The Effects of Operating Conditions on Speed and Power of Replica – Based SRAM Circuits

Nika Sharif Vaghefi and Nicholas Kumar
Department of EECS, University of California at Berkeley
Email: nikasv@berkeley.edu, nicholas.kumar@berkeley.edu

Abstract - With an increased demand for low-power portable devices, circuit designers continue to scale down supply voltages to minimize power dissipation. The progression towards low supply voltages, especially in SRAM designs, results in the larger impact of threshold and supply voltage fluctuations on the speed and power specifications of SRAMs. To minimize the PVT variations on the speed and performance of SRAM sense amplifiers, replica circuits provide a solution. This paper explores the tradeoffs between different replica-based SRAM circuits in terms of their level of tolerance towards PVT variations.

Index Terms – replica based SRAM, PVT variations, low-power SRAM, Sense Amplifiers.

I. INTRODUCTION

The growth of technology in recent years has placed the high-speed and low-power SRAM in greater demand for various applications. The most effective method of reaching low power operation is through the downward scaling of supply voltages. Additionally, the use of SRAM sense amplifiers has resulted in lower power consumption and higher performance. With the use of sense amplifiers, the wordline pulswidth can be smaller, only wide enough to generate a bitline swing detectable by the sense amplifier, resulting in less power consumption.

The read access time, an important performance parameter, is also reduced because the read operation does not need to wait for a large bitline swing. For reliable operation, the sense amplifier should track the bitline delay of the SRAM across process, voltage, and temperature variations. If the SAE triggers the sense amplifier before the difference in the voltage of the bitlines exceeds the sense-amp offset voltage (Vos), a read failure may occur. On the other hand, if the sense-amp is triggered by the SAE after the differential bitline voltage exceeds the offset voltage, both the access time and power consumption increase. Therefore, a careful approach should be taken to ensure that sense-amp is triggered at the right time.

Traditionally, sense-amps were timed through an inverter delay chain, which was not effective because process and operating conditions affect SRAM cells and the delay chain differently. Use of replicas for triggering sense-amps was introduced as an effective method to account for delay variations due to process and environmental conditions. The basic idea is to use memory cells connected to dummy bitlines whose combined capacitance is fixed to be a fraction of the core bitline capacitance. The replica bitlines are discharged much faster than the main bitlines. Therefore they can be used as triggers for the sense-amp. Furthermore, since dummy cells are actual SRAM cells that always contain zero, the process and environmental conditions have the same effect on replica and main bitline delays.
In section II of this paper, the conventional replica technique for wordline and sense control in low-power SRAM’s will be discussed. In section III, the multi-stage, digitized, and configurable replica circuits will be explored in the ways in which they attempt to reduce the variations in SAE sense amplifier timings. Section IV concludes this paper.

II. CONVENTIONAL REPLICA

Replica-based timing was introduced in [1]. A column of memory cells are used as dummy cells which always contain zero. The replica bitline is cut in order to have only a segment of memory cells connected to the replica bitline, which results in lower capacitance and thus lower delay for the replica bitline (see Figure 2). The exact number of cells in the segment is determined by the ratio between the sense-amp offset voltage and the difference between the supply voltage and the sense-amp trigger voltage. By using N cells in the replica, the replica bitline gets discharged N times as fast as the main bitline. Therefore the following equation must be satisfied in order to trigger the sense-amp at the right time:

\[(V_{dd} - V_{trigger}) = N \times V_{os}\]

Using replicas effectively solved the problem faced by the prevalent method at the time: use of an inverter delay chain. The replica delay chain is affected by process and environmental conditions the same way the SRAM cells are. In the traditional method, if one wanted to account for these variations, one had to add a large margin to the delay produced by the chain. Figure 1 shows that the replica delay chain behaves more consistently than the inverter delay chain.

III. COMPARISON

Figure 1: The bitline delay with Vos = 120 mV compared to two delay elements, one based on an inverter chain and the other on replica cell-bitline combination [1].

The main problem faced by replica delay chains is that the delay produced by the chain has variations itself. This is due to the Vth variations across different transistors. In order to allow for correct SRAM behavior we have to take these variations into account and hence, increase the mean delay so that in the worst case we will have the correct behavior. This in turn results in an increased read access time and leads to higher power consumption. Other papers discuss different approaches that can be used to lower the delay variations in SAE timing.

One of these approaches is called the multi-stage replica technique where the replica bitline is divided into n stages. Therefore, the replica bitline delay per stage is 1/n times that of the conventional bitline. The current through each of these replica stages remains
unchanged, as it is equivalent to the current through the conventional bitlines. Consider a multi-stage replica design with four stages. The replica bitline is thus, a quarter of a conventional bitline, which results in \( \frac{C_{\text{bitline}}}{4} \) capacitance per stage. Due to this decrease in capacitance, the delay of each individual stage is \( \frac{1}{4} \) times the delay of the conventional bitline or \( \frac{T_{\text{delay}}}{4} \). But the total delay of all stages in the multi-stage replica technique adds up to the delay of the conventional bitline. However, variation of the multi-stage replica bitline delay becomes half (i.e. \( \left( \frac{1}{4} \right) \frac{1}{2} \)) that of the conventional replica bitline delay [1], assuming that the delay variations in different stages are independent. By increasing the number of replica stages, the ability to suppress the SAE timing variation increases. However, this results in the usage of more inverters between the replica stages, which contributes to a higher delay. This increase in inverter delay will, in turn, require a decrease in the replica bitline delay to maintain a constant total delay. Therefore, a tradeoff exists between the effect of suppressing SAE timing variations and inverter delay overhead [1]. The disadvantage of this technique lies in the operation of the SRAM in low voltage conditions where \( V_{\text{th}} \) variations require a large replica stage count. As noted earlier, the inverter count increases with the number of replica stages resulting in a larger difference in delay between the replica and conventional bitlines, which is undesirable.

Since the number of replica stages needed to suppress SAE timing variations increases at advanced process nodes, power consumption and area constraints become an issue. This is clear when the number of stages grow to the extent that multiple columns are required. Therefore, reducing the SAE timing variations using the multi-stage replica technique becomes impractical at advanced process nodes due to the large increase in power consumption and area. To tackle this problem, a new design using configurable replica bitlines was proposed. Using post-silicon tests, cells that best suppresses SAE timing variations are selected. Therefore, a single column of replica cells with the best immunity to timing variations can be selected even for different process nodes. Thus, configurable replica bitline incurs no extra area and power consumption penalties for suppressing timing variation for different processes. The configurable replica cells are superior to the inverter-delay chains, as it is able to trace variations in temperature and voltage of the bitlines more accurately through its use of post-silicon test. The disadvantage of the cBRL design lies in the rising cost of post-silicon tests, which it uses to select variant-tolerant replica cells. In current SoC designs, which contain hundreds of SRAM modules, the cBRL technique proves impractical, as the post-silicon test for each of these SRAMs will greatly increase the cost of the chip.

The need for suppressing SAE timing variation at low supply voltages led to the design of the digitized replica bitline technique. This scheme uses \( k \) times as many cells as the conventional replica design, where \( k \) is chosen to suppress the SAE timing variation.
variation effectively. With k times as many cells, the standard deviation of the SAE timing is reduced by a factor of $k^{3/2}$. Therefore, the variation in SAE timing is greatly reduced. However, by having k times as many cells, the delay of the replica bitline is also reduced by a factor of k. A timing multiplier circuit (TMC) is then used to generate the SAE timing that is equivalent to the conventional SAE delay. This TMC circuit is composed of a forward and backward path. Adjusting the forward and backward path delays, we can obtain quantized multiples of the replica bitline timing. Figure 3 shows the schematic of the TMC circuit. Using TMC to multiply the replica bitline delay by k, we will have an SAE delay that has the same mean as the conventional scheme has, but has $k^{1/2}$ times less variations. Despite succeeding in reducing the SAE timing variation and cycle time in low Vdd operation, the addition of the TMC circuitry increases the area of the replica circuit significantly.

![Figure 3: Schematic of Timing Multiplier Circuit (TMC) [2]](image)

IV. CONCLUSION

The CBRL technique, which uses post-silicon tests, should achieve the lowest area and power consumption. However, the increasing cost of post-silicon tests makes it an impractical solution for reducing SAE variation. For low voltage operations, the DRBD technique proves to be a viable solution, however, it includes a large area overhead due to the TMC circuitry. The power consumption due to the increased TMC circuitry should also be considered before employing this technique. The multi-stage replica bitline technique is effective in reducing SAE timing variations and cycle time. For low voltage operations, this technique should be avoided due to the increased inverter delay, which will dominate the SAE timing.

REFERENCES


