A Double Sensing Scheme With Selective Bitline Voltage Regulation for Ultralow-Voltage Timing Speculative SRAM

Jun Yang, *Member, IEEE*, Hao Ji, Yichen Guo, Jizhe Zhu, Yuan Zhuang, *Member, IEEE*, Zhi Li, Xinning Liu, and Longxing Shi, *Senior Member, IEEE*

***Abstract*— A double sensing with selective bitline voltage regulation (DS-SBVR) scheme is proposed to improve the throughput of ultralow-voltage static random access memory (SRAM). It senses the bitline voltage swing twice and compares two samples for confirmation. The bitline voltage is dynamically regulated by charge sharing between two sensing steps. Different from other timing speculative SRAMs, its error flag is generated much earlier; therefore, it achieves a higher reading throughput. Meanwhile, a digitized timing scheme is proposed to generate configurable timing pulses for the DS-SBVR. Compared with other timing techniques, it has a better ability to process, voltage, temperature (PVT) tracking and variance suppression. For fair comparison of performance/power/area, three different column-based timing speculative designs are implemented in the same technology. A 28-nm test chip including 40 SRAM macros (128 × 32) is fabricated to demonstrate the scheme. Compared with the conventional design, measurements show that DS-SBVR achieves 1.45× throughput gain at 0.6-V SS corner. The figure of merit (FOM) is introduced for power, performance, and area (PPA) gain comparison. Compared with the conventional design, the FOMs of PPA gain are 1.54 and 2.33 in 128-row and 512-row memories, respectively. Compared with other timing speculative SRAMs, it achieves 1.83×–2.24× improvement.**

***Index Terms*— Double sensing (DS), static random access memory (SRAM), timing speculation, ultralow voltage.**

I. Introduction

With the increasing demand of Internet of Things (IoT) devices in the market, it is critical to decrease the power of system on chip (SoC). Reducing the supply voltage is one of the most commonly used methods. As the supply voltage approaches the threshold voltage of CMOS transistors, energy efficiency is near to the optimal point [1]. As the conventional 6T static random access memory (SRAM) active V_{ddmin} is bounded by opposing constraints on read stability and write ability [2], ultralow-voltage SRAM designs have been widely investigated in the literature. Recent works have developed new SRAM bitcell topologies [3]–[8] and various read-/write-assist techniques [9]–[17] to enable the robust operation at low voltage.

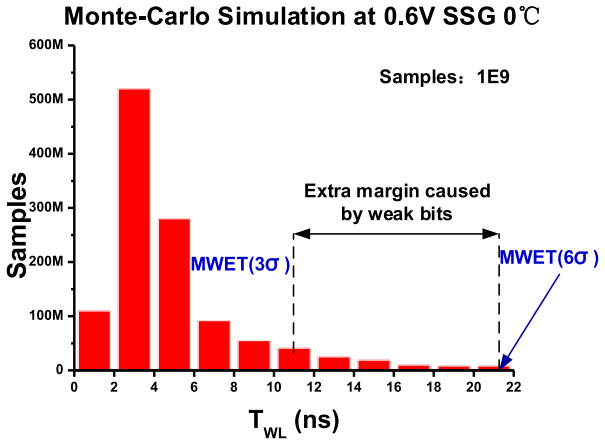


Fig. 1. Non-Gaussian distribution of read latency due to process variation.

Although these methods reduce the operating supply voltage, SRAM's performance degrades more significantly than logic in ultralow voltage. The fundamental problem is that memory bitcell has both larger process variation and higher robust requirements than logic due to its small transistor size and large capacity. When supply voltage is reduced to the near-threshold region, memory read latency T_{cq}, which is mainly composed of wordline enable time T_{wl}, exhibits a non-Gaussian distribution, as shown in Fig. 1. Minimum wordline enable time (MWET) is defined as an SRAM performance indicator, which is a function of bitline capacitance C_{bl}, sense amplifier (SA) offset voltage V_{offset}, and yields requirement, as shown in Fig. 1. For example, MWET(6σ) is the wordline enable time for correct reading of 6σ bitcell. T_{wl} is evaluated by 1×10^9 runs Monte Carlo simulation with supply voltage 0.6 V. The longest latency is 21 ns and corresponds to the weakest bitcell in memory, while most read latencies are less than 10 ns and correspond to strong bitcells. Hence, extra timing margins are required for reading these weak bits, and

Manuscript received January 20, 2018; revised March 25, 2018 and May 12, 2018; accepted May 12, 2018. Date of publication June 6, 2018; date of current version July 20, 2018. 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 61474022 and in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2015AA016601. *(Corresponding author: Yuan Zhuang.)*

J. Yang, H. Ji, Y. Guo, J. Zhu, Y. Zhuang, X. Liu, and L. Shi are with the National ASIC Research Center, Southeast University, Nanjing 210096, China (e-mail: zhy.0908@gmail.com).

Z. Li is with Semiconductor Manufacturing International Corporation, Shanghai 201203, 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.2018.2837862

the throughput of the low-voltage SRAM is limited by these weak bits.

A. Timing Speculative SRAM

The concept of timing speculation is first proposed in digital circuit to eliminate the over-design margins by *in situ* timing error detection. Ernst *et al.* [18] use the flip-flop and shadow latch to double sample input data at different edges. The main flip-flop samples at the clock positive edge, whereas the shadow latch samples at the negative edge. The error flag is triggered if these two samples are different. The scheme is often used in dynamic voltage scaling (DVS) system to reduce the voltage margin. Similar to the above method, timing error detection circuit is also used in domino register file [20] for DVS and voltage droop mitigation.

Karl *et al.* [19] apply the above idea to SRAM, which contains shadow SA in addition to main SA. The main SA is triggered speculatively at the clock negative edge. After a while, the shadow SA re-samples the bitline to confirm the result. In the normal case, the two samples are the same. As the supply voltage decreases, the error flag is generated when two sample are different. System detects the number of errors while reducing the voltage. When the number of errors exceeds the threshold, the supply voltage cannot be further reduced. This method can also improve the SRAM performance. Khayatzadeh *et al.* [21] propose a timing speculative SRAM that reads memory twice with dual ports in a pipelined method. In most cases, the read output is available after one clock cycle, and is then confirmed by comparison with a second sample in the next cycle. For weak bits, the second sample will be found to be different from the first one, and error flag will be triggered. Because of the rare weak bits, the clock frequency is greatly improved.

In summary, previous research were based on double sensing (DS), the first sensing corresponds to speculative reading, and the second sensing corresponds to confirm reading. The above methods have a common disadvantage: the interval between speculative reading and confirm reading is large and the error flag is generated too later especially in low-voltage applications, which limits its application in SoC systems.

B. Contribution

The main contributions of this paper are described as follows.

1) A DS with selective bitline voltage regulation (DS-SBVR) technique is proposed to eliminate the large design margin in ultralow-voltage SRAM. Compared with previous research, the SA double senses the bitline voltage swing continuously, in which bitline voltage is dynamically regulated. Due to the short time of voltage regulation and error detection, the error flag is generated much earlier; therefore, it is more suitable for SoC applications.
2) In order to facilitate the DS-SBVR, a digitized timing scheme is proposed to generate the configurable timing of SRAM. Compared to other techniques, its process, voltage, temperature (PVT) tracking and variation suppression ability are significantly improved.

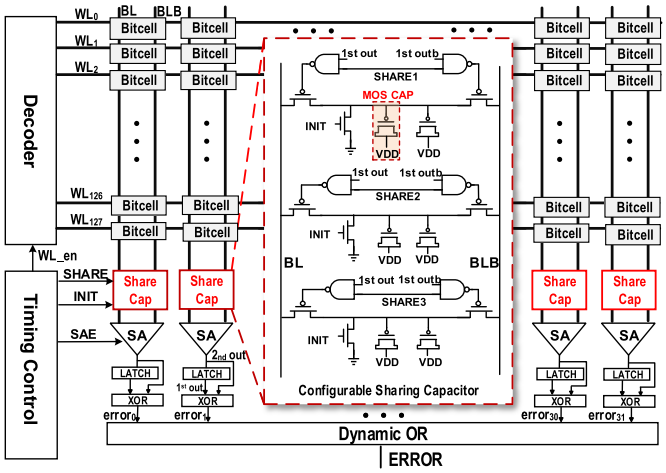


Fig. 2. Schematic of the DS-SBVR SRAM.

The remainder of this paper is organized as follows. Section II describes the proposed the DS-SBVR scheme in detail. Section III fairly compares the proposed DS-SBVR scheme with previous works. The measured results based on the 28-nm CMOS process are provided in Section IV. Section V concludes this paper.

II. Double Sensing With Selective Bitline Voltage Regulation

The architecture of DS-SBVR SRAM is shown in Fig. 2. It includes a configurable sharing capacitor, a latch style SA, an XOR logic, a dynamic latch in each column, and a dynamic OR logic across all columns. The sharing capacitor regulates the voltage of BL/BLB by charge sharing between BL/BLB and the capacitor. The dynamic latch temporarily stores data, and XOR logic compares the results in two read operations. Column's errors_i are merged by OR logic, and an error flag is generated.

A. Principle of DS-SBVR

The reading process of the DS-SBVR scheme is shown with a flowchart in Fig. 3. In conventional SRAMs, SA is enabled until MWET(6σ). In the DS-SBVR scheme, for timing speculation, a margin-less wordline enable time is applied for the risk sensing. In risk sensing, the voltage difference of BL–BLB is ΔBL_1 which is not enough for the particularly weak bits and slight risks still exist. According to the results of the first sensing, the voltage of corresponding bitline is regulated by sharing with configurable capacitor in share phase. Then, the BL–BLB voltage difference ΔBL_2 is sensed again which is called a confirm sensing.

The core idea of the bitline voltage regulation mechanism is to detect the weak bits. Fig. 3 (right) shows the relationship between V_{offset} and V_{swing} of different samples. The weak bits are the particular slow bitcells whose voltage swings are difficult to be sensed by the SA. If risk sensing result is "1," it means that $V_{\text{swing}} = V_{\text{BL}} - V_{\text{BLB}} > V_{\text{offset}}$ and the bitcell

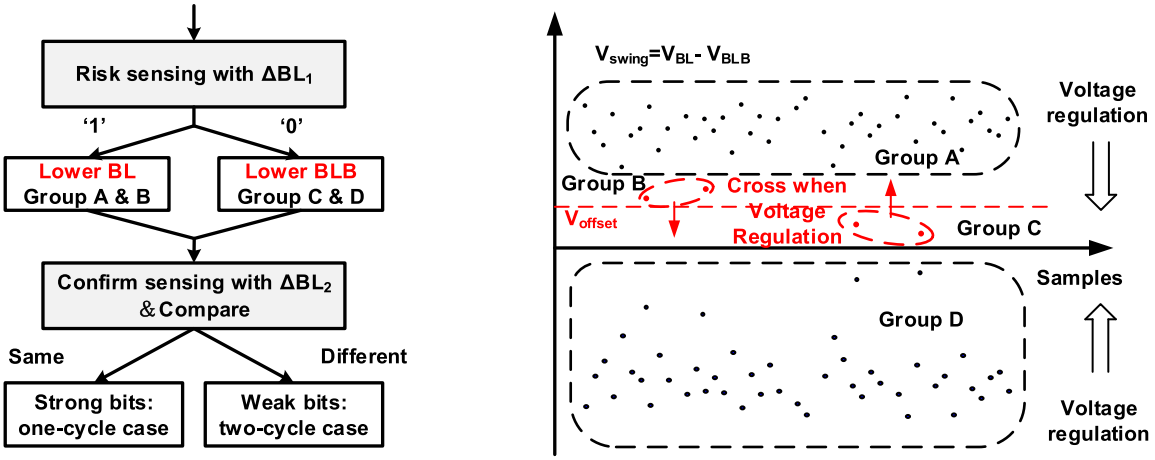


Fig. 3. Reading process of DS-SBVR SRAM.

belongs to Group A or Group B in Fig. 3. Then, the BL will be lowered by charge sharing so that the V_{swing} will be reduced. For the bitcells in Group A, the reduced ΔBL_2 is still larger than V_{offset} and the second sensing result will not change. For the bitcells in Group B, the reduced ΔBL_2 is lower than the V_{offset} and the second sensing is different with the first sensing. The bitcells in Group B are identified as weak bits. Similarly, if risk sensing result is "0," it means that $V_{swing} = V_{BL} - V_{BLB} < V_{offset}$ and the bitcell belongs to Group C or Group D. The BLB will be lowered and V_{swing} will be enlarged in this case. The bitcells in Group D are similar to the bitcells in Group A; the second sensing result is the same as that of the first risk sensing. The bitcells in Group C are similar to the bitcells in Group B. The enlarged ΔBL_2 is larger than V_{offset} and a different confirm sensing result will be obtained. In summary, for the strong bits, the voltage regulation does not change the relationship between V_{swing} and V_{offset}. However, for the weak bits, the bitline voltage swing is close to V_{offset}; therefore, the relationship between V_{swing} and V_{offset} is changed by the voltage regulation. Finally, the weak bits can be dynamically picked out by comparing the two sensing results.

As shown in the timing diagrams in Fig. 4(a), if the risk sensing is a correct prediction, BL (assuming the bitcell value is "1") is lowered and the confirm sensing will check the risk result with reduced voltage difference ΔBL_2. It can be ensured that the risk sensing is reliable if there is no discrepancy between the two samples. On the other hand, as shown in Fig. 4(b), the risk sensing is suspicious if the confirm sensing result is different from the risk sensing, and the bitcell will be identified as a weak bit. An extra cycle will be extended to ensure correct reading.

As shown in Fig. 4(c), if the risk sensing is a wrong prediction, which means that ΔBL_1 is not large enough for the SA, the BLB is lowered so that the voltage difference is enlarged to ΔBL_2 for confirm sensing. Obviously, it is easier to sense ΔBL_2 and identify the weak bits. Error reading only occurs when ΔBL_2 is still not large enough for correct reading, which will be analyzed in Section II-C in detail.

The two sensing results are latched and compared with XOR logic. The error signal is triggered if the confirm sensing

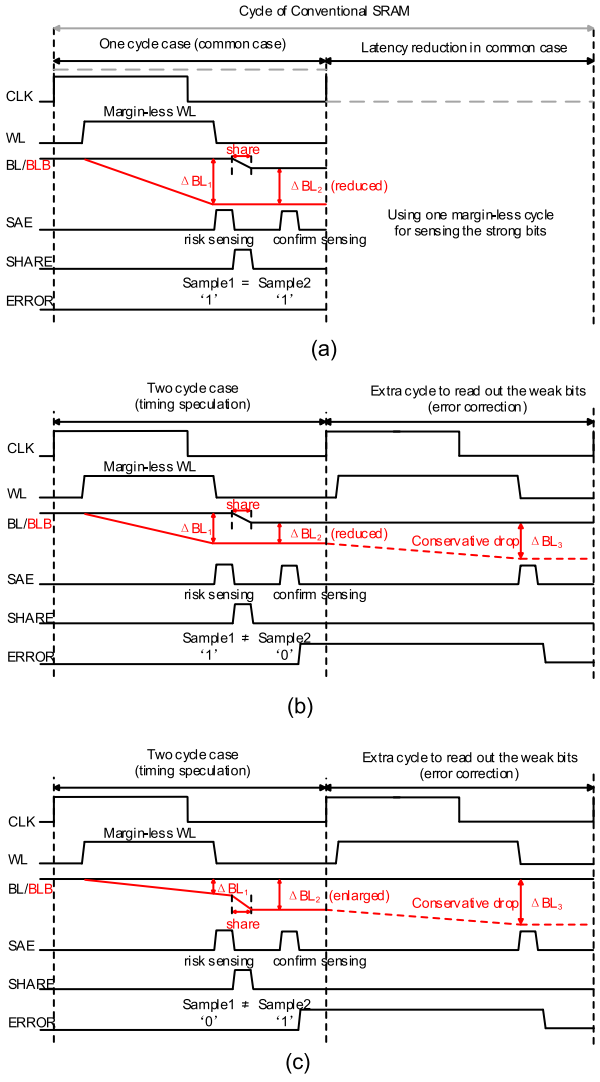


Fig. 4. Timing diagrams of DS-SBVR SRAM (assume reading "1"). (a) One-cycle case for strong bit. (b) V_{swing} is reduced when risk sensing is correct for weak bit. (c) V_{swing} is enlarged when risk sensing is wrong for weak bit.

is different from the risk sensing and the bitcell will be identified as a weak bit. The wordline and SA are enabled again, and one extra cycle is extended for sensing to ensure

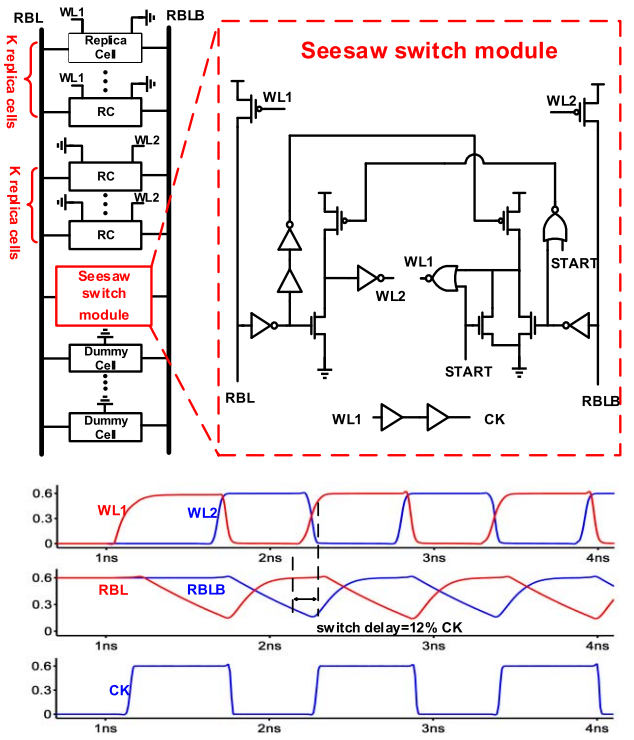


Fig. 5. Schematic and timing of digitized timing scheme.

correct reading (two-cycle cases). All column error signals are merged to gate the random logic's clock, so that the speculative memory outputs cannot affect subsequent registers if an error reading occurred. The speculative data are immediately sent out to subsequent combinational logic after the risk sensing. The weak bits are rare because they are at the tails of non-Gaussian statistical distribution. Thus, the read latency reduces significantly with a margin-less WL for most bitcells. It is particularly meaningful at near-threshold region, where local process variation causes a more conservative timing margin.

The schematic of the configurable sharing capacitor is shown in Fig. 2. In the pre-charge phase of read operation, all of the charges on the capacitor are discharged. In the share phase, signal SHARE1, SHARE2, or SHARE3 is active and the voltage of BL or BLB is regulated by charge sharing between BL/BLB and the capacitor according to the risk sensing. The magnitude of the voltage regulation is controlled by SHARE1/SHARE2/SHARE3 for test purpose.

B. Digitized Timing Scheme

An inaccurate SA enable (SAE) signal can lead to a decrease in the throughput benefit of the proposed DS-SBVR scheme. The replica bitline tracking circuit is often used in a conventional SRAM for generating an SAE signal; however, the timing of SAE also varies greatly and introduces an extra design margin at low voltage. In order to generate the configurable timing pulse for the DS-SBVR scheme, suppress timing variation to generate an accurate SAE signal and improve the PVT tracking ability of bitline discharging, a digitized timing scheme is proposed, as shown in Fig. 5.

The replica bitline consists of $2K$ replica cells (RCs) ($K = 32$ in this design), 64 dummy cells, and a switch control unit. The wordline of the RCs is split to WL1 and WL2, and the values of RCs are all fixed to high. In the reset phase, the start signal is high while WL1 and WL2 are low. The pair of replica bitlines (RBL and RBLB) are also charged. When the start signal changes to low, WL1 is enabled and RBL begins to discharge through K RCs. When RBL is sufficiently low, WL2 is enabled and WL1 is closed through the switch control unit. RBLB then begins to discharge, and RBL is charged to high at the same time. The switch control module is triggered by the low voltage of RBL or RBLB. The switch module operates like a seesaw; therefore, it is called seesaw switch logic. Finally, RBL and RBLB discharge alternately and generate signal CK. It makes the SRAM timing digitized and is able to generate a complex and configurable pulse in the DS scheme, as shown in Fig. 5. For example, the timing control signal for WL, SAE, and SHARE can be generated by the shift registers with the periodic CK.

Due to local process variation, SAE is approximately a Gaussian distribution. As shown in (2), j is the number of RCs in the conventional replica bitline scheme, k is the number of RCs in the proposed timing scheme and μ_{conv} and σ_{conv} are the mean and standard deviation of the conventional scheme. $X \sim N$ means that X conforms a Gaussian distribution. When k/j times of conventional RCs discharge simultaneously, the mean and standard deviation are divided to j/k and $j/k\sqrt{j/k}$, respectively [22]. Let X_1 and X_2 be the discharging distribution of RBL and RBLB which each discharge $M/2$ times. The proposed delay variation is decreased $j/k\sqrt{j/k} \times M/\sqrt{2}$ compared with the conventional scheme, as shown in (3). $j/k\sqrt{j/k}$ is caused by increasing the number of RCs in the discharging step. $M/\sqrt{2}$ is related to the accumulation of the variance of each discharging step

$$X_{\text{conv}} \sim N(\mu_{\text{conv}}, \sigma_{\text{conv}}) \tag{1}$$

$$X_1 \sim N\left(\frac{j}{k}\mu_{\text{conv}}, \left(\frac{j}{k}\sqrt{\frac{j}{k}}\sigma_{\text{conv}}\right)^2\right) \tag{2}$$

$$X_2 \sim N\left(\frac{j}{k}\mu_{\text{conv}}, \left(\frac{j}{k}\sqrt{\frac{j}{k}}\sigma_{\text{conv}}\right)^2\right) \tag{3}$$

$$\frac{M}{2}X_1 + \frac{M}{2}X_2 \sim N\left\{\left(\frac{Mj}{2k}\mu_{\text{conv}} \times 2\right), \left(\frac{Mj}{2k}\sqrt{\frac{j}{k}}\sigma_{\text{conv}}\right)^2 \times 2\right\} = N\left(\frac{jM}{k}\mu_{\text{conv}}, \left(\frac{M}{\sqrt{2}}\frac{j}{k}\sqrt{\frac{j}{k}}\sigma_{\text{conv}}\right)^2\right). \tag{4}$$

Fig. 6 shows the timing simulation compared with those of existing methods in 28-nm CMOS process, with the condition of SS corner and 0 °C. For fair comparison, the number of bitcells in one column is set to 128 in each method. The RC number in each stage of the multi-stage replica bitline (MRB) [24] is equal to that of the conventional replica bitline (CONV), which is "2" in this paper. The MBR stage N and the K value of the digitalized bitline delay replica (DRBR) [22] are both set to 4. The pipeline replica

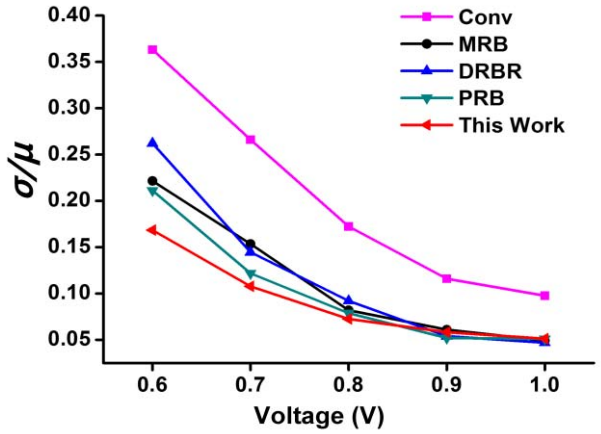


Fig. 6. Comparison between digitized timing scheme and other methods in suppressing timing variation of SAE.

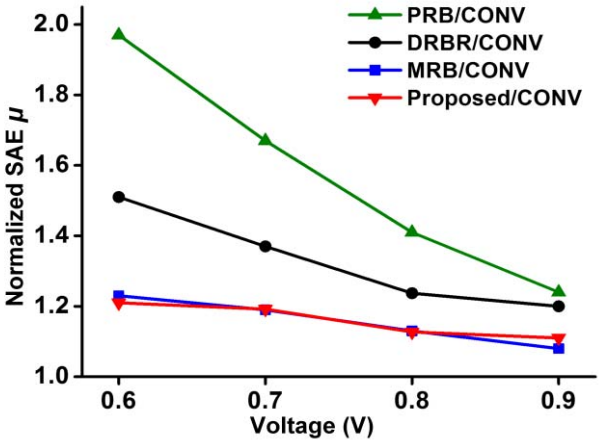


Fig. 7. Normalized mean value of SAE at different voltages.

bitline (PRB) [23] uses 16 cells in each column, which are grouped every two adjoining cells. The extra logic apart from the replica bitline will introduce extra variation in low voltage which limits the amount that discharging replica bitcells can increase. The delay of a switch operation is only 12% of the CK period, so that almost no variation is caused by the extra logic. Thus, the number of RCs (K is set to 32 in the design) can be greatly increased while the tracking ability performs better at the same time. The mean of normalized timing delay is the characterization of timing tracking ability (as shown in Fig. 7), and σ/μ represents its variation. MRB has good tracking ability, but it has a large variation due to a small number of replica bitcells. PRB has a good performance of timing variation suppression, but the tracking ability is poor in low voltage since its switching module is too complicated. Thanks to more discharging bitcells and exquisite switching logic, the proposed method in this paper shows the best tracking ability and variation suppression effect.

C. Probability Simulation of DS-SBVR

It is obvious that one-cycle case reading improves the read latency significantly and two-cycle case reading has almost no performance gain. Thus, a detailed probability analysis of the DS scheme is presented in this section to explain why the scheme is reliable and efficient.

Suppose the value of the bitcell is "1." Four possible results of the risk and confirm sensing are "11/01/10/00." Case "11" means that both ΔBL_1 and reduced ΔBL_2 are sufficiently large for the SA. Thus, as long as the voltage difference is large enough, it can be ensured that the risk sensing is reliable and only one cycle is required. Case "10" and case "01" mean that weak bits are identified by the confirm sensing with reduced or enlarged voltage difference ΔBL_2. An extra cycle will be extended to ensure the correct sensing. Case "00" is a forbidden case as amplifier cannot distinguish "0" and "1" in memory. It means that the enlarged ΔBL_2 is still not sufficiently large for the SA. The probability of case "00" is zero by configuring the suitable sharing voltage.

The result of SRAM reading is related to three variables: V_{swing}, V_{share}, and V_{offset}. V_{swing} is the voltage difference between BL and BLB. V_{share} is the voltage magnitude regulated by charge sharing. Assume that V_{share} is linearly changed in different configurations, ranging from 5 to 40 mV at 0.6 V and modeled by (5). As shown in (6), V_{swing} is proportional to cell discharge current I_{cell} and wordline enables time T_{wl} which can be configured as j cycles of T_{ck} (see Section II-B). I_{cell} is a random variable which can be simulated by Monte Carlo simulation based on the CMOS 28-nm process. V_{offset} is SA offset voltage which obeys the Gaussian distribution, and also can be obtained by Monte Carlo simulation. The three variables are modeled with MATLAB to analyze the probability of DS scheme

$$V_{\text{share}}(i) = 5i \; i \in [1:8] \tag{5}$$

$$V_{\text{swing}}(j) = \frac{I_{\text{cell}} \times T_{wl}}{C_{\text{bl}}} = \frac{I_{\text{cell}} \times T_{\text{ck}} \times j}{C_{\text{bl}}} \; j \in [1:8]. \tag{6}$$

As described in case "11," one cycle correct sensing should meet the following condition, and the corresponding bits are regarded as strong bits

$$V_{\text{swing}}(j) - V_{\text{offset}} - V_{\text{share}}(i) > 0. \tag{7}$$

Similarly, the condition of case "00" is

$$V_{\text{swing}}(j) - V_{\text{offset}} + V_{\text{share}}(i) < 0. \tag{8}$$

Supposing that variable $\#S$ is the number of strong bits and $\#W$ is the number of weak bits, the two variables can be calculated from the Monte Carlo simulation based on the MATALB models above. The probability of one cycle reading P(11) (Case "11") and the probability of error reading P(00) (case "00") are shown in the following pseudocode where w is word size.

Considering (8), $V_{\text{swing}}(j) + V_{\text{share}}(i) < V_{\text{offset}}$ is the condition to guarantee no error reading, and the probability of error reading can be ignored with the suitable configuration of i and j. Fig. 8 shows the relationship between T_{wl} and V_{share} when P(00) equals zero. The region above the line represents the situation where no error occurs. Wordline must last for more than 20 ns to guarantee no error reading in the conventional SRAM when V_{share} equals zero. T_{wl} can be reduced with the increase of V_{share}.

Algorithm 1 Monte-Carlo Simulation Pseudocode

```
mc_num ← 1000000000
w ← 32
for  j ← 1 to 8
  for  i ← 1 to 8
    for mc_index ← 1 to mc_num   do
          //random select one bit
          I_index = random(1E9); I_c = I_cell(I_index);
          //random select one SA
          SA_index =random(w); Vsa = Vs(SA_index);
          if  Vswing[I_c, j] - Vsa − Vshare[i] > 0 then
            # S ← #S + 1
          if   Vswing[I_c, j] - Vsa + Vshare[i] < 0 then
            # W ← #W + 1
    end
    P11(j, i) = (#S/mc_num)^w
    P00(j, i) = 1 − (1 − #W/mc_num)^w
  end
end
```

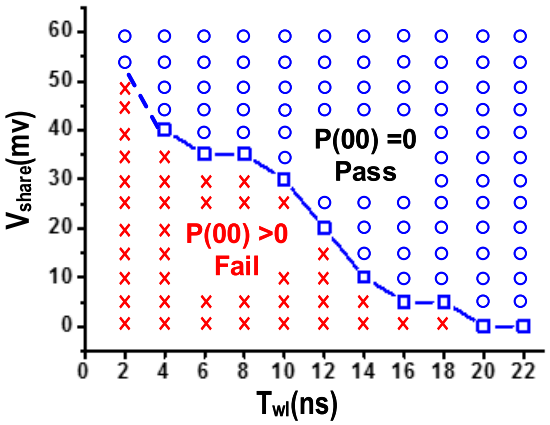


Fig. 8. Relationship between T_{wl} and V_{share} in certain read yield requirement.

For the sake of convenient analysis, we define P(10&01) as the probability of two cycles. P(10&01) equals 1−P(11)−P(00). The T_{cq} of the conventional SRAM has to be 20 ns to guarantee no error. The averaged read latency is

$$T_{\text{cq\_avr}} \approx P(11) \times T_{wl} + P(10\ \&\ 01) \times 2 \times T_{wl}. \quad (9)$$

The T_{cq} improvement is defined as

$$(T_{\text{cq\_conv}} - T_{\text{cq\_avr}})/T_{\text{cq\_conv}}. \quad (10)$$

Fig. 9 shows the T_{cq} improvement and the probability of case "11" in different T_{wl}. V_{share} is configured as the minimum value shown in Fig. 8 to guarantee P(00) = 0. When T_{wl} is small, most of the words belong to the two-cycle case which means that P(11) is near to zero. Thus, the T_{cq} improvement is significantly limited. With the increase of T_{wl}, the T_{cq} improvement reaches its peak. When T_{wl} has further increment, almost all words are read out within one cycle and the throughput approximately equals $1 - T_{wl}/T_{\text{cq\_conv}}$. Thus, the throughput benefit decreases gradually.

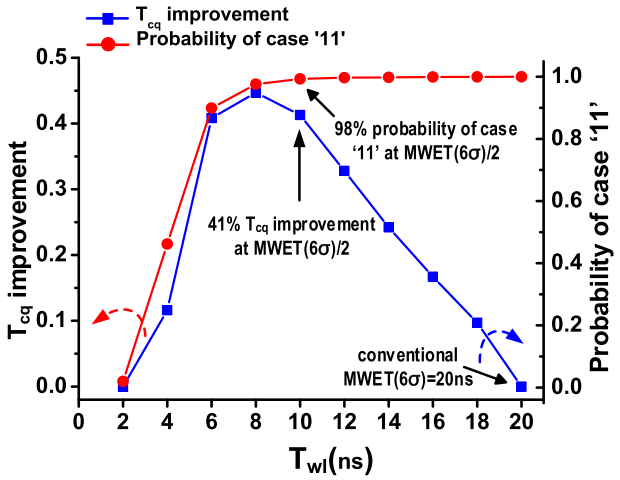


Fig. 9. T_{cq} improvement and probability of case "11" in different T_{wl}.

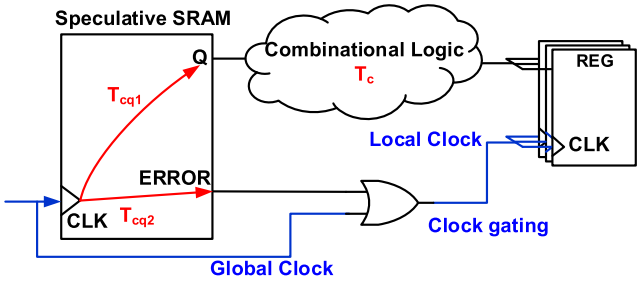


Fig. 10. Typical path in timing speculative system.

As shown in Fig. 9, T_{wl} should be configured as half of the conventional design to achieve the maximum T_{cq} improvement. V_{share} should be configured at about 35 mV to ensure no error reading with a certain design margin. P(10&01) can be statistically analyzed by the signal of ERROR in the time domain. It can be used to indicate whether the SRAM is close to the voltage–frequency limit, i.e., DVFS adjustment. In Fig. 9, P(10&01) is 2% when T_{wl} is 10 ns; that is, only two words are two-cycle cases when reading 100 words.

III. Comparisons

A. Timing Speculative System

A typical path of timing speculative system is shown in Fig. 10. Different from the conventional system, the clock of registers is controlled by the SRAM speculative signal ERROR. When detecting a suspicious SRAM reading, the following registers are clock gated to wait the correct data. From timing point of view, the speculative SRAM has two delay parameters: one is delay of speculative outputs T_{cq1} and the other is delay of confirmation T_{cq2}.

Fig. 11 shows the timing parameters in different kinds of speculative SRAMs. In the conventional SRAM, SA is enabled until the voltage difference of BL–BLB is sufficiently large. Its delay parameter is T_{cq} which is composed of time of wordline driven, BL–BLB discharging, and SA sensing. Karl *et al.* [19] release the speculative outputs at half of wordline enable time. The ideal T_{cq1} is only 50% of T_{cq}, and T_{cq2} is equal to T_{cq}.

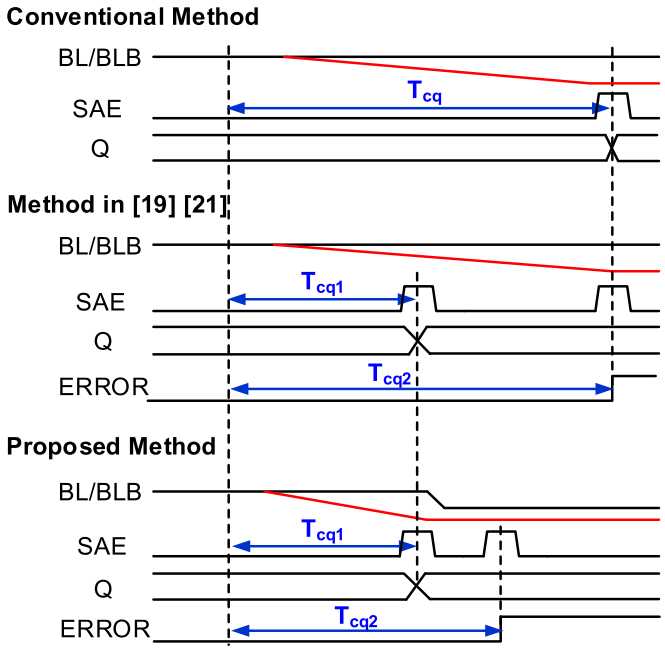


Fig. 11. Timing parameters in different speculative SRAMs.

TABLE I
TIMING CONSTRAINTS OF SPECULATIVE SRAM

	Constraint 1	Constraint 2	Max throughput gain
Paper [19]	$T_{cq1}+T_c<T$	$T_{cq2}<T$	$1+T_{cq1}/T_{cq2}$
Paper [21]	$T_{cq1}+T_c<T$	$T_{cq2}<2*T$	2
Proposed	$T_{cq1}+T_c<T$	$T_{cq2}<T$	$1+T_{cq1}/T_{cq2}$

There are two existed timing paths; one is CLK–Q–COMB–REGs and the other is CLK–ERROR–REGs. It is only suitable for logic dominant path in which logic delay T_c occupies most of clock period, but not suitable for SRAM dominant path, such as cache in processors.

The principle of Razor SRAM [21] is similar to [19]; therefore, its parameters are the same as that of [19]. In order to increase the clock frequency, it speculatively reads data through two independent ports and achieves great throughput gain at the cost of huge area overhead. The CLK–ERROR–REGs path in [21] is a multi-cycle timing path. Thus, the constraint 2. of [21] is $T_{cq2} < 2 \times T$. Thanks to the bitline voltage regulation, the confirmation in this paper is much earlier than Razor SRAM [21] and the research [19]. Thus, the ideal T_{cq2} is only a little larger than $T_{cq}/2$.

Table I summaries the timing constraints of different timing speculative SRAMs. Considering the above two timing constraints comprehensively, the highest clock frequency of the system can be obtained, which corresponds to the maximum throughput. The maximum throughput gain is defined as the ratio of the maximum throughput to that of the conventional solution. The theoretical maximum throughput gain of [21] is 2 and that of [19] and this paper is $1 + T_{cq1}/T_{cq2}$. It can be found that when T_{cq2} is closer to T_{cq1}, the throughput gain is greater.

B. Comparison of Area Overhead

Considering different implementations of peripheral circuits, such as decoders and timing circuits, it is unfair to compare the area overhead, which is the ratio of extra logic to the SRAM macro. The reading of timing speculative SRAM is column-based operation; therefore, the number of columns in one macro is not a key factor. Two types of column layout are presented to demonstrate the area overhead, in which the 128-row layout is the real implementation in the test chip, and the 512-row layout is a larger expanded implementation, which are shown in Fig. 12(a) and (b), respectively. To fairly compare their area, the layouts of one column in [19] and [21] are also realized with the same SMIC 28-nm process technology. The bitcells in the layouts including the push-rule 6T single port (SP) and 8T dual port (DP) are provided by foundry. It should be pointed out that the column MUX is not included in the layouts because of its little impact on area overhead. The bitcell array and SA including pre-charge circuit are regarded as the baseline layout.

In [19], the shadow SA, MUX, and XOR are extra logic. The area overhead is 25.7% and 7.8%, respectively, in 128-row layout and 512-row layout. Most of the area overhead is caused by the shadow SA which is already an area efficient SA. Thus, the area overhead is an optimistic estimation.

In Razor SRAM [21], DP SRAM is configured and used as a timing speculative SRAM. Therefore, there is a larger bitcell array, in which latch, MUX, and XOR are extra logic. Its corresponding area overheads are 77.64% and 79.62%, respectively, in 128-row and 512-row layouts. Most of area overhead is caused by 8T DP bitcell. (Its area is 0.315 μm^2 and that of the 6T SP bitcell is 0.155 μm^2.)

In this paper, the sharing capacitor, latch, and XOR are extra logic. In the test chip, three kinds of sharing capacitors are used for test purpose, but in the real implementation, only Configuration 3 is essential. The area overheads are 20.8% and 7.6%, respectively, in 128-row layout and 512-row layout. To achieve the same magnitude of the voltage regulation, the capacitance is proportional to the row number. Long-channel transistor is a more efficient implementation in MOS capacitor. The sharing capacitor is implemented by short-channel transistor for 128-row layout and long-channel transistor for 512-row layout. Thus, the area of 512-row sharing capacitor is 1.3× larger than that of 128 row. Other components, such as NAND gate, latch, and XOR, do not change as the increase of row number. Therefore, the area overhead is greatly reduced as the increase of the SRAM size.

C. Comparison of Power Consumption Overhead

Similar to the analysis of area overhead, the analysis of power consumption uses the same method, including 128-row and 512-row column structures. The power consumption overhead refers to the ratio of increased reading energy to that of baseline version. The reading energy consumption

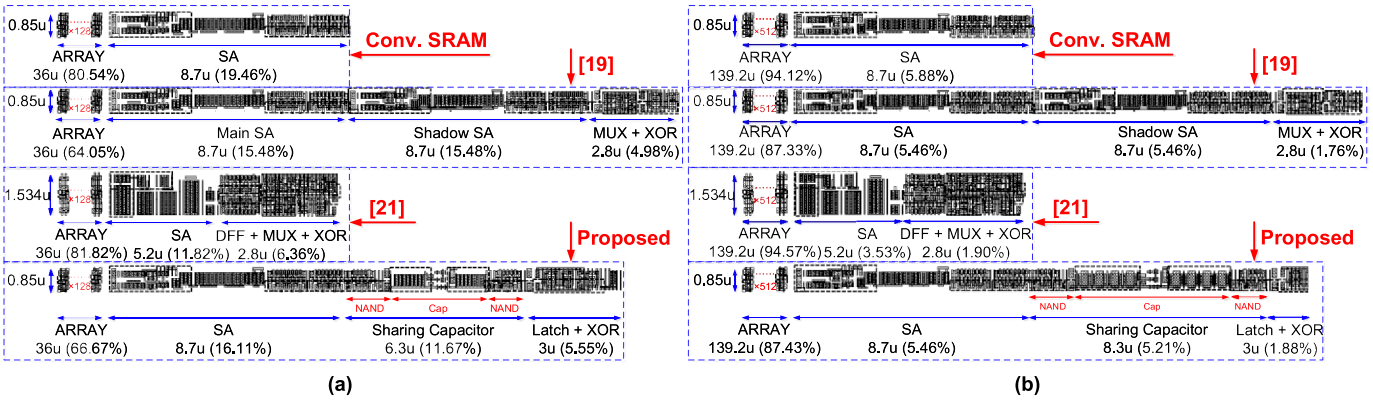


Fig. 12. Column-based layouts of (a) 128 row and (b) 512 row.

mainly includes the energy consumed by bitline discharging, error detection including voltage regulation, and the second sensing. According to the previous simulation, the T_{wl} should be larger than 20 ns at the 0.6-V SS corner to ensure the reading yield in the conventional SRAM. Due to local process variation, the bitline voltage swing in different columns is not the same.

Thus, the energy consumption in the result of this paper means the averaged energy which is estimated by 1000 times of Monte Carlo simulation. The averaged energy is shown in Fig. 13.

The conventional reading energy is about 6.33 and 19.15 fJ, respectively, in 128 row and 512 row, where 91% and 97% are consumed on BL–BLB, and the rest is consumed on the SA. It should be pointed out that most of bitlines have discharged to ground at MWET(6σ).

The read energies of [19] are 8.17 and 20.99 fJ, respectively, in 128 row and 512 row. Since the voltage swing of bitline does not change, the energy consumption of BL–BLB is the same. Due to the error detection, the energy overhead is 29% and 9.6%, respectively. The shadows SA, XOR, and MUX consume the extra energy. The principle of [21] is similar to that of [19]; therefore, its reading energy overhead is almost the same.

The reading energies of this paper are 5.35 and 13.63 fJ in most cases (corresponding to one-cycle case). Compared with the conventional SRAM, the energy consumption on the BL–BLB is reduced by about 37% due to the shorter wordline enable time. Finally, the energy reduces 15.5% and 28.8%, respectively, in 128 row and 512 row.

D. Comparison of Performance

Similar to the analysis of area overhead and energy overhead, the analysis of performance uses the same method, including 128-row and 512-row column circuits. The performance comparison which is shown in Fig. 14 mainly includes T_{cq1} and T_{cq2}.

In [19] and [21], the error detection principle is similar, which causes similar T_{cq1} and T_{cq2}. Compared to the conventional reading, T_{cq1} is almost reduced by 50%.

In the proposed scheme, T_{cq1} is almost not changed, but T_{cq2} is reduced by 36% and 43%, respectively, in 128 row and 512 row when compared with those in [19] and [21]. T_{cq2} depends on T_{cq1} and error detection delay. T_{cq1} will increase with the depth of SRAM, while the error detection delay remains almost unchanged. In 128 row and 512 row, the ratios of error detection delay to T_{cq2} are 14.8% and 4.1%, respectively. When the depth of SRAM increases, the delay of error detection can be ignored; therefore, T_{cq2} improves greatly in larger SRAM macro.

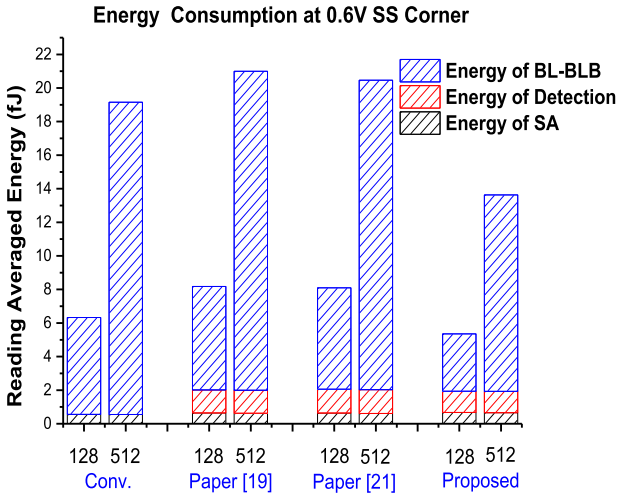


Fig. 13. Comparison of column reading energy consumption.

IV. Measurements

A. Fabrication

The SRAM test chip with DS-SBVR was fabricated in SMIC 28-nm PolySiON process. In addition to the one wafer in normal 28-nm process, referred as TT corner, two other wafers, i.e., FF corner and SS corner, were fabricated by changing doping density, oxide thickness, and other parameters. The three wafers were measured at different temperatures: 0 °C(LT), 25 °C (NT), 70 °C (HT). About 50 dies were located in different wafer regions. Each test chip included 40 SRAM macros (array size was 128 × 32); therefore, there were about 8 Mb = 128 × 32 × 40 × 50 bits in one wafer. Fig. 15 shows the die photograph and the testing logic of the chip. Apart from

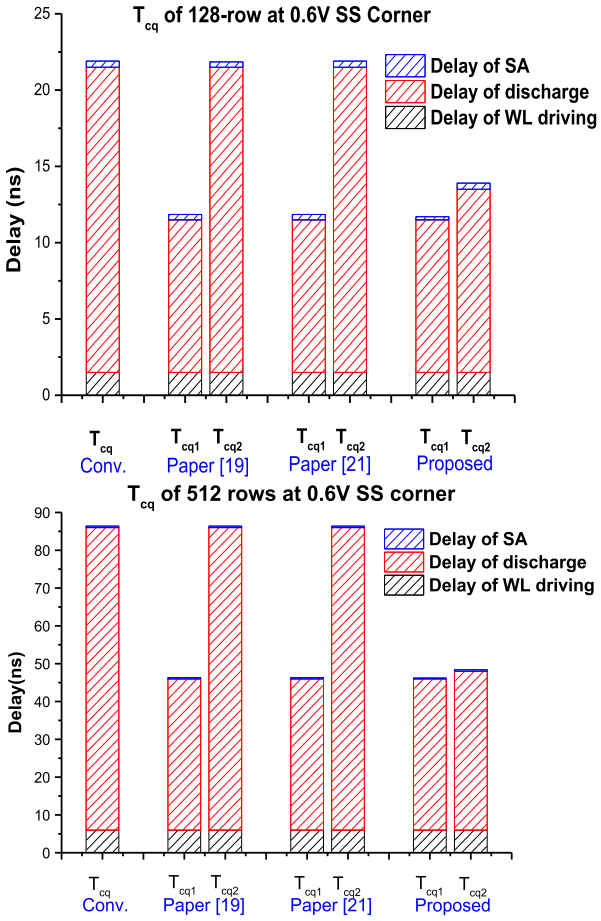


Fig. 14. Comparison of T_{cq1} and T_{cq2} in different timing speculative SRAMs.

the memory macros, a ring oscillator, digital timing modules, inverter chains, and serial peripheral interface (SPI) interface are embedded in the test chip. An inverter chain is inserted outside of each SRAM to represent the combinational logic in the path and the delay of inverter chain is tunable. With digitized timing scheme, the SRAM can work in two modes: DS-SBVR mode and conventional mode. T_{wl} and V_{share} can be configured by SPI interface and control unit. The testing harness counts the number of weak bits and error bits to calculate the throughput gain.

B. Measurements of DS-SBVR Scheme

The DS-SBVR scheme improves ultralow-voltage SRAM throughput by removing the extra design margin caused by weak bits. Thus, a die-to-die comparison with the conventional SRAM design in the same PVT condition is performed. The number of error bits in different V_{share} and T_{wl} configurations at supply voltage 0.6 V is measured with a typical die in Fig. 16. Configuration 3 has a larger voltage regulation magnitude than that of Configurations 1 or 2. As the increase of V_{share} and T_{wl}, the number of error bits gradually decreases which has a consistent correlation with the MATLAB simulation above. In the conventional mode, T_{wl} should be larger than 24 ns for the weakest bit. However, with the DS-SBVR scheme, it is reduced to 15 ns with no error bits.

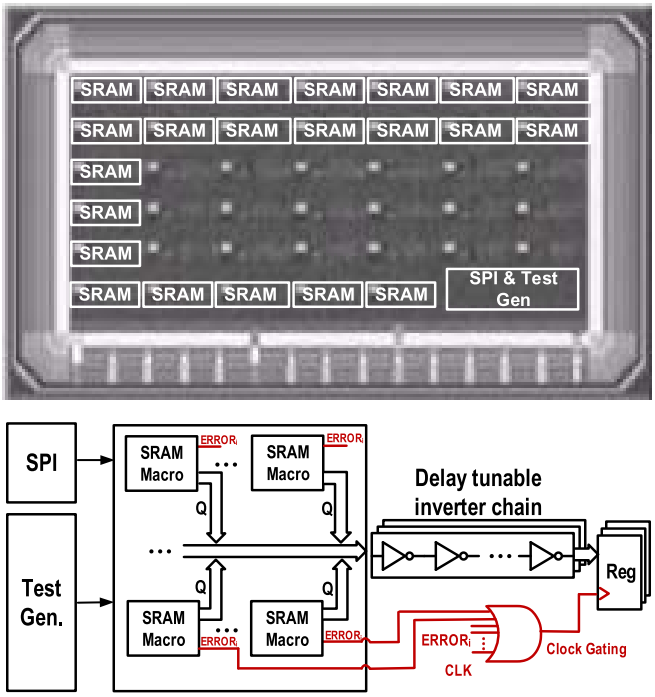


Fig. 15. Die photograph and schematic of testing logic.

C. Measurements of Digitized Timing Scheme

Fig. 17 shows the scatter diagram of measurements, including the proposed digitized timing scheme, CONV. scheme, and PRB scheme [16]. These three schemes are all fabricated in the same 28-nm process so that the comparison is meaningful. In order to compare the timing variation clearly, the mean values of different techniques are normalized to 1. For lack of measurements at low voltage in [16], measurements at 0.8 V in this paper are compared with that at 1.05 V in [16]. It shows that the digitized timing scheme performs better in suppressing timing variation although the supply voltage of the proposed scheme is lower than the others.

PVT tracking ability is also an important metric of timing unit. Fig. 18 shows the mean value of SAE in different process corners, temperatures, and supply voltages. All the timing corners are normalized to that of CONV at a supply voltage of 0.9 V, corresponding to the process corner and temperature. Considering the tracking ability of supply voltage, the maximum variation of SAE (which is the average among 40 dies) is 1.3× at FF corner, LT when supply voltage changes from 0.9 to 0.6 V. Considering the tracking ability of global process corner and temperature, the normalized timing delay variation is less than 10% in different supply voltages.

D. Performance Improvement

The DS and digitized timing scheme greatly improve the performance of low-voltage SRAM. By gradually reducing the delay of inverter chains in test chip, the maximum clock frequency can be obtained, which corresponds to the maximum throughput.

Fig. 19 shows the measurements of maximum operating frequency in different process corners at 0.6 V. In the case

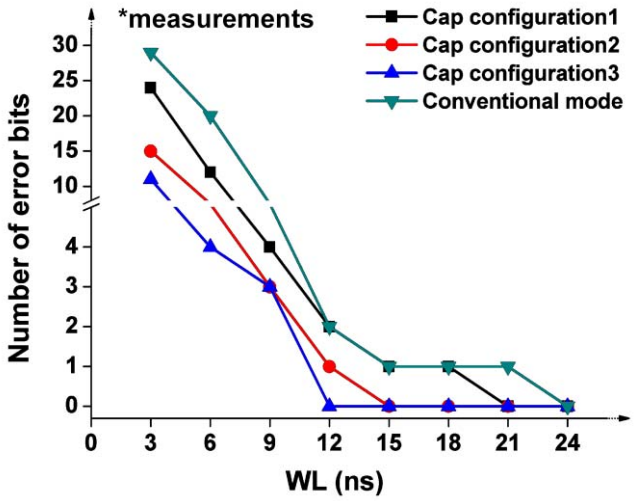


Fig. 16. Numbers of error bits in different Cap and WL enable time configurations.

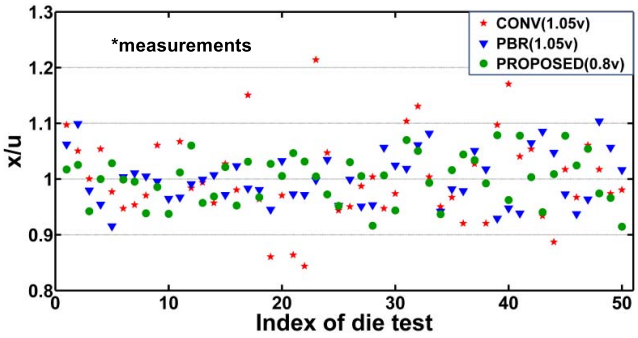


Fig. 17. Scatter diagram of different timing schemes in TT corner.

of the TT process corner and room temperature, the average clock frequency of the 40 chips is 81 MHz with DS-SBVR, which has a 31% increment when compared with that of the conventional mode. In the case of SS corner, the average clock frequency increases 46% while in the case of FF corner, the average frequency increases 29%.

When considering the performance of SoC with large capacity SRAM, the worst frequency in the wafer (8-Mb SRAM) should be used for comparison. As shown in Fig. 19, which takes TT corner as the example, the worst frequency of 8-Mb SRAM in the conventional mode is 41 MHz and the worst frequency with DS-SBVR scheme is 56 MHz. Thus, the performance improvement is about 35% in TT corner, 41% in FF corner, and 43% in SS corner.

The following is the discussion of measurements.

1) In the same PVT condition, different performance improvements are achieved in different dies due to their local process variations. The proposed DS-SBVR scheme always improves the operating frequency compared to that of the conventional method.
2) The simulation shows that the T_{cq1} and T_{cq2} are 21.8 and 13.9 ns, respectively, at 0.6-V SS 0 °C in Section III, the theoretical maximum throughput gain is 1.57×. Thus, it is consistent with the maximum throughput gain 1.45× by measurements. Similarly, the theoretical maximum throughput gains are 1.35× and 1.45× at 0.6-V FF 70 °C and 0.6-V TT 25 °C, respectively. They are all consistent with the measurements.
3) The SRAM throughput gain is the most profitable in SS corner and is the least in FF corner. The reason lies in two points: 1) design margin caused by weak bits is larger in SS corner than that in other corners and 2) DS-SBVR scheme only reduces T_{wl} which is part of T_{cq2}. The T_{wl} is ∼80% of T_{cq2} in SS corner, while ∼50% in other corners. Thus, throughput gain in SS corner is larger than that in other corners.

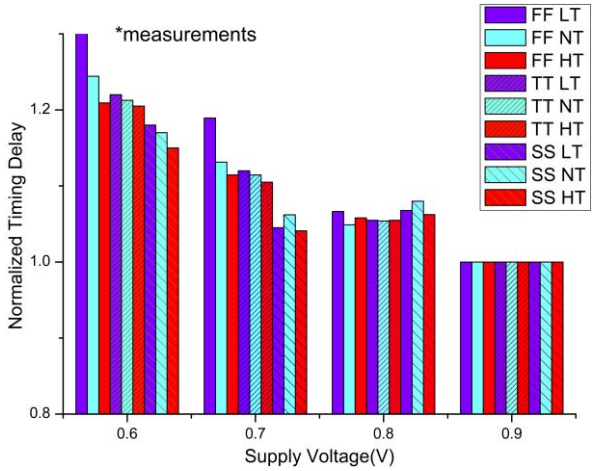


Fig. 18. PVT tracking ability of digitized timing scheme.

E. Discussion

Table II summarizes three timing speculative SRAMs with 128-row organization (represents small macro) and 512-row-type organization (represents large macro) at 0.6-V SS 0 °C corner (worst in signoff corners) in SMIC 28 nm. Compared to [19] and [21], the T_{cq2} in the proposed scheme is quite shorter due to the voltage regulation. Considering SRAM in a timing speculative system, the maximum throughput can be calculated based on T_{cq1} and T_{cq2}, as shown in Table I. The maximum throughput gains of [19] and [21] and this paper are 1.5×, 2×, and 1.57×, respectively, in 128-row SRAM, and 1.5×, 2×, and 1.78× in 512-row SRAM.

The throughput gain depends on the memory column organization, and leads to different area and power overhead. Therefore, the figure of merit (FOM) of power, performance, and area (PPA) gain is introduced for fair comparison. The FOM of PPA gain is defined as the ratio of throughput gain and (area overhead × energy overhead). As shown in Table II, the FOMs of PPA gain in [19] are 0.93 and 1.27. It shows that the throughput gain is larger than area-energy overhead in large macro, while is poor in small macro. The FOMs of Razor SRAM are 0.88 and 1.04. It shows that its PPA gain is very small. The FOMs in this paper are 1.54 and 2.33. In 128-row organization, it is 1.66× and 1.75× than those of [19] and [21], respectively. And it is 1.83× and 2.24× in 512-row organization.

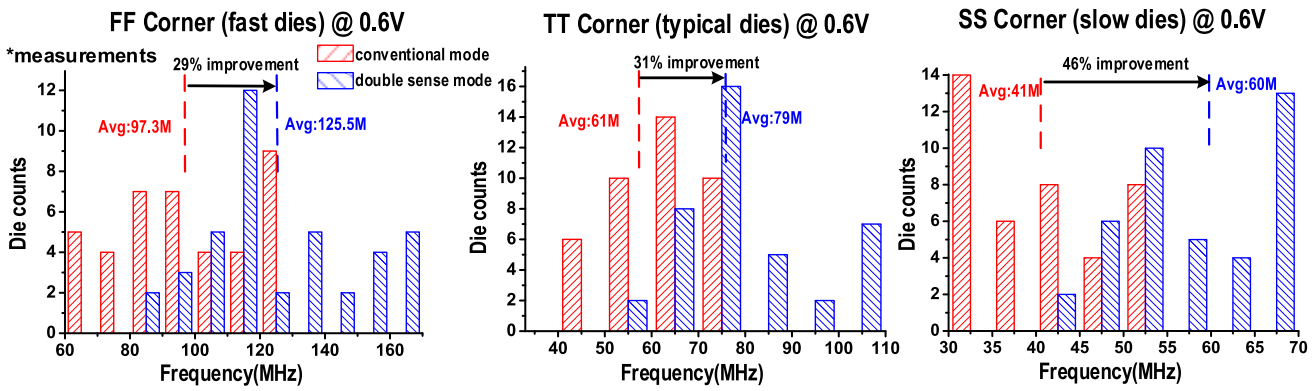


Fig. 19. Performance comparison between the conventional mode and the DS mode.

TABLE II
COMPARISON OF DIFFERENT TIMING SPECULATIVE SRAMS

		Paper [19]		Razor SRAM [21]		Proposed	
Column organization		128-Row	512-Row	128-Row	512-Row	128-Row	512-Row
Bitcell		6T SP		8T DP		6T SP	
Speculative reading method		Double sensing with main and shadow SA		Double sensing with dual ports in two consecutive cycle		Double sensing with selective bitline voltage regulation	
Area overhead		25.7%	7.8%	77.6%	79.6%	20.8%	7.6%
Energy overhead		29.0%	9.6%	27.9%	6.8%	-15.5%	-28.8%
Performance	T_{cq}/T_{cq1}	2x	2x	2x	2x	2x	2x
	T_{cq}/T_{cq2}	1x	1x	1x	1x	1.57x	1.78x
	Max. throughput gain	1.5x	1.5x	2x	2x	1.57x	1.78x
FOM of PPA gain		0.93	1.27	0.88	1.04	1.54	2.33

V. CONCLUSION

The DS-SBVR scheme is proposed to eliminate the extra design margin caused by the weak bitcells in low-voltage SRAM design. A digitized timing scheme is also proposed to generate a configurable timing pulse and suppress the timing variation of SAE signal. A 28-nm test chip including 40 SRAM macros whose size is 128 × 32 is fabricated to demonstrate the proposed scheme. Measurements show that the throughput improvement of the proposed DS-SBVR scheme is 31% at 0.6-V TT corner and 46% at 0.6-V SS corner when compared with the conventional design. Compared with the previous timing speculative SRAM, the FOM of PPA gain achieves 1.83×–2.24× improvement in 512-row organization.

REFERENCES

[1] M. Alioto, "Ultra-low power VLSI circuit design demystified and explained: A tutorial," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 1, pp. 3–29, Jan. 2012.

[2] R. Joshi *et al.*, "6.6+ GHz low Vmin, read and half select disturb-free 1.2 Mb SRAM," in *Symp. VLSI Circuits Dig.*, Jun. 2007, pp. 250–251.

[3] M.-F. Chang *et al.*, "A differential data-aware power-supplied (D^2AP) 8T SRAM cell with expanded write/read stabilities for lower VDDmin Applications," *IEEE J. Solid-State Circuits*, vol. 45, no. 6, pp. 1234–1245, Jun. 2010.

[4] J.-J. Wu *et al.*, "A large σV_{TH}/VDD tolerant zigzag 8T SRAM with area-efficient decoupled differential sensing and fast write-back scheme," *IEEE J. Solid-State Circuits*, vol. 46, no. 4, pp. 815–827, Apr. 2011.

[5] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200 mV 6T SRAM in 0.13 μm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 332–333.

[6] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 388–389.

[7] A. Raychowdhury *et al.*, "PVT-and-aging adaptive wordline boosting for 8T SRAM power reduction," in *ISSCC Dig. Tech. Papers*, Feb. 2010, pp. 352–353.

[8] M.-F. Chang, S.-W. Chang, P.-W. Chou, and W.-C. Wu, "A 130 mV SRAM with expanded write and read margins for subthreshold applications," *IEEE J. Solid-State Circuits*, vol. 46, no. 2, pp. 520–529, Feb. 2011.

[9] K. Takeda *et al.*, "Multi-step word-line control technology in hierarchical cell architecture for scaled-down high-density SRAMs," *IEEE J. Solid-State Circuits*, vol. 46, no. 4, pp. 806–814, Apr. 2011.

[10] S. Mukhopadhyay, R. M. Rao, J.-J. Kim, and C.-T. Chuang, "SRAM write-ability improvement with transient negative bit-line voltage," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 1, pp. 24–32, Jan. 2011.

[11] J. Chang *et al.*, "A 20 nm 112 Mb SRAM in high-κ metalgate with assist circuitry for low-leakage and low-V_{MIN} applications," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2013, pp. 316–317.

[12] Y. Wang *et al.*, "Dynamic behavior of SRAM data retention and a novel transient voltage collapse technique for 0.6V 32 nm LP SRAM," in *IEDM Tech. Dig.*, Dec. 2011, pp. 32.1.1–32.1.4.

[13] C.-F. Chen, T.-H. Chang, L.-F. Chen, M.-F. Chang, and H. Yamauchi, "A 210 mV 7.3 MHz 8T SRAM with dual data-aware write-assists and negative read wordline for high cell-stability, speed and area-efficiency," in *Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2013, pp. c130–c131.

[14] T. Song *et al.*, "A 14 nm FinFET 128 Mb SRAM with V_{MIN} enhancement techniques for low-power applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 158–169, Jan. 2015.

[15] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, and M. Alioto, "SRAM for error-tolerant applications with dynamic energy-quality management in 28 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 5, pp. 1310–1323, May 2015.

[16] 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.
[17] M.-F. Chang, C.-F. Che, T.-H. Chang, C.-C. Shuai, Y.-Y. Wang, and H. Yamauchi, "A 28 nm 256 Kb 6T-SRAM with 280 mV improvement in V_{MIN} using a dual-split-control assist scheme," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 314–315.
[18] D. Ernst *et al.*, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proc. MICRO*, 2003, p. 7.
[19] E. Karl, D. Sylvester, and D. Blaauw, "Timing error correction techniques for voltage-scalable on-chip memories," in *Proc. IEEE Int. Symp. Circuits Syst. Circuits Syst. (ISCAS)*, May 2005, pp. 3563–3566.
[20] J. P. Kulkarni *et al.*, "A 409 GOPS/W adaptive and resilient domino register file in 22 nm tri-gate CMOS featuring *in-situ* timing margin and error detection for tolerance to within-die variation, voltage droop, temperature and aging," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 117–129, Jan. 2016.
[21] M. Khayatzadeh, M. Saligane, J. Wang, M. Alioto, D. Blaauw, and D. Sylvester, "A reconfigurable dual-port memory with error detection and correction in 28nm FDSOI," in *ISSCC Dig. Tech. Papers*, Jan./Feb. 2016, pp. 310–311.
[22] 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.
[23] 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.
[24] 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 CICC*, Sep. 2009, pp. 701–704.

Jun Yang (M'15) received the B.S., M.S., and Ph.D. degrees from Southeast University, Nanjing, China, in 1999, 2001, and 2004, respectively.

He is currently a Professor with the School of Electronic Science and Engineering, Southeast University. He has co-authored over 50 academic papers, and holds 40 patents. His current research interests include near-threshold circuit design and Global Navigation Satellite System (GNSS) algorithm.

Hao Ji received the B.S. degree from Southeast University, Nanjing, China, in 2015, where he is currently pursuing the M.S. degree with the School of Electronic Science and Engineering.

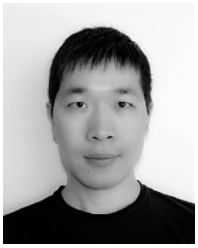
His current research interests include low-voltage static random access memory (SRAM) circuit design.

Yichen Guo received the B.S. degree from Anhui University, Hefei, China, in 2015. He is currently pursuing the M.S. degree with the School of Electronic Science and Engineering, Southeast University, Nanjing, China.

His current research interests include low-voltage static random access memory (SRAM) circuit design.

Jizhe Zhu received the B.S. degree from Southeast University, Nanjing, China, in 2015, where she is currently pursuing the M.S. degree with the School of Electronic Science and Engineering.

Her current research interests include low-voltage static random access memory (SRAM) circuit design.



Yuan Zhuang (M'16) received the B.S. and M.S. degrees from Southeast University, Nanjing, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Calgary, Canada, in 2015.

He has co-authored over 30 academic papers, and holds 11 patents. His current research interests include Internet of Things (IoT)-based asset tracking, ultralow-power IC design, and multi-sensors integration.

Dr. Zhuang received over ten academic awards.

Zhi Li received the B.S. and M.S. degrees 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 Assistant Director of the IP Development Division, Design Service Center. He is responsible for SMIC's in-house IP including memory compiler and customized memory IP development.

Xinning Liu received the B.S., M.S., and Ph.D. degrees from Southeast University, Nanjing, China, in 2000, 2003, and 2015, respectively.

He is currently a Lecturer with the School of Electronic Science and Engineering, Southeast University. His current research interests include low-power design technology and always-on circuits.

Longxing Shi (SM'06) received the B.S., M.S., and Ph.D. degrees from Southeast University, Nanjing, China, in 1984, 1987, and 1992, respectively.

From 1992 to 2000, he was an Associate Professor with the School of Electronic Science and Engineering. Since 2001, he has been a Professor and the Dean of the National ASIC System Engineering Research Center. He has authored one book and over 130 articles. His current research interests include ultralow-power IC design.