Comparative Analysis on Replica Techniques for Bit-Line Tracking in 14-nm node

Se-Hyeok Oh[1], Han-wool Jeong[2] and Seong-Ook Jung[3]
[1, 2, 3] Department of Electrical and Electronic Engineering, Yonsei University, Republic of Korea
50 Yonsei-ro Seodaemun-gu, Seoul, 03722, Republic of Korea
E-mail: [1]im5sh@yonsei.ac.kr, [2]hanwool87@yonsei.ac.kr, [3]sjung@yonsei.ac.kr

Abstract: Replica bit-line technique is used for generating sense amplifier enable signal (SAE) to track the bit-line delay of static random access memory (SRAM). However, threshold voltage variation in the replica bit-line circuit changes the current in the replica bit-cell, which results in variation of the SAE time, T_{SAE}. The variation of T_{SAE} makes the sensing operation unstable. In this paper, in addition to conventional replica bit-line delay (RBL_{conv}), dual replica bit-line delay (DRBD) and multi-stage dual replica bit-line delay (MDRBD), which are used for reducing T_{SAE} variation, are briefly introduced, and the optimal number of on-cells, which can minimize bit-line delay with satisfying 6σ sensing yield, is determined through simulation at a supply voltage of 0.6 V with 14 nm FinFET technology. As a result, it is observed that delay of DRBD (MDRBD) is improved by 24.4% (48.3%) compared with that of RBL_{conv} and energy consumption of DRBD (MDRBD) is reduced by 8% (32.4%) compared with that of RBL_{conv}.

Keywords—**Replica bit-line technique, T_{SAE}, 14nm FinFET, delay, energy consumption**

1. Introduction

As technology scales down, the variation in process, temperature and voltage (PVT) becomes severe [1]. This PVT variation significantly degrades the stability of digital circuit operation. In particular, the static random access memory (SRAM) is highly sensitive to the PVT variation because SRAM cell is typically composed of minimum sized transistors. Especially, the timing variation of control signals in the SRAM, such as a sense amplifier enable (SAE) signal, can degrade the performance and stability of SRAM operation.

In SRAM, SAE determines the moment when the sense amplifier captures the bit-line voltage difference. It is crucial to generate SAE appropriately because the total performance, power consumption and operational yield of SRAM are determined by this SAE.

In order to properly generate SAE with less variation, replica bit-line technique is widely used. [2]-[5]. Because the replica bit-line circuit is composed of the replica bit-cells which have the identical structure with the data bit-cell, T_{SAE} can track the bit-line delay. In this paper, various previously proposed replica bit-line (RBL) techniques are compared in aspects of sensing yield, energy consumption, and read delay.

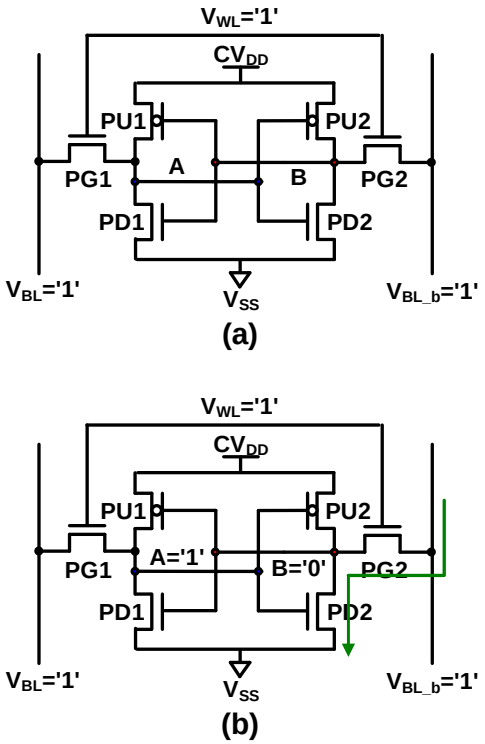

Fig. 1. (a) 6T memory cell schematic of SRAM (b) read operation

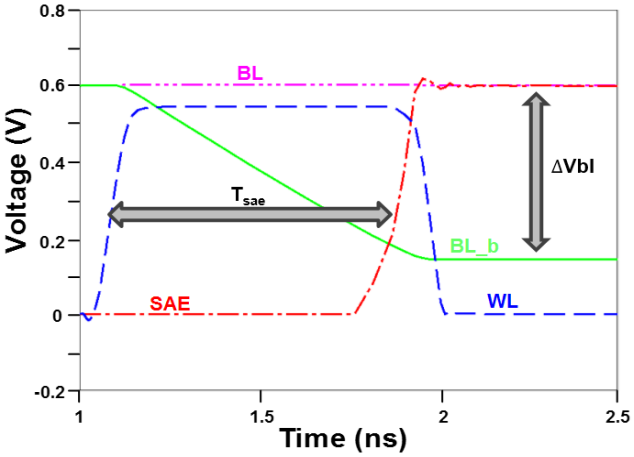

Fig. 2. Waveforms of SRAM read operation

2. Backgrounds

Fig. 1 (a) shows the structure of the conventional 6T SRAM cell which consists of four nMOS transistors and

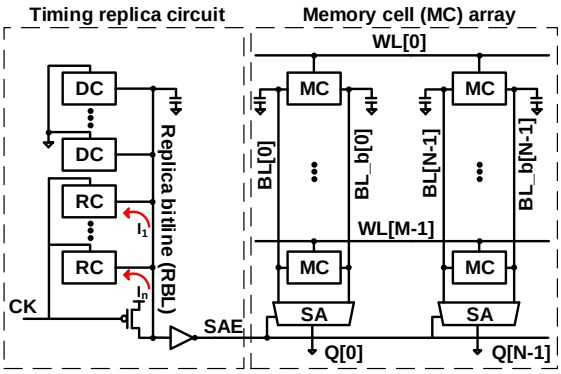

Fig. 3. Block diagram of the SRAM array with conventional RBL delay circuit (RC : replica cell; DC: dummy cell; MC : memory cell)

two pMOS transistors. Two cross-coupled inverters form a latch, which consists of two pull up transistors (PU) and two pull down transistors (PD). Two pass gate transistors (PG) connect the bit-line pairs with two storage nodes of the bit-cell, A and B.

Fig. 1(b) shows the status of the 6T SRAM cell during the read operation, and Fig. 2 shows the waveforms of read operation. The precharged BL or BL_b to V_{DD} is discharged according to storage node voltage as WL becomes '1'. For example, when the voltage of two storage nodes, V_A and V_B, are '1' and '0', respectively, as shown in Fig. 1(b), BL_b voltage (V_{BL_b}) is discharged and BL voltage (V_{BL}) is maintained near V_{DD}.

After the voltage difference of between BL and BL_b ($\Delta V_{bl,SRAM}$) is sufficiently large, SAE is enabled. For a stable read operation, when the SAE is enabled, $\Delta V_{bl,SRAM}$ should be larger than offset voltage (V_{OS}) of the SA. Thus, the time from the moment when the WL is enabled to the moment when the SAE is enabled, T_{SAE} should be sufficiently large. However, if T_{SAE} is too large, the energy consumption increases because of the unnecessarily large $\Delta V_{bl,SRAM}$. Therefore, the T_{SAE} should be appropriately determined considering energy consumption, read delay, and sensing yield [6].

3. Issues of Conventional Replica Bit-line

Generally, RBL technique or inverter chain is used to generate a SAE. When using inverter chain for generating SAE, T_{SAE} is determined by the inverter chain which consists of logic transistors.

However, inverter chain cannot track the cell current variation by PVT variation due to its different circuit structure and layout style compared with SRAM cell and bit-line. On the other hand, RBL technique can track the cell current variation related to the global variation because it has the same circuit structure and layout as the bit-line. However, because there is a local variation, the variation of T_{SAE} affects the sensing yield, delay, and energy consumption. [2].

Fig. 3 shows block diagram of the SRAM array with conventional RBL delay circuit. Timing replica circuit exists to make SAE signal. Memory cell (MC) array is consisted of MC and SA. Each column and each row

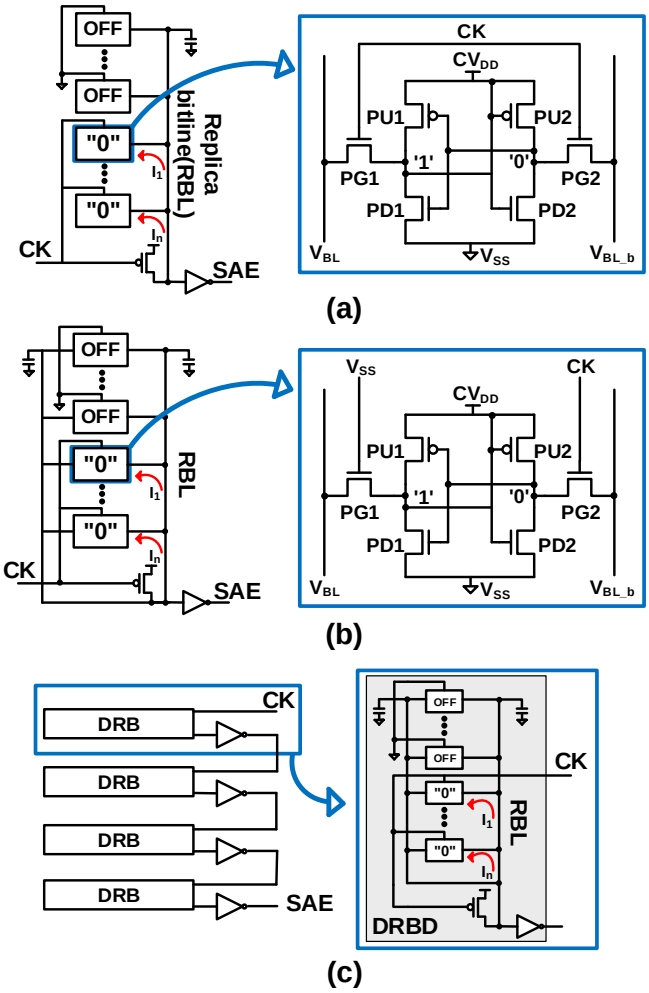

Fig. 4. Replica technique structures of (a) RBL_{conv}, (b) DRBD and (c) MDRBD in SRAM.

in SRAM array have two BL and one WL, respectively.

Fig. 4(a) shows the structure of conventional RBL (RBL_{conv}) technique which has the same structure as a column of SRAM array. RBL_{conv} is divided into on-cell and off-cell. The on-cell is turned on when CK is '1' but the off-cell is always turned off because WL is driven V_{SS}. The structure of the on-cell is similar to that of MC. The difference between replica on-cell and SRAM cell is data node which connects cell V_{DD} (CV_{DD}), which makes data node always '1'. Thus, the data node never flips to '0'. When on-cell is turned on, RBL begins to be discharged. Total current of all on-cell (I_{RBL}) is defined as

$$I_{RBL} = I_1 + I_2 + \ldots + I_n \quad (2.1)$$

I_{RBL} is proportional to the number of on-cells and each on-cell has the same mean of current (μ_I) and standard deviation of current (σ_I) because all on-cells have same structure.

$$\mu_{I1} = \mu_{I2} = \ldots = \mu_{In} = \mu_I \quad (2.2)$$
$$\sigma_{I1} = \sigma_{I2} = \ldots = \sigma_{In} = \sigma_I$$

Both μ_I and σ_I of each on-cell are independent of those of other on-cells because each on-cell has its own transistors. Thus, when the number of on-cells is N, the mean and standard deviation of I_{RBL} are defined as

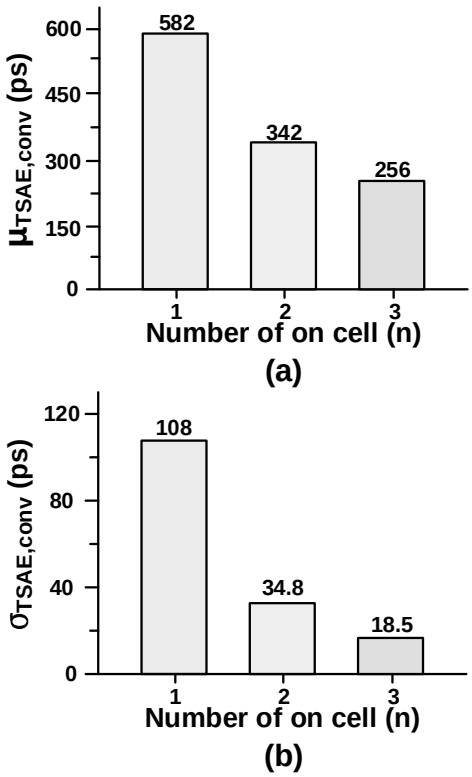

Fig. 5. Simulation result when RBL_{conv} uses 1, 2 and 3 on-cells for (a) $\mu_{TSAE,conv}$ and (b) $\sigma_{TSAE,conv}$.

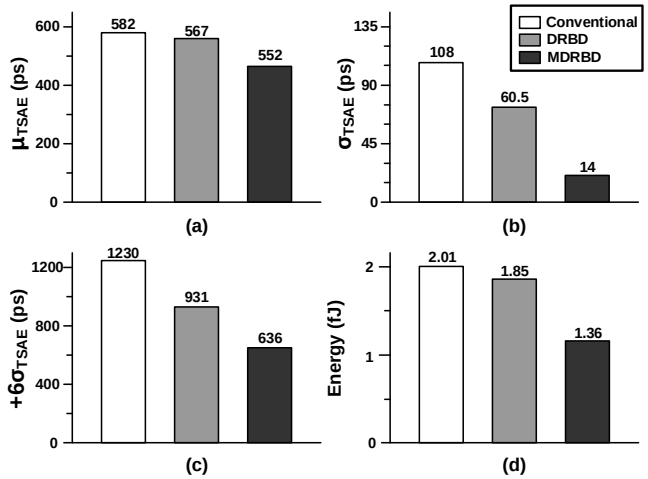

Fig 6. Simulation result of replica techniques for (a) μ_{TSAE}, (b) σ_{TSAE}, (c) Delay and (d) Consuming energy in SRAM.

$$\mu_{IBL} = \mu_{I1} + \mu_{I2} + ... + \mu_{In} = n \times \mu_I \tag{2.3}$$
$$\sigma_{IBL} = \sqrt{(\sigma_{I1})^2 + (\sigma_{I2})^2 + ... + (\sigma_{In})^2} = \sqrt{n} \times \sigma_{I1}$$

Meanwhile, T_{SAE} is inversely proportional to I_{RBL} as

$$T_{SAE} = \frac{CV_{DD}}{I_{RBL}} \tag{2.4}$$

Therefore, the mean and standard deviation of T_{SAE} in RBL_{conv} ($\mu_{TSAE,conv}$ and $\sigma_{TSAE,conv}$) are defined as (2.5) and (2.6) [3], respectively.

$$\mu_{TSAE,conv} = \frac{CV_{DD}}{\mu_{IRBL}} = \frac{CV_{DD}}{n \times \mu_I} \tag{2.5}$$

$$\sigma_{TSAE,conv} = \frac{\triangle CV_{DD}}{\triangle I_{RBL}} * \sigma_{IRBL} \approx \left| \frac{CV_{DD}}{\mu^2_{IRBL}} \right| * \sigma_{IRBL} \tag{2.6}$$
$$= \left| \frac{CV_{DD}}{n^2 \mu^2_I} \right| * \sigma_{IRBL} = \frac{1}{n\sqrt{n}} * \left| \frac{CV_{DD}}{\mu^2_I} \right| * \sigma_I$$

$\mu_{TSAE,conv}$ and $\sigma_{TSAE,conv}$ affect the delay and energy consumption. To reduce the delay and energy consumption, $\sigma_{TSAE,conv}$ needs to be reduced.

As shown as in Fig. 4(b), dual replica bit-line delay (DRBD) is proposed to reduce $\sigma_{TSAE,conv}$. One of PG in on-cell of DRBD connects V_{SS}, and thus the PG is always turned off and another PG is turned on when CK is '1' because the PG connects CK. DRBD has a two times larger capacitance than RBL_{conv}. The increased capacitance increases μ_{TSAE} and σ_{TSAE} compared with those of RBL_{conv} according to (2.4) [4]. Fig. 4(c) shows the structure of multi-stage dual replica bit-line delay (MDRBD) [5].

In the MDRBD, the number of capacitance is twice larger than that in RBL_{conv}, and the RBL is divided into several sub-RBLs (m stages novel dual replica bit-lines, DRBs). When RBL voltage (V_{RBL}) in the first DRB becomes threshold voltage of pMOS in inverter ($V_{pth,inv}$), V_{RBL} in the second DRB begins to be discharged and when V_{RBL} in the last DRB is discharged to $V_{pth,inv}$, SAE of MDRBD is enabled [6]. Each DRB has the same structure and sub DRBs are independent to each other, and thus the μ_{TSAE} and σ_{TSAE} of DRBD and MDRBD are defined as (2.7) and (2.8), respectively.

$$\mu_{TSAE,DRBD} = \frac{2}{n} * \mu_{TSAE,conv} \tag{2.7}$$
$$\sigma_{TSAE,DRBD} = \frac{2}{n} * \sigma_{TSAE,conv}$$

$$\mu_{TSAE,MDRBD} = \mu_{TSAE,DRBD} = \frac{2}{n} * \mu_{TSAE,conv} \tag{2.8}$$
$$\sigma_{TSAE,MDRBD} = \frac{1}{\sqrt{m}} * \sigma_{TSAE,DRBD} = \frac{2}{n\sqrt{nm}} * \sigma_{TSAE,conv}$$

where n is the number of on-cells and m is the number of sub-DRB states.

4. Simulation Result and Discussion

$\mu_{TSAE,conv}$ and $\sigma_{TSAE,conv}$ the number of on cells in RBL_{conv} is 1, 2, and 3 on-cells are shown in Fig. 5. In each case, the maximum number of on-cells which can satisfy 6σ sensing yield are used, when the supply voltage is 0.6V and 14nm FinFET technology is used in 32-kb SRAM. Fig. 5 shows simulation results of $\mu_{TSAE,conv}$ and $\sigma_{TSAE,conv}$ when the number of on-cells are 1, 2 and 3. $\mu_{TSAE,conv}$ and $\sigma_{TSAE,conv}$ are defined as (2.5) and (2.6), respectively. $\mu_{TSAE,conv}$ should be decreased by 1/n according to (2.5) compared with RBL_{conv} which turns on one of on-cell. But $\mu_{TSAE,conv}$ takes a long time more than expected value because drain to source current (I_{ds}) of transistor is defined by (2.9), where, V_A* is called the Early voltage [7].

$$I_{ds} = \frac{b}{2}(Vgs - V_{th})^2 \times (1 + \frac{Vds}{V_A*}) \tag{2.9}$$

When SAE is enabled, the drain to source voltage (V_{ds}) in PG is decreased proportionally to V_{BL} and source voltage

is increased by β-ratio, and thus I_{ds} is decreased and $\mu_{TSAE,conv}$ takes a long time more than expected value.

When on-cell in RBL_{conv} increases from one to two or three, $\mu_{TSAE,conv}$ is decreased by 41.3% and 56% in Fig5. (a), respectively and $\sigma_{TSAE,conv}$ is decreased by 59.5% and 75.9% in Fig. 5(b), respectively.

As shown in Fig. 6, the simulation results for delay and energy are compared. The maximum number of on-cells for RBL_{conv}, DRBD and MDRBD are 1, 2 and 3, respectively. μ_{TSAE} and σ_{TSAE} of three RBL techniques are shown in Fig. 6(a) and (b). In case of μ_{TSAE}, MDRBD should be decreased by 1/3 times according to (2.7) compared with RBL_{conv}, but the decrease in μ_{TSAE} is smaller than expectation due to the inverter between DRBs. σ_{TSAE} of DRBD and MDRBD are decreased by 44% and 87.1% compared with RBL_{conv} due to the increased number of on-cells, respectively. Delay is defined as $\mu_{TSAE} + 6\sigma_{TSAE}$ in Fig. 6(c). The delay of DRBD and MDRBD are decreased by 24.4% and 48.3% compared with RBL_{conv} because μ_{TSAE} and σ_{TSAE} are decreased due to the increased number of on-cells. Energy consumption in the SRAM array measured when T_{SAE} is μ_{TSAE} in Fig. 6(d). Energy consumptions of DRBD and MDRBD are reduced by 8% and 32.4% compared with $RBL_{conv,}$ respectively, due to the decrease in μ_{TSAE}.

5. Conclusion

In three representative replica techniques, RBL_{conv}, DRBD and MDRBD, the maximum number of on-cells which can satisfy 6σ sensing yield are determined. The delay and energy consumption of DRBD and MDRBD are significantly improved compared to those of RBL_{conv}. This is because σ_{TSAE} is decreased by using DRBD and MDRBD, the replica techniques can turn on more on-cells to maintain target sensing yield. As a result, additional on-cells reduce μ_{TSAE} and σ_{TSAE}, which improves the energy and delay. However, σ_{TSAE} of DRBD or MDRBD is still large and the delay and energy consumption of SRAM are degraded due to σ_{TSAE}. Thus, a new replica bit-line technique which can further decrease σ_{TSAE} should be developed as a future work.

References

[1] S. Pal and A. Islam, "Device bias technique to improve design metrics of 6T SRAM cell for subthreshold operation," in *Signal Processing and Integrated Networks (SPIN), 2015 2nd International Conference on*, 2015, pp. 865-870.

[2] Y. Niki, A. Kawasumi, "A Digitized Replica Bitline Delay Technique for Random-Variation-Tolerant Timing Generation of SRAM Sense Amplifiers," *Solid-State Circuits, IEEE Journal* of, vol. 46, pp. 2545-2551, 2011.

[3] U. Arslan, M. P. McCartney, M. Bhargava, L. Xin, M. Ken, and L. T. Pileggi, "Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, 2008, pp. 415-418.

[4] W. Jianhui, "A Multiple-Stage Parallel Replica-Bitline Delay Addition Technique for Reducing Timing Variation of SRAM Sense Amplifiers," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 61, pp. 264-268, 2014.

[5] C.-y. Peng, "Multi-stage dual replica bit-line delay technique for process-variation-robust timing of low voltage SRAM sense amplifier," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, pp. 700-706, 2015.

[6] 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 *Custom Integrated Circuits Conference, 2009. CICC '09. IEEE*, 2009, pp. 701-704.

[7]Neil H. E. Weste and David Money Harris, Integrated Circuit Design, PEARSON, pp. 74-78, 4th edition.