. 5, 20150102, Feb. 2015.

[7] 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, Express Briefs*, vol. 61, no. 4, pp. 264–268, Apr. 2014.

[8] B. Nikolic *et al.*, "Technology variability from a design perspective," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 9, pp. 1996–2009, Sep. 2011.

[9] S. Komatsu, M. Yamaoka, M. Morimoto, N. Maeda, Y. Shimazaki, and K. Osada, "A 40-nm low-power SRAM with multi-stage replica-bitline technique for reducing timing variation," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2009, pp. 701–704.

[10] B. S. Amrutur and M. A. Horowitz, "A replica technique for wordline and sense control in low-power SRAM's," *IEEE J. Solid-State Circuits*, vol. 33, no. 8, pp. 1208–1219, Aug. 1998.

[11] C. D. C. Arandilla and J. A. R. Madamba, "Comparison of replica bitline technique and chain delay technique as read timing control for low-power asynchronous SRAM," in *Proc. Asia Int. Conf. Modelling Symp.*, May 2011, pp. 275–278.

[12] U. Arslan, M. P. McCartney, M. Bhargava, X. Li, K. Mai, and L. T. Pileggi, "Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2008, pp. 415–418.

[13] S. Ishikura *et al.*, "A 45 nm 2-port 8T-SRAM using hierarchical replica bitline technique with immunity from simultaneous R/W access issues," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 938–945, Apr. 2008.

[14] Y.-H. Chen *et al.*, "Compact measurement schemes for bit-line swing, sense amplifier offset voltage, and word-line pulse width to characterize sensing tolerance margin in a 40 nm fully functional embedded SRAM," *IEEE J. Solid-State Circuits*, vol. 47, no. 4, pp. 969–980, Apr. 2012.

[15] K. R. Viveka and B. Amrutur, "Digitally controlled variation tolerant timing generation technique for SRAM sense amplifiers," in *Proc. 5th Asia Symp. Quality Electron. Design (ASQED)*, Aug. 2013, pp. 233–239.

[16] W. Lu, C. Peng, Y. Tao, and Z. Li, "Efficient replica bitline technique for variation-tolerant timing generation scheme of SRAM sense amplifiers," *Electron. Lett.*, vol. 51, no. 10, pp. 742–743, May 2015.

[17] A. Kawasumi *et al.*, "A 47% access time reduction with a worst-case timing-generation scheme utilizing a statistical method for ultra low voltage SRAMs," in *Proc. Symp. VLSI Circuits (VLSIC)*, Jun. 2012, pp. 100–101.

[18] M.-F. Chang, S.-M. Yang, and K.-T. Chen, "Wide V_{DD} embedded asynchronous SRAM with dual-mode self-timed technique for dynamic voltage systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 8, pp. 1657–1667, Aug. 2009.

[19] 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.

[20] Y. Li *et al.*, "An area-efficient dual replica-bitline delay technique for process-variation-tolerant low voltage SRAM sense amplifier timing," *IEICE Electron. Exp.*, vol. 11, no. 3, 20130992, Feb. 2014.

[21] Y. D. Ye, X. L. Wu, and Z. T. Lin, “A 4T dual replica-bitline delay technique for process-variation-tolerant low voltage SRAM sense amplifier timing,” (in Chinese), *Microelectron. Comput.*, vol. 32, no. 3, pp. 28–30, Mar. 2015.

[22] Z. Li *et al.*, “Variation-resilient pipelined timing tracking circuit for SRAM sense amplifier,” *IEICE Electron. Exp.*, vol. 13, no. 7, 20150951, Jan. 2016.

[23] S.-B. Tan, W.-J. Lu, C.-Y. Peng, Z.-P. Li, Y.-W. Tao, and J.-N. Chen, “Multi-stage dual replica bit-line delay technique for process-variation-robust timing of low voltage SRAM sense amplifier,” *Frontiers Inf. Technol. Electron. Eng.*, vol. 16, no. 8, pp. 700–706, Aug. 2015.

[24] V. K. Rajanna and B. Amrutur, “A variation-tolerant replica-based reference-generation technique for single-ended sensing in wide voltage-range SRAMs,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 5, pp. 1663–1674, May 2016.

[25] Y.-C. Lai and S.-Y. Huang, “Robust SRAM design via BIST-assisted timing-tracking (BATT),” *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 642–649, Feb. 2009.

[26] M. J. M. Pelgrom, A. C. J. Duinmaijer, and A. P. G. Welbers, “Matching properties of MOS transistors,” *IEEE J. Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, Oct. 1989.

Zhiting Lin (SM’16) received the B.S. and Ph.D. degrees in electronics and information engineering from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively.

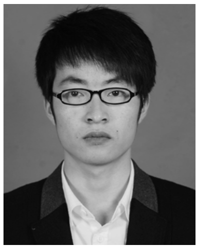
In 2011, he joined the Department of Electronics and Information Engineering, Anhui University, Hefei. From 2015 to 2016, he was a Visiting Scholar with the Electronics and Computer Science Department, Baylor University, Waco, TX, USA. He is currently an Associate Professor with Anhui University. He has authored about 40 papers and holds over 10 Chinese patents. His current research interests include pipeline ADC and high performance SRAM.

Xiulong Wu received the B.S. degree in computer science from the University of Science and Technology of China, 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 at the Engineering Department, University of Texas, Dallas, TX, USA. He is currently a Professor with Anhui University. He has authored about 60 papers and holds over 10 Chinese patents. His current research interests include high performance SRAM and mixed-signal IC.

Zhi Li received the B.S. and M.S. degrees in electrical engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2003, respectively.

In 2003, he joined Semiconductor Manufacturing International Corporation (SMIC), Shanghai, China, where he is currently the Senior Manager of the Memory Group with the Design Service Center. He is responsible for SMICs in house memory compiler and customized memory IP development. He also leads the technology qualification vehicle design support for new process technology development.



Lijun Guan received the B.S. degree in electronics and information engineering from Anhui University, Hefei, China, in 2013. He is currently pursuing the M.S. degree in electronics and information engineering at the same university.

His current research interests include low-voltage SRAM circuit design.

Chunyu Peng received the B.S. degree in communication engineering and the M.S. degree in circuit and system from Anhui University, Hefei, China, in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree in microelectronics and solid state electronics.

He is currently an Assistant Lecturer with Anhui University. His current research interests include signal processing, analog IC design, and high-performance memory technology.

Changyong Liu received the B.S. degree in microelectronics and the M.S. degree in electronic circuit and system from Anhui University, Hefei, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree.

Junning Chen received the Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 1993.

From 1993 to 1996, he held a post-doctoral position with the CAD Laboratory, Fudan University, Shanghai, China. In 1996, he became a Professor at Anhui University, Hefei, China. He has authored about 100 papers and holds 15 Chinese patents. His current research interests include the very large scale integration design and semiconductor device physics and process.

Dr. Chen is a senior member of the Chinese Institute of Electronics and the China Instrument and Control Society. He was a recipient of the first prize of the Ministry of Education Natural Science of China, and was the Vice-Chairman of the Circuit and System Society, IEEE Nanjing branch. He was the Principal Investigator of the National Science and Technology Major Projects.