An area-efficient dual replica-bitline delay technique for process-variation-tolerant low voltage SRAM sense amplifier timing

Yi Li\textsuperscript{1a)}, Liang Wen\textsuperscript{1)}, Yuejun Zhang\textsuperscript{2)}, Xu Cheng\textsuperscript{1)}, Jun Han\textsuperscript{1)}, Zhiyi Yu\textsuperscript{1)} and Xiaoyang Zeng\textsuperscript{1b)}

\textsuperscript{1} State Key Laboratory of ASIC and System, Fudan University, Shanghai, 201203, China
\textsuperscript{2} Institute of Circuits and Systems, Ningbo University, Ningbo, Zhejiang, 315211, China

\text{a)} 11110720045@fudan.edu.cn
\text{b)} xyzeng@fudan.edu.cn

Abstract: A novel area-efficient dual replica-bitline delay technique is proposed in this brief to improve process-variation-tolerance of low voltage SRAM application. This strategy suppresses the timing variation by adding one another replica-bitline and introducing novel replica cell which has the same size as conventional. Simulation results in TSMC 65 nm LP technology show that more than 32.3\% timing variation is reduced and 18\% cycle time is saved at low supply voltage without any area overhead.

Keywords: low voltage SRAM, process variation, sense amplifier enable, replica bit-line delay

Classification: Integrated circuits

References

[1] G. Gammie, et al.: IEEE ISSCC Dig. Tech. Papers (2011) 132.
[2] F. Zhang, et al.: IEEE ISSCC Dig. Tech. Papers (2012) 298.
[3] U. Arslan, et al.: IEEE Custom Integrated Circuits Conference, CICC (2008) 415.
[4] S. Komatsu, et al.: IEEE Custom Integrated Circuits Conference, CICC (2009) 701.
[5] Y. Niki, et al.: IEEE J. Solid-State Circuits 46 [11] (2011) 2545.

1 Introduction

The low voltage SRAM for its ultralow power feature is widely used for energy-constraint applications such as mobile products, wireless body area...
networks (WBANs), and portable biomedical equipments [1, 2]. Sense amplifier (SA), as the key component of SRAM, generally is introduced to amplify the voltage difference between the long bit-line pair to reduce the read cycle and save the energy consumption on the bit-line. Moreover, if the SA-enable (SAE) arrives early before the bit-line difference reaches the SA offset, a read functional failure may occur. Contrarily, a late-arrived SAE would consume more unnecessary read cycle, thereby leading to power-waste [3].

However, in low voltage operation, the read control timing of SAE in conventional design is more vulnerable due to the process variation, thus leading to read failures and timing deterioration. The technique of suppressing the SA timing in previous work [3, 4] is not considered to be suitable for the low voltage operation because of the impractical test cost and timing mismatch between the replica bitline and normal one. The state-of-the-art digitized bitline delay replica technique is proposed in [5] for solving the SAE timing problem. Unfortunately, the logic delay of the timing multiplier circuit not only introduces large area overhead, but also brings significant timing variation due to the quantization error and the process variation of logic units.

In this paper, to solve the issues mentioned above, we propose a Dual Replica-Bitline Delay (DRBD) technique for SA read control timing in low voltage SRAM design. The novel scheme improves robustness of SAE timing nearly without any area overhead. The simulation results using TSMC 65 nm LP CMOS technology show that DRBD reduces the timing variation from 51.0% to 32.3% at 600 mV with 2 to 4 replica cells, and more than 40.8% at low supply voltage.

The remainder of this paper is organized as follows. In Section 2, the timing control technique of the conventional design and previous work are presented. Section 3 describes and analyzes the proposed dual replica-bitline delay technique. Simulation results and performance comparisons are presented in Section 4. Finally, Section 5 concludes this paper.

2 Conventional design and previous work

As shown in Fig. 1, the conventional replica-bitline scheme has one additional column of replica cells for generating the SAE timing. At initial state, the read bit-line and replica bit-line are precharged to the supply voltage level. Firstly, the read control and read word-line at the same time activate the replica cells and memory cells, respectively. Then, the voltage of replica bit-lines and normal bit-lines drop by the drawing current of the replica cells and memory cells. When the replica bit-line voltage level decreases below the logic input threshold of the inverter, the SAE will rise and become available. At this time, if the normal bit-line voltage difference is larger than $V_{offset}$ of SA, the output could be correct and if not, the read operation is failure.

For suppressing the variation of SAE, the general method of previous designs is to increase the number of the replica cell (assume $n$). However, this strategy in low voltage operation faced with some new challenges. Firstly, as mentioned in [5], there are some limits on replica cell count due to the
lower supply voltage. Secondly, although some timing matching schemes by adding logic gate delay could indeed increase the replica cell count and reduce the variation significantly, they also bring additional area cost. For example, the stat-of-the-art design of a digitized bit-line delay replica technique in [5] is proposed for suppressing the SAE timing variation. This technique introduces logic gate delay circuits called Timing Multiplier Circuit (TMC) for postponing the SA-enable signal. Unfortunately, a large number of logic gates in TMC are required for the SAE timing delay quantization. This leads to a serious area overhead. Additionally, on one hand, the quantization error of TMC could have significant impacts on SAE timing delay variation. On the other hand, in low voltage operation, the timing variation of logic gate delay could not be negligible, and even larger than the replica-bitline delay’s. Fig. 2 shows 500 times monte carlo simulation of the timing delay for the two delay strategies of the replica bit-line delay and the inverters chain logic delay. It indicates that the normalized deviation (Std/Mean) of the inverter chain is 0.28 which is larger than the replica bit-line’s 0.21.

3 Dual replica-bitline delay technique

A novel technique called Dual Replica-Bitline Delay (DRBD) is proposed in this paper. It could improve the process-variation-tolerance of the SAE timing without introducing any area overhead. Fig. 3 shows the scheme of the proposed strategy and the conventional one. In conventional design, there is only one replica bit-line for generating SAE delay timing. The replica bit-line
delay is determined by the total capacitance load $C_{bitline}$ and the replica cell discharged current $nI_{read}$. In proposed DRBD, each dummy cell and replica cell has two bit-line entries for connecting with two replica bit-lines and the two replica bit-lines are tied together. Thus, compared with the conventional design, the bit-line capacitance load and replica cell discharged current are $2nI_{read}$ and $2C_{bitline}$, respectively. Actually, the mean value of SAE timing delay in DRBD is the same as the conventional one’s. However, the standard deviation $\sigma$ of SAE timing in DRBD is suppressed to $\sigma/\sqrt{2}$ because of the addition of the discharged paths and transistors.

**Fig. 3.** Proposed dual replica-bitline scheme.

**Fig. 4.** Conventional single-ended and proposed dual-ended replica cell.

Fig. 4 shows the schematic and layout of conventional and proposed replica cell. In conventional design [3], 6 transistors are utilized for realizing one replica cell. N4 gate input is tied to supply voltage level. Transistor N2 and N4 form one discharge circuit path which connects with the replica bit-line. In proposed strategy, the replica cell also requires only 6 transistors.
but two discharged paths are realized by N1-N4. The gate input of N3 and N4 is tied to the supply voltage. The doubled discharged paths suppress the SAE timing variation as aforementioned. The cell layout of both conventional and proposed design are illustrated in the Fig. 4. Compared with the conventional design, although there are some changes of metal connections in the proposed one, the two design still have the same layout area. Thus, the proposed DRBD will not bring any area overhead.

4 Simulation results

Fig. 5 shows the 1000 times Monte Carlo simulation results of the proposed DRBD and the conventional single replica-bitline scheme at the supply voltage of 600 mV. The number of replica cells varies from 2 to 4 for the total 64 rows memory cells. Compared with the conventional design, the proposed DRBD reduces the standard deviation of the SAE delay variation from 51.0% to 32.3%. Noticed that the standard deviation of proposed 2 cells is larger than conventional 4 cells, even if they have both 4 discharging paths. This is because heavy BL capacitor load can degrade the standard deviation.

The cycle time improvement cross from 500 mV to 700 mV are illustrated in Fig. 6. Ideal cycle time is generally twice times as SAE timing delay without any timing margin. Conventional cycle time and proposed cycle time both have 3 times the standard deviation for the SAE timing margin. As

![Fig. 5. Standard deviation of SAE delay under different replica cells count.](image)

![Fig. 6. Cycle time improvement under different supply voltage.](image)
Fig. 6 shows, compared with the ideal cycle time, there is 4.8x-1.8x timing penalty for process variation in conventional scheme across 500 mV to 700 mV. However, by using the proposed DRBD, there is only 2.4x-1.5x timing penalty and about 50%-18% cycle time savings compared with the conventional one.

Table I. shows the comparison between different SA timing strategies. Unlike the previous work, the proposed DRBD doesn’t require any extra area cost but realizes the suppression for SAE timing delay variation.

| Variation Suppression Strategy       | Area Overhead |
|-------------------------------------|---------------|
| CICC, 2008 [3] configurable replica-bitline | Yes (configurable circuits) |
| CICC, 2009 [4] multi-stage replica-bitline delay | Yes (inserted buffers) |
| JSSC, 2011 [5] digital replica-bitline delay | Yes (TMC) |
| Proposed DRBD dual replica-bitline delay | No |

The new added replica-bitline in DRBD may bring additional power consumption. However, this power cost can be acceptable due to the significant reduction of SAE timing variation. Moreover, the conventional strategy requires higher supply voltage to achieve the same SAE timing variation as the proposed one. (For example, as Fig. 6 shows, the proposed DRBD needs only 550 mV supply voltage for realizing about 700 ps timing variation while the conventional one needs 600 mV.) Since the power consumption of SRAM is proportional to the supply voltage, higher supply voltage can lead to more power cost. Thus, for achieving some yield, from one perspective, the proposed DRBD can save overall power consumption of SRAM.

5 Conclusion

A novel area-efficient dual replica-bitline delay technique is proposed in this brief for improving the process-variation-tolerance of sense amplifier read control timing of low voltage SRAM. Unlike the conventional strategy, this scheme doesn’t require any extra area cost but realizes the suppression for SAE timing delay variation. The simulation is implemented in TSMC 65 nm LP technology. Compared with the conventional design, there are 51.0% to 32.3% timing variation reduction at 600 mV with different count of replica cells and more than 18% cycle time reduction when the supply voltage is lower than 700 mV.

Acknowledgments

The authors would like to thank State Key Laboratory of ASIC and System for providing the conducive research environment. This work is supported by the National Natural Science Foundation of China (Grant No.61234002).