Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Multi-stage dual replica bit-line delay technique for process-variation-robust timing of low voltage SRAM sense amplifier*

Shou-biao TAN†, Wen-juan LU†, Chun-yu PENG†‡, Zheng-ping LI, You-wu TAO, Jun-ning CHEN

(*School of Electronics and Information Engineering, Anhui University, Hefei 230601, China*)

†E-mail: tsb@ustc.edu; luwenjuan@yeah.net; cyupeng@ahu.edu.cn

Received Dec. 28, 2014; Revision accepted June 7, 2015; Crosschecked July 20, 2015

Abstract: A multi-stage dual replica bit-line delay (MDRBD) technique is proposed for reducing access time by suppressing the sense-amplifier enable (SAE) timing variation of low voltage static random-access memory (SRAM) applications. Compared with the traditional technique, this strategy, using statistical theory, reduces the timing variation by using multi-stage ideas, meanwhile doubling the replica bit-line (RBL) capacitance and discharge path simultaneously in each stage. At a supply voltage of 0.6 V, the simulation results show that the standard deviations of the SAE timing and cycle time with the proposed technique are 69.2% and 47.2%, respectively, smaller than that with a conventional RBL delay technique in TSMC 65 nm CMOS technology (Taiwan Semiconductor Manufacturing Company, Taiwan).

Key words: Process-variation-robust, Sense amplifier (SA), Replica bit-line (RBL) delay, Timing variation
doi:10.1631/FITEE.1400439 **Document code:** A **CLC number:** TN43

1 Introduction

At present, the static random-access memory (SRAM) is widely used for many emerging portable electronic devices (Chang *et al.*, 2011; Gammie *et al.*, 2011). Generally, to reduce the power consumption and speed up the operation during the reading phase of SRAM, a sense amplifier (SA) is used. Fig. 1 presents the read operation of SRAM. Considering the mismatch between the cross-coupled NMOSs (N_1 and N_2, Fig. 1b) due to random dopant fluctuations (RDF) (Keyes, 1975; Pelgrom *et al.*, 1989; Johnson *et al.*, 2008), sense-amplifier enable (SAE) signal should be activated after there is a tiny differential voltage swing level that is detectable and larger than the offset voltage (V_{OS}) of SA, ΔV_{BL}, between the bit-line pair (Lovett *et al.*, 2000; Song *et al.*, 2010). The smaller the bit-line swing, the less the access time and power consumption. From this point, the accurate timing of SAE is necessary. If the SAE signal arrives earlier than the time when the differential voltage starts to exceed V_{OS}, SA will not perform the amplifying operation properly and this leads to read failure. On the other hand, if the SAE signal is asserted too late, the additional access time and extra power consumption increase unnecessarily. Hence, the optimum timing of the SAE signal has a critical impact on the design of high-speed and low-power SRAM. However, the timing of SAE is sensitive to process, voltage, and temperature (PVT) variations (Amrutur and Horowitz, 1998; Osada *et al.*, 2001; Arslan *et al.*, 2008; Komatsu *et al.*, 2009; Niki *et al.*, 2010; 2011; Arandilla and Madamba, 2011;

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 61474001)
ORCID: Chun-yu PENG, http://orcid.org/0000-0003-2408-5048

Kawasumi *et al.*, 2012; Li *et al.*, 2014; Wu *et al.*, 2014). Thus, for low-supply voltage applications, the timing variation of SAE will deteriorate.

To obtain the optimum timing with minimal slack for SAE, the replica bit-line (RBL) technique is commonly adopted (Amrutur and Horowitz, 1998; Arandilla and Madamba, 2011). In this method, the replica cells (RC), dummy cells (DC), and replica bit-line capacitance, rather than the logic gate, are employed for tracking the normal bit-lines discharge delay to obtain the suitable timing. However, with fabrication technology scaling, the threshold voltage (V_{TH}) of the transistor is more prone to shift caused by RDF. Meanwhile, the RBL technique cannot track the V_{TH} variation, which leads to read failure and increase of access time, particularly for low-supply applications.

To solve this issue, further suppression of the timing of SAE is necessary. In this paper, using statistical theory, a multi-stage dual replica bit-line delay technique is proposed, which is divided by M stages, while in each stage doubling the replica bit-line capacitance and discharge path simultaneously. Thus, the optimized timing of SAE is suitable for the low-voltage SRAM application.

2 Related RBL delay techniques

The conventional RBL technique, presented in Fig. 2, was proposed for the timing of control path tight tracking the time of the read discharge (Amrutur and Horowitz, 1998). Compared with the tradition inverter chain delay technique, the RBL delay technique uses a replica bit-line and replica cells to replicate normal bit-line capacitance and bit-cell current. In standby mode, the normal bit-lines (BL[n]) and RBL are precharged to V_{DD}. During a read operation, the word line (WL[m]) and the CLK signal are activated. RBL and the normal bit-line selected are discharged simultaneously. Finally, when the differential voltage between the worst case normal bit-line pair exceeds the V_{OS} of SA, the SAE is activated to speed up reading. Simulation results (Amrutur and Horowitz, 1998) showed that the RBL technique is more robust than the inverter chain delay technique under process variations (PVs).

The robustness characteristic of the RBL delay technique depends on having the same systematic variations. However, due to local variations, the RBL delay can be decreased, reducing the bit-line swing and causing a read failure, or increased, causing an increase in power consumption. Fig. 3 presents the effect of SAE timing variation caused by local variation of RBL on performance.

To obtain the optimum timing of SAE, the configurable replica bit-line delay (CRBD) technique (Arslan *et al.*, 2008) has been proposed, based on the conclusion that the timing variation of SAE is reduced, compared to that of a conventional RBL circuit, by using plural replica cells in the RBL column, according to Osada *et al.* (2001). The method shrinks the timing variation of SAE by $\sim 14\times$. However, the CRBD technique increases the implementation costs because additional post-silicon tests are

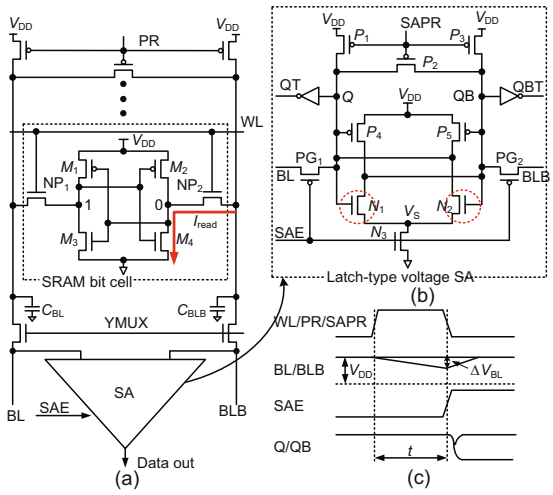

Fig. 1 Read operation of SRAM: (a) block diagram of the read operation path; (b) conventional voltage latch-type SA; (c) signal waveforms during the read operation

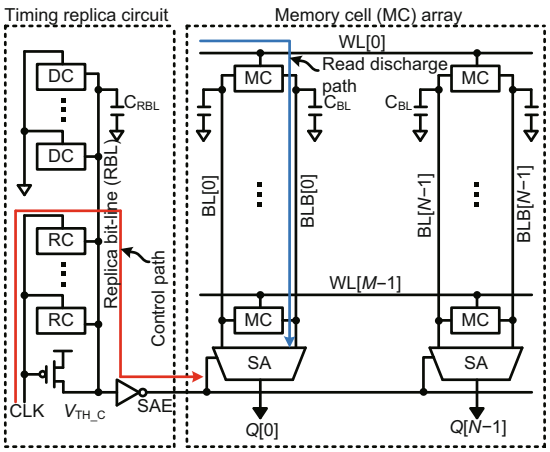

Fig. 2 Block diagram of the SRAM array with conventional RBL delay circuit (RC: replica cell; DC: dummy cell; MC: memory cell)

required. As a further improvement, a multi-stage replica bit-line delay (MRBD) technique (Fig. 4a), based on statistical principles has been developed in Komatsu *et al.* (2009). The MRBD technique divides the RBL into several sub-RBLs, while keeping the RCs in each stage the same as that in conventional RBL. As a result, the theoretical total standard deviation of timing variation, σ, is divided by $\sqrt{M}$ (M is the number of the stages) compared to that of conventional RBL. However, with the increase of M, the mismatch between global RBL and normal bit-line becomes obvious owing to the gate delay of the inverters inserted in every two stages. From this, a digitized replica bit-line delay (digitized-RBD) technique (Fig. 4b) was proposed by Niki *et al.* (2010; 2011). The digitized-RBD technique uses K times RCs in the RBL column compared to that of conventional RBL to obtain the standard deviation of RBL delay variation divided by $K\sqrt{K}$. Then the goal timing variation of SAE is multiplied by K for using the time multiply circuit (TMC) to guarantee that the delay time is the same as that of normal bit-line. That is to say, the theoretical goal standard deviation of timing variation is divided by $\sqrt{K}$ compared to conventional RBL. The disadvantage of this technique is that, as the RCs count increases, the quantified error caused by the delay unit used in TMC becomes larger, and there is an increase in total area overhead. Another technique, called multiple-stage parallel replica bit-line delay (MPRBD) (Fig. 4c) (Wu *et al.*, 2014), whose essence is K multiple RBLs

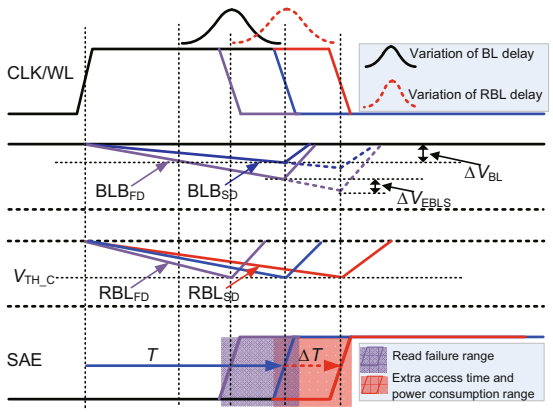

Fig. 3 Effect of SAE timing variation related to local variation of RBL on performance (ΔV_{BL}: differential voltage swing between the bit-line pair; ΔV_{EBLS}: extra bit-line swing; BLB_{FD} and BLB_{SD}: fastest and slowest discharging situations for BLB, respectively; RBL_{FD} and RBL_{SD}: fastest and slowest discharging situations for RBL, respectively)

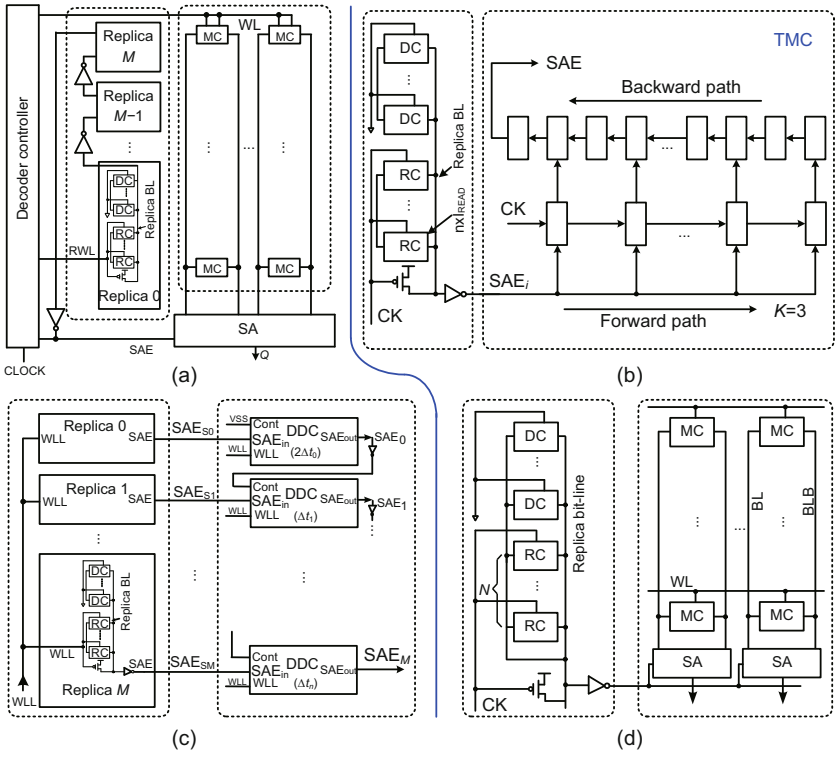

Fig. 4 Existing replica bit-line techniques: (a) MRBD; (b) digitized-RBD; (c) MPRBD; (d) DRBD

being divided into $M \times K$ stages, faces the same problems as the digitized-RBD technique. Considering the trade-off between area overhead and performance, the two sets of replica bit-line (main and reservoir) delay technique is employed (Kawasumi *et al.*, 2012). To further improve the area overhead, the dual replica bit-line delay (DRBD) technique (Fig. 4d) has recently been proposed by Li *et al.* (2014). Without extra area overhead, this design proposal improves the target standard deviation of timing variation of the SAE by $1/\sqrt{2}$ times. However, the capacitance of the dual-RBL is doubled by connecting the two sides of the replica column. Consequently, compared to normal bit-line, the charging time of dual-RBL is doubled, which leads to the increase of access time. In addition, the new replica cell proposed by Li *et al.* (2014) cannot well simulate the real 6T bitcell owing to the destruction of cross coupling.

3 Principle and structure of the proposed MDRBD technique

According to Arslan *et al.* (2008), the scaling of conventional RBL delay variation (σ/μ) is a function of the number of RCs, which is described by

$$\frac{\sigma_n/\mu_n}{\sigma_0/\mu_0} = \frac{1}{\sqrt{n}}, \tag{1}$$

where σ_0 and μ_0 are the standard deviation and mean of timing variation of conventional RBL, respectively. The distinctions σ_n and μ_n are those of RBL with n times RCs compared to the conventional one. Owing to the n times RCs being activated (i.e., I_{read} of RBL increases by n times), μ_n is equal to μ_0/n. Keeping the capacitance (C_{RBL}) invariant, σ_n is approximately equal to $\sigma_0/(n\sqrt{n})$ (i.e., the theoretical basis of $k\sqrt{k}$ directly given in Niki *et al.* (2010; 2011)). However, the time delay of RBL is directly proportional to the ratio of $C_{\mathrm{RBL}}/I_{\mathrm{read}}$. Therefore, keeping I_{read} invariant (i.e., the number of RCs is invariant) and C_{RBL} being multiplied by m, σ_m is approximately equal to $m\sigma_0$ (i.e., $C_{\mathrm{RBL}}/I_{\mathrm{read}} \sim (\mu_0, \sigma_0^2)$, $mC_{\mathrm{RBL}}/I_{\mathrm{read}} \sim (m\mu_0, (m\sigma_0)^2)$), according to Komatsu *et al.* (2009). Furthermore, to ensure that the mean of timing variation is the same as that of the conventional RBL, extra circuits such as TMCs or accumulative circuits are necessary.

According to the aforementioned fundamental principle, the principle of the proposed MDRBD technique shown in Fig. 5 could be explained as follows: first, using twice the number of RCs compared to conventional RBL, the standard deviation and mean of timing variation are $\sigma_0/(2\sqrt{2})$ and $\mu_0/2$, respectively. Then they become $\sigma_0/\sqrt{2}$ and μ_0 owing to C_{RBL} multiplied by two on both sides. Further, divide the RBL into several sub-RBLs (i.e., M stages novel dual replica bit-lines, DRBs), while keeping the RCs in each stage the same as those before division. Those of DRB become $\sigma_0/(M\sqrt{2})$ and μ_0/M. Finally, accumulating the M stages using inverters, they become $\sigma_0/\sqrt{2M}$ and μ_0.

Fig. 6 presents the distribution comparison between different techniques, where σ is the standard deviation of the timing variation of DRB used in the proposed MDRBD technique. Compared with the conventional RBL, the standard deviation of the timing variation of MDRBD is divided by $\sqrt{2M}$.

The proposed MDRBD technique solves the problem of the double charging time necessary in the DRBD technique using the multi-stage idea. Moreover, to achieve the same effect, $2M$ stages are necessary according to the MRBD technique, which means

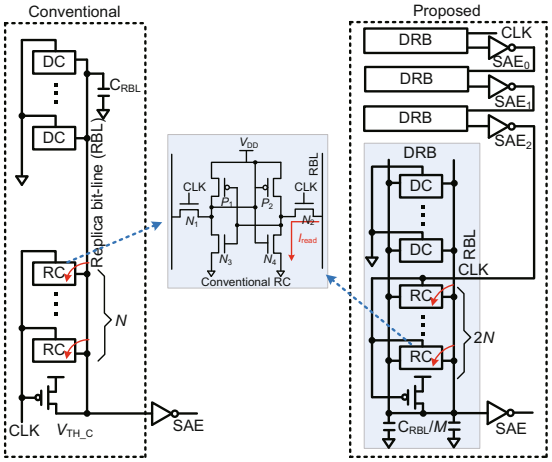

Fig. 5 Block diagram of the conventional RBL and proposed MDRBD circuits

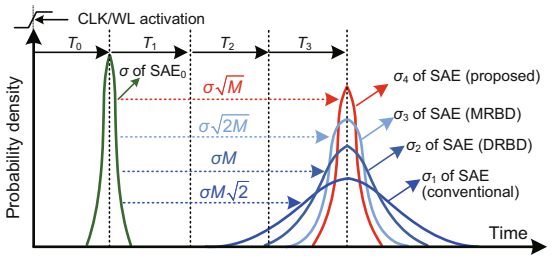

Fig. 6 Relations of distribution of the SAE timing among different techniques

$2M$ inverters are necessary. However, only M inverters are necessary in the proposed MDRBD technique owing to, in each stage, doubling the replica bit-line capacitance and discharge path simultaneously. Consequently, to a certain degree, the mismatch between global RBL and normal bit-line owing to the gate delay of the inverters decreases.

4 Simulation results and discussion

Fig. 7 presents the standard deviation of SAE timing with different stages using MRBD and the proposed MDRBD technology. Keeping the same stage, the standard deviation of SAE timing of the RBL with the proposed MDRBD technology will be reduced approximately by $1/\sqrt{2}$, compared with that of the RBL with the MRBD technology. However, with theoretically the same improvement, the standard deviation of SAE timing of the RBL with the proposed MDRBD technology is better than that of the RBL with the MRBD technology owing to the half inverters utilized (Table 1).

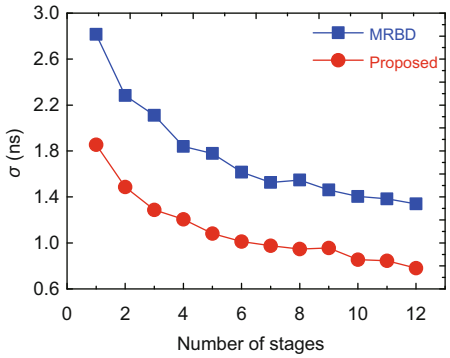

Fig. 7 Standard deviation of SAE timing with different numbers of stages, using MRBD and the proposed MDRBD technique, respectively (i.e., 0.8 V, SS, −40 °C, $N = 2$)

As shown in Fig. 8, a comparative study, in the condition of 0.6 V supply voltage, slow-slow (SS) corner, −40 °C, and 128-row memory cells, is made among the conventional RBL, MRBD, DRBD, and the proposed MDRBD based on the TSMC 65 nm CMOS technology (Taiwan Semiconductor Manufacturing Company, Taiwan). Here, RBLs are divided into four stages for MRBD and MDRBD, and $N = 2$. The Monte Carlo simulation results show that the proposed MDRBD has the lowest standard deviation σ (8.4 ns), resulting in improvements of 69.2%, 45.1%, and 11.6% compared with those of the conventional RBL, DRBD, and MRBD, respectively. Generally, without the timing margin, the desired access cycle time is double the SAE timing delay (Niki *et al.*, 2010; 2011; Li *et al.*, 2014). Considering random V_{TH} variation, a large timing margin is necessary. Assuming three times standard deviation for the SAE timing margin, the conventional and proposed access timing margins are 163.8 ns and 50.4 ns (i.e., $3\sigma \times 2$), respectively. The timing margin is reduced by about 113.4 ns owing to a 69.2% reduction of the SAE timing variation. Consequently, the access time is reduced by 227 ns. Thus, a 47.2% cycle time improvement is expected using the proposed MDRBD technique.

Table 1 Comparisons of the number of stages divided, inverter used, and standard deviation of SAE timing between MRBD and the proposed MDRBD technique*

Number of stages		Number of inverters used		σ (ns)	
MRBD	MDRBD	MRBD	MDRBD	MRBD	MDRBD
2	1	2	1	2.283	1.855
4	2	4	2	1.840	1.486
6	3	6	3	1.617	1.287
8	4	8	4	1.547	1.205
10	5	10	5	1.405	1.081
12	6	12	6	1.340	1.012

* With theoretically the same improvement, 0.8 V, SS, −40 °C, $N = 2$

Fig. 9 shows the standard deviation of these current techniques and the proposed design with different supply voltages. The simulation condition remains the same except for the supply voltage (i.e., SS,−40 °C). When the supply voltage varies from 0.6 V to 1.0 V, the standard deviation of the timing variation of the proposed design is reduced by about 63.2% to 69.2%, 44% to 48%, and 6.7% to 13.3% compared with those of the conventional RBL, DRBD, and MRBD, respectively. In particular, at the supply voltage of 0.6 V, the standard deviation is decreased by 69.2% compared with that of the conventional RBL, which indicates that the proposed design is more effective in low-supply applications.

Fig. 10 shows the standard deviation of these current techniques and the proposed design with a different process corner. Keeping the supply voltage at 0.6 V and the temperature at −40 °C, the variation tendency of the standard deviation in different RBL techniques shows that the standard deviation of the proposed design is reduced by 59.2% to 69.2%,

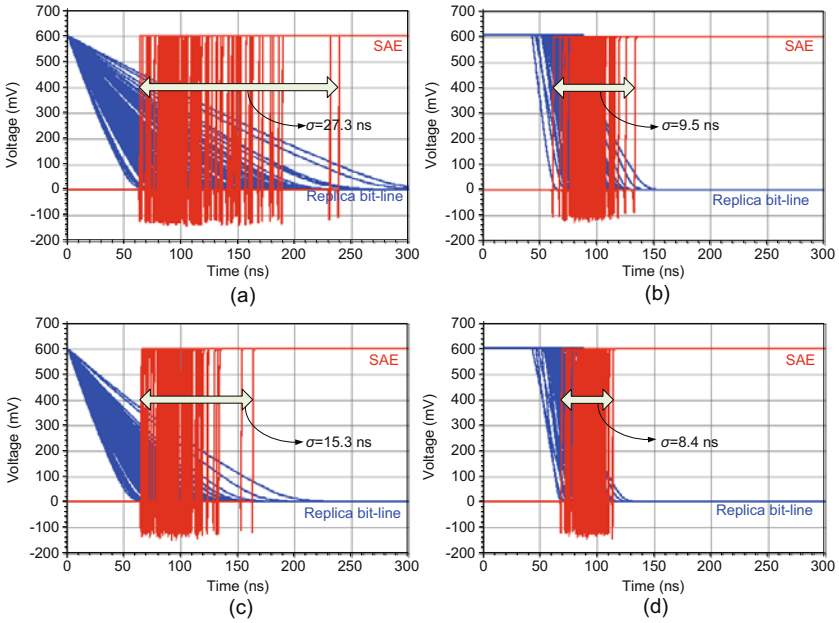

Fig. 8 Standard deviation comparison of SAE timing in the worst case (i.e., 0.6 V, SS, −40 °C): (a) conventional RBL; (b) MRBD; (c) DRBD; (d) proposed MDRBD

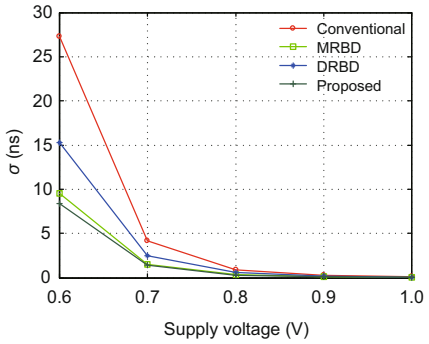

Fig. 9 Standard deviation comparison of SAE timing with different voltages (i.e., SS, −40 °C)

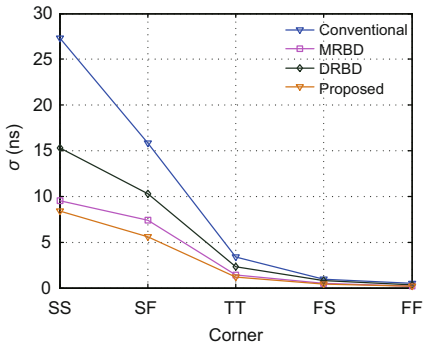

Fig. 10 Standard deviation comparison of SAE timing with different process corners (i.e., 0.6 V, −40 °C)

45.1% to 49.8%, and 11.6% to 24.3% compared with those of the conventional RBL, DRBD, and MRBD, respectively. Therefore, the proposed MDRBD is more robust under process variation compared with the other techniques.

Keeping the supply voltage at 0.6 V and the process corner at SS, the standard deviation with changing temperature is as presented in Fig. 11. As can be seen from the results, as the temperature changes from −40 °C to 125 °C, the standard deviation of the proposed design is reduced by 62.8% to 69.2%, 40.5% to 45.1%, and 10% to 12.8% compared with those of the conventional RBL, DRBD, and MRBD techniques, respectively.

Table 2 shows the comparison between different timing strategies. As mentioned above, the proposed technique requires the same area overhead as MRDB. In terms of power consumption, there is an increase of 5.67% with respect to DRBD.

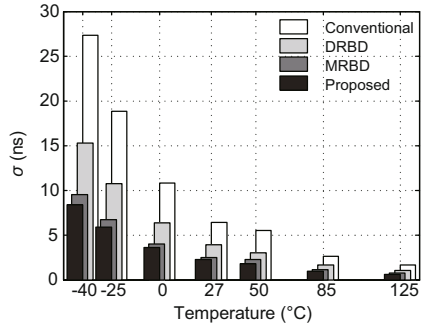

Fig. 11 Standard deviation comparison of SAE timing with different temperatures (i.e., 0.6 V, SS)

Table 2 Comparison of different timing strategies for SA (SS, −40 °C)

Strategy	Power consumption (nW)				Area
	0.6 V	0.8 V	1.0 V	1.2 V	overhead
Conventional	89.82	120.6	153.8	190.0	–
MRBD	98.70	133.6	170.9	210.1	Yes
DRBD	175.90	234.7	293.6	354.3	No
Proposed	185.20	247.7	310.8	375.5	Yes

5 Conclusions

A multi-stage dual replica bit-line delay technique has been proposed to further optimize the SA timing of low-supply SRAM. By using multi-stage and dual replica bit-line, the proposed MDRBD technique, to some degree, decreases the mismatch between global RBL and normal bit-line owing to the gate delay of inverters. Simulation results indicate that the proposed design can obtain the smallest SAE timing variation in TSMC 65 nm CMOS technology, particularly at a low supply voltage. At 0.6 V, SS process corner, and −40 °C, the standard deviation in the proposed design has decreased by 69.2% relative to that of the conventional scheme, and the cycle time has reduced by 47.2%.

References

Amrutur, B.S., Horowitz, M.A., 1998. A replica technique for wordline and sense control in low-power SRAM's. *IEEE J. Sol.-State Circ.*, **33**(8):1208-1219. [doi:10.1109/4.705359]

Arandilla, C.D.C., Madamba, J.A.R., 2011. Comparison of replica bitline technique and chain delay technique as read timing control for low-power asynchronous SRAM. Proc. 5th Asia Modelling Symp., p.275-278. [doi:10.1109/AMS.2011.58]

Arslan, U., McCartney, M.P., Bhargava, M., *et al.*, 2008. Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines. Proc. IEEE Custom Integrated Circuits Conf., p.415-418. [doi:10.1109/CICC.2008.4672108]

Chang, I.J., Mohapatra, D., Roy, K., 2011. A priority-based 6T/8T hybrid SRAM architecture for aggressive voltage scaling in video applications. *IEEE Trans. Circ. Syst. Video Technol.*, **21**(2):101-112. [doi:10.1109/TCSVT.2011.2105550]

Gammie, G., Ickes, N., Sinangil, M.E., *et al.*, 2011. A 28 nm 0.6 V low-power DSP for mobile applications. Proc. IEEE Int. Solid-State Circuits Conf., p.132-134. [doi:10.1109/ISSCC.2011.5746251]

Johnson, J.B., Hook, T.B., Lee, Y.M., 2008. Analysis and modeling of threshold voltage mismatch for CMOS at 65 nm and beyond. *IEEE Electr. Dev. Lett.*, **29**(7):802-804. [doi:10.1109/LED.2008.2000649]

Kawasumi, A., Takeyama, Y., Hirabayashi, O., *et al.*, 2012. Energy efficiency deterioration by variability in SRAM and circuit techniques for energy saving without voltage reduction. Proc. IEEE Int. Conf. on IC Design & Technology, p.1-4. [doi:10.1109/ICICDT.2012.6232859]

Keyes, R.W., 1975. Effect of randomness in the distribution of impurity ions on FET thresholds in integrated electronics. *IEEE J. Sol.-State Circ.*, **10**(4):245-247. [doi:10.1109/JSSC.1975.1050600]

Komatsu, S., Yamaoka, M., Morimoto, M., *et al.*, 2009. A 40-nm low-power SRAM with multi-stage replica-bitline technique for reducing timing variation. Proc. IEEE Custom Integrated Circuits Conf., p.701-704. [doi:10.1109/CICC.2009.5280731]

Li, Y., Wen, L., Zhang, Y., *et al.*, 2014. An area-efficient dual replica-bitline delay technique for process-variation-tolerant low voltage SRAM sense amplifier timing. *IEICE Electron. Expr.*, **11**(3):1-6. [doi:10.1587/elex.11.20130992]

Lovett, S.J., Gibbs, G.A., Pancholy, A., 2000. Yield and matching implications for static RAM memory array sense-amplifier design. *IEEE J. Sol.-State Circ.*, **35**(8):1200-1204. [doi:10.1109/4.859510]

Niki, Y., Kawasumi, A., Suzuki, A., *et al.*, 2010. A digitized replica bitline delay technique for random-variation-tolerant timing generation of SRAM sense amplifiers. Proc. IEEE Asian Solid State Circuits Conf., p.1-4. [doi:10.1109/ASSCC.2010.5716633]

Niki, Y., Kawasumi, A., Suzuki, A., *et al.*, 2011. A digitized replica bitline delay technique for random-variation-tolerant timing generation of SRAM sense amplifiers. *IEEE J. Sol.-State Circ.*, **46**(11):2545-2551. [doi:10.1109/JSSC.2011.2164294]

Osada, K., Shin, J., Khan, M., *et al.*, 2001. Universal-VDD 0.65-2.0 V 32 KB cache using voltage-adapted timing-generation scheme and a lithographical-symmetric cell. Proc. IEEE Int. Solid-State Circuits Conf., p.168-169. [doi:10.1109/ISSCC.2001.912589]

Pelgrom, M.J.M., Duinmaijer, A.C.J., Welbers, A.P.G., 1989. Matching properties of MOS transistors. *IEEE J. Sol.-State Circ.*, **24**(5):1433-1439. [doi:10.1109/JSSC.1989.572629]

Song, T., Lee, S.M., Choi, J., *et al.*, 2010. A robust latch-type sense amplifier using adaptive latch resistance. Proc. IEEE Int. Conf. on IC Design and Technology, p.182-185. [doi:10.1109/ICICDT.2010.5510258]

Wu, J., Zhu, J., Xia, Y., *et al.*, 2014. A multiple-stage parallel replica-bitline delay addition technique for reducing timing variation of SRAM sense amplifiers. *IEEE Trans. Circ. Syst. II*, **61**(4):264-268. [doi:10.1109/TCSII.2014.2304893]