Effects of Process Variation on the Access Time in SRAM Cells

Vicent Lorente and Julio Sahuquillo

Department of Computer Engineering (DISCA)
Universitat Politècnica de València, Spain
{vlorente,jsahuqui}@disca.upv.es

Abstract. As technology advances continue reducing transistor features, microscopic variations in number and location of dopant atoms in the channel region induce increasing electrical deviations in device parameters such as the threshold voltage. Deviations refer to mismatches with respect to device parameters at design time. These deviations are specially important in SRAM cells whose transistors are constructed with minimum geometry to fulfill area constraints, since they can cause some cells to fail.

In this paper, we study the impact of threshold voltage variations in the stability of the cell for a 16nm technology node. The failure probability has been studied for the four types of SRAM failures: write, access, read, and hold. We found that, under the assumed experimental conditions, the two former types of failures can be reduced by increasing the wordline pulse width of the cell. Experimental results show that access failures can be reduced up to 43.9% and write failures around 23.4% by enlarging the wordline pulse by 5 times the nominal width.

1 Introduction

As technology node continues to shrink, dealing with manufacturing imperfections is a major design concern since they affect the manufacturing yield, the energy consumption, and the performance of current and incoming processors.

Because of process variation, the manufacturing process provides transistors with different features (e.g. threshold voltage, channel length, or channel width) so that not all the transistors in a chip are able to properly work with the same voltage and frequency conditions. Due to this fact, manufacturers opt for relaxing the conditions to introduce in the market those chips having a noticeable amount of transistors not able to properly work under the target design goal. For instance, some processors are sold at a lower price when their speed is lowered below the target one to avoid process variation errors.

Cache structures occupy a large area of the microprocessor die. To reduce this area, SRAM cells of caches are designed with the minimum transistor size. This design condition makes these cells particularly sensitive to voltage scaling.

As a consequence, dealing with process variation in cache memories is a critical design issue. Failures due to the manufacturing process in caches mainly rise in

I. Caragiannis et al. (Eds.): Euro-Par 2012 Workshops, LNCS 7640, pp. 347–356, 2013.

destructive reads, unsuccessful writes, an increase in the access time, and content destruction in stand-by mode; known as read, write, access time and hold failures, respectively.

The number of failures due to process variation is determined by the processor working conditions (power supply and frequency). In other words, most errors usually appear in a subset of the working conditions range. For instance, it may happen that an error reading a memory cell appears when the processor works at a given frequency, but that error might not appear when working at a slower frequency (i.e. access time failure). On the other hand, a destructive read could be avoided by increasing the voltage supply.

Understanding why errors appear and which conditions should be done to avoid them (or most of them) is important for microprocessor architects in order to take the proper architectural design choices to achieve the best tradeoff between performance and power.

Among the different transistor features, the most significant source of random intra-die variation is the threshold voltage. In this paper we characterize the four mentioned failures, varying the power supply for a wide range of voltage values.

In this paper we focus on time dependent types of failures, which can be reduced by enlarging read and write operations (i.e. longer WL-pulse width). At first sight, it may seem that longer operations may increase execution time, but indeed, it may not. The cell probability of failure affects directly to the cache capacity. That is, the higher the probability of failure, the lower the effective cache capacity. As the effective cache capacity reduces, more cache misses are produced and therefore, the execution time increases. However, if we are able to reduce the probability of failure by reducing frequency, the execution time could not be affected because although a longer cache access time is used, the effective cache capacity is also larger. Therefore there is a tradeoff among cache access time, probability of failure (effective cache capacity) and energy. This paper is aimed at providing some preliminary results addressing these issues in order to help computer architects to devise the best design choice according to storage requirements of the running workloads.

This paper provides an study of the relation between WL pulse widths and SRAM time dependent failures. The remainder of this paper is organized as follows. Section 2 introduces the different types of SRAM cell failures. Section 3 characterizes SRAM cell failures at different power supply voltages, taking into account Vth variations. Section 4 presents the related work. Finally, Section 5 concludes the paper.

2 Background on SRAM Failures

Manufacturing process produces variations in the transistor parameters mainly due to physical factors caused by processing and masking imperfections [1]. Variations affect the channel length, channel width, oxide thickness, threshold voltage, line-edge roughness, and random dopant fluctuations, and are typically classified in *inter-die* and *intra-die* variations.

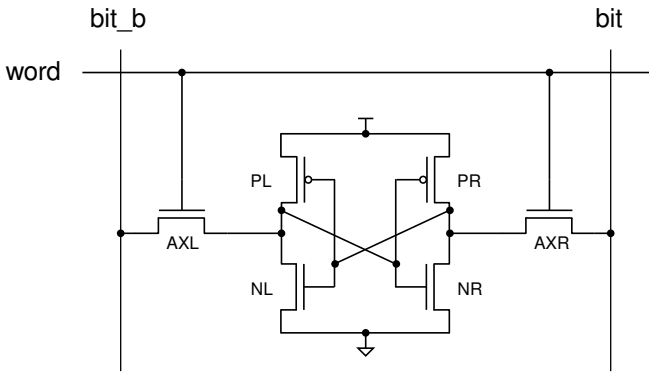


Fig. 1. 6T SRAM cell details

Inter-die variations affect all the transistors of a given die in the same way (e.g. threshold voltage of all the transistors either increase or reduce). As opposite, *intra-die* variations may affect transistors in the same chip in a different way (e.g. the Vth of some transistors can increase with respect to the nominal one whereas some others can have a Vth lower than the nominal one). In turn, *intra-die* variations can be either systematic or random. *Systematic* are variations depending on the variations of neighboring transistors while *random* variations are independent of the neighboring transistors.

Variations in different device parameters result in large spread in transistor threshold voltage [2]. Among them, the random placement of dopants causes threshold voltage mismatches among transistors that are spatially close to each other. Because of the small geometry of the SRAM cell, the main source of the device mismatch is the intrinsic fluctuation of the Vth of different transistors due to random dopant fluctuations [3–5], that is, random intra-die variations. These device parameters mismatches severely affect SRAM cells in sub-50nm technologies [6].

Each SRAM cell contains two pairs of transistors forming a logical not. Figure 1 shows these two pairs, one formed by PR-NR and the other formed by PL-NL. Any mismatch between devices of a pair degrades the stability of the cell and results in a cell failure when working at voltages lower than the design one. These mismatches between the variations of close transistors caused by intra-die variations can result in the failure of the cell in four different ways: hold failure, read failure, write failure, and access failure. Below we discuss them.

2.1 Hold Failure

Each of the mentioned transistor pairs is referred to as a node. One of them contains a “1” and the other a “0”. The voltage of the node storing “1” is the same as the power supply of the cell. Most current microprocessors implement a low power mode which highly reduces the power supply to save energy.

When working at this mode, if the voltage of the node storing "1" is reduced below the trip-point[1] of the node storing "0" then a flip occurs, so loosing the stored value and producing a hold failure.

2.2 Read Failure

Before the read is performed, both bitlines (i.e. *bit* and *bit_b*) must be precharged to Vdd. When the wordline is activated, the pass-transistors communicate both bitlines with the nodes of the cell. Then, the node storing "0" discharges the associated bitline while the node storing "1" remains to Vdd. The voltage increases for a while in the node storing "0" to a positive value due to the voltage divider action. When this increase is greater than the trip-point of the node storing "1", a flip is produced, which is known as a read failure, since it occurs when the cell is being read.

2.3 Write Failure

In a write operation, the bitline is precharged to "0" or "1" according to the value to be written. A write failure is produced when a "0" cannot be written in the cell. When the wordline is activated, the pass-transistor communicates the node storing "1" (Vdd) with the bitline (0V). To be a successful write operation, the node storing "1" must reduce the voltage below the trip-point of the node storing "0" while the wordline is active. Due to process variation, this decrease may be too slow. In other words, causing the time the wordline is active is not longer enough to decrease the voltage below the trip-point.

2.4 Access Failure

The cell access time is defined as the time required to produce the necessary voltage difference to excite the sense amplifier in a read operation. This voltage difference is typically by 10% the Vdd and must be reached while the wordline is active. To perform a read, both bitlines are precharged to Vdd and the bitline of the node storing "0" is discharged to 0V. The time needed to discharge that bitline depends on the pass transistor and the NMOS features. Due to process variation, the mismatch in these transistors can affect to the discharging speed . If this speed is too slow, the difference needed to excite the sense amplifier is not achieved.

3 Experimental Evaluation

3.1 Methodology

Failure probabilities have been estimated simulating an SRAM cell with the HSPICE circuit level simulator. Experiments assumed SRAM transistors based

[1] The trip-point is the voltage necessary at the input of the node to change the output.

on 16nm nodes with high-performance profile from the Predictive Technology Model (PTM) [7]. Simulations used the BSIM4 MOSFET model that addresses the MOSFET physical effects into sub-100nm regime.

Transistor sizes have been chosen according to [8] to ensure read and write-ability as well as to provide a good layout. Regarding area, device parameters (channel width W and channel length L) relationships (W/L) for the different types of transistors in the cell have been modeled as $6/2\ \lambda$ for the access transistors [2], $4/2\ \lambda$ for the pull-up PMOS transistors, and $8/2\ \lambda$ for the pull-down NMOS transistors. To simulate read and write operations we assumed 90ps word-line pulses.

As stated in Section 2, the different random intra-die variations can be summarized in Vth fluctuations. These fluctuations have been modeled for each transistor (NMOS and PMOS) of the cell as an independent Gaussian random variable with μ and σ_{VT0} equal to 0 and 37.3% , respectively, and assuming a maximum Vth deviation equal to 112% [9]. For the sake of accuracy, we gathered 40000 samples of cells generated using the Monte Carlo simulation.

3.2 Error Failure Characterization

Power consumption strongly depends on the power supply. More precisely, dynamic and static power grow quadratically and linearly with the power supply, respectively. Thus many research has focused on reducing the power supply to reduce energy. Nevertheless, reducing power supply negatively impacts on the number of errors so hurting the performance. Therefore, process variation introduces a tradeoff between performance and power.

This section characterizes the behavior of the SRAM cells, quantified in probability of failure, varying the power supply and taking into account Vth variations.

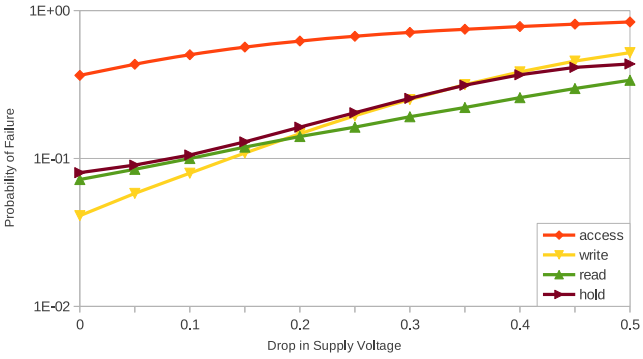


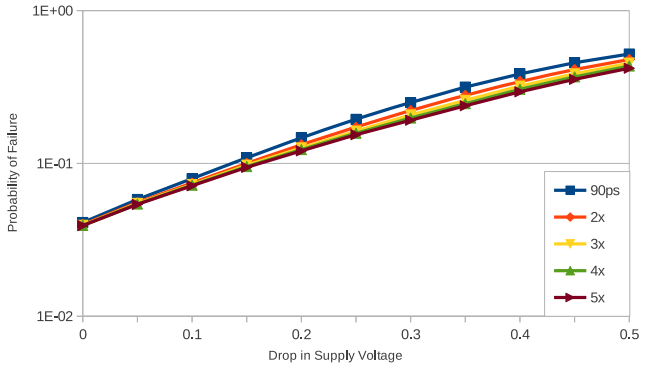
Fig. 2. Probability of failure reducing the power supply from 0.7V

[2] λ is defined as half the feature size (for 16nm nodes, λ=8nm).

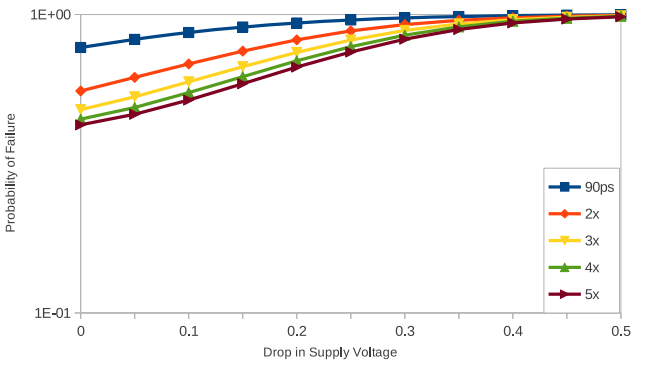
Figure 2 shows the probability of failure for the four types of SRAM errors described above as the voltage supply drops. The nominal power supply has been set to 0.7V and drops are drawn in steps of 50mV; that is, the X axis shows the power supply ranging from 0.7V to 0.2V.

As observed, the access failure is the predominant one. In fact, this operation already shows a noticeable amount of failures at nominal voltage. Read, write and hold failures have a probability much smaller. This means that the major performance benefits can be achieved by attacking and reducing the access failures.

On the other hand, access failures, as well as write failures, are affected by the WL pulse time. That is, these errors can be reduced by using a larger WL pulse.



(a) Write



(b) Access

Fig. 3. Probability of write and access failures varying the WL pulse

3.3 Impact of the Access Time on Write and Access Failures

To estimate the pulse width the cell was tested using transistors with no variations and a power supply of 0.7V. The adequate pulse width is that that meets the specifications, that is, a WL pulse that gets a voltage difference between bitlines higher than 10% of Vdd. Attending to the results, we used a 90ps as the nominal pulse width.

Nevertheless, as discussed above, write and access failures are produced because, due to transistor variations, there is not enough time to carry out the operation. This section analyzes how longer pulses can help reducing write and access failures. To this end, we enlarged the WL pulse in a 2-, 3-, 4-, and 5x factor.

Figure 3 shows the probability of failure for access and write operations varying the WL pulse. Results are shown for WL pulses as large as 2-, 3-, 4- and 5x the original WL pulse length. As observed, differences among pulse lengths curves rise with low voltage drops in access operations (left side of Figure 3(b)) and with high voltage drops in write operations (right side of Figure 3(a)). Regarding access failures, in it can be observed (Figure 3(b)) that curves for the different pulse lengths begin to converge in a 0.4V ranging power-supply drop. That is, from such a drop no improvement can be done. On the other hand, in Figure 3(a), improvements appear with power supply drops from 0.15V to 0.5V.

Computer architects need insights on access time and probability of failures since depending on the workload behavior it could be a better choice a larger access time but with less failures than a shorter access time with a larger amount of failures or vice versa. These design issues are addressed next. To this end, Figures 4 and 5 present a zoom of the most interesting parts of Figure 3. Results on these figures can be analyzed in two main ways. By drawing a vertical line crossing the different curves, it can be estimated how much the pulse length can improve the probability of failure. On the other hand, by drawing an horizontal

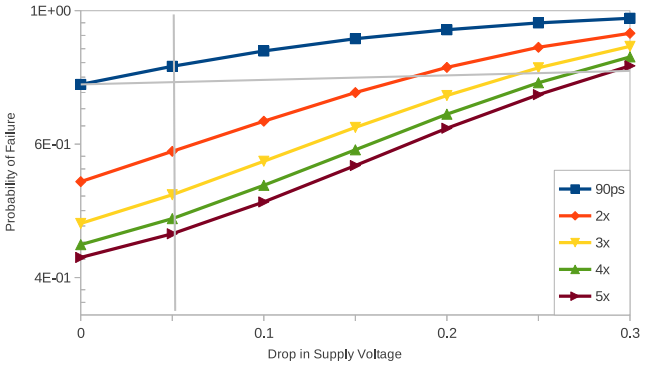


Fig. 4. Probability of access failure (zoom)

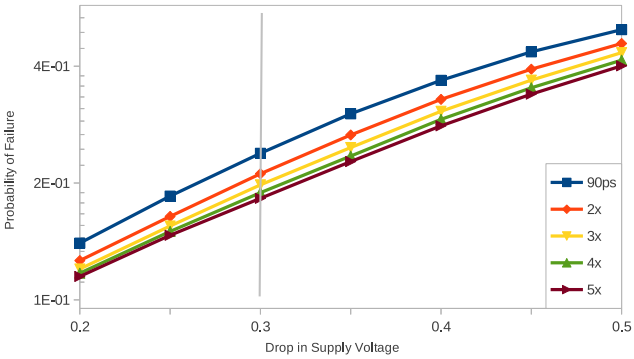


Fig. 5. Probability of write failure (zoom)

line, the curve traversing the crossing point identifies the pulse width required to work at a given voltage for a given probability of failure.

For instance, the vertical line drawn in the Figure 4 indicates that, for 0.65V, the probability of failure can be reduced around 25.5% by doubling pulse length, and as much as 43.9% by enlarging 5 times the nominal pulse.

Although the probability of failure increases with the voltage drop, the plotted horizontal line shows WL pulse length that would be required in order to keep the same probability of failure. For instance, for a 0.2V voltage drop, the required pulse to keep the same probability of failure as the original voltage should be twice as large the original one.

Regarding write operations Figure 5 shows that enlarging the WL pulse provides a rather more limited improvement than in access operations.

The vertical line shows that the probability of failure can be reduced about 11.4% reducing frequency 0.5X, and up to 23.4% by reducing it 5 times.

4 Related Work

Memory arrays blocks can be identified as unreliable at the post-fabrication process. Most of current approaches can be classified in two main types: those that correct unreliable blocks and those that avoid the use of it.

Regarding the former approaches type, Kim et al. [10] use error correcting codes (ECC) and redundancy to improve memory cell failures in SRAM caches, considering both systematic and random intra-die variations. This study takes into account three device parameters; the channel width, the channel length, and the threshold voltage. In contrast, we study how to reduce the number of two types of SRAM failures (access and write) in some cells increasing the WL-pulse width. In [11] Somnath et al. classify memory blocks in three main groups and apply different error correcting codes (ECC) to restore blocks according to the group they belong to. They also simulate effects of both inter and intra die variations.

Regarding the latter approaches type, Wilkerson et al. [12] propose two architectural techniques that reduce the effective cache storage capacity by 50%, based on probability of failure of a 65nm SRAM cell and taking into account Vth variations.

In [8] Mukhopadhyay et al. model failure probability of an SRAM cell in a 50nm technology node. For this purpose, random intra-die variations are taken into account resulting in Vth fluctuations. An exhaustive analysis of the different types of SRAM failures and their correlations is also provided. Their objective is to size the transistors of the SRAM cell at design time to minimize the failure probability of a memory chip considering area and leakage constraints. In contrast, our work keeps the same transistor size and varies the required time to access the cell to minimize the error failures caused by the reduction of power supply.

A model of timing errors due to parameter variation is proposed in [13] by Sarangi et al. This work takes into account both systematic and random intra-die variations, and apply the model to estimate timing error rates for pipeline stages in a processor with variations. Unlike this research, our work studies the effects of random intra-die variations focusing on SRAM cells

5 Conclusions

In this paper we have characterized the four types of failures that can be produced in SRAM cells (read, write, access and hold) analyzing how threshold voltage variation affects the probability of failure ranging the power supply voltage from 0.7V to 0.2V, and for 16nm feature size. Reducing the voltage, the power consumption can be highly reduced but at expense of suffering a higher amount of failures and reduced effective cache capacity. Therefore, there is a tradeoff between performance and power.

Simulation results showed that access failure is the predominant type of failure. A deeper analysis has been conducted to explore the impact of those types of failures that are dependent of WL-pulse width, that is, access and write failures. The main goal behind this work is to help computer architects to take architectural decisions to deal with the aforementioned tradeoff. In this context it would be worth to know the relationship between probability of failure and pulse length. Results have shown that the access failure probability can be reduced up to 25.5% and write failure probability by 11.4% when doubling WL-pulse width. In addition such pulse width also allows to drop the voltage 0.2V without increasing the probability of failure.

Acknowledgment. This work was supported by Spanish MICINN, Consolider Programme and Plan E funds, as well as European Commission FEDER funds, under Grants CSD2006-00046 and TIN2009-14475-C04-01.

References

1. Nassif, S.: Modeling and analysis of manufacturing variations. In: IEEE Conference on Custom Integrated Circuits, pp. 223–228 (2001)
2. Hocevar, D., Cox, P., Yang, P.: Parametric yield optimization for MOS circuit blocks. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 7(6), 645–658 (1988)
3. Bhavnagarwala, A., Tang, X., Meindl, J.: The impact of intrinsic device fluctuations on CMOS sram cell stability. IEEE Journal of Solid-State Circuits 36(4), 658–665 (2001)
4. Heald, R., Wang, P.: Variability in sub-100nm SRAM designs. In: IEEE/ACM International Conference on Computer Aided Design, ICCAD 2004, pp. 347–352 (November 2004)
5. Burnett, D., Erington, K., Subramanian, C., Baker, K.: Implications of fundamental threshold voltage variations for high-density SRAM and logic circuits. In: 1994 Symposium on VLSI Technology, Digest of Technical Papers, pp. 15–16 (June 1994)
6. Agarwal, A., Paul, B., Mukhopadhyay, S., Roy, K.: Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture. IEEE Journal of Solid-State Circuits 40(9), 1804–1814 (2005)
7. Zhao, W., Cao, Y.: Predictive Technology Model for Nano-CMOS Design Exploration. Journal on Emerging Technologies in Computing Systems 3(1), 1–17 (2007)
8. Mukhopadhyay, S., Mahmoodi, H., Roy, K.: Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS. IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems 24(12), 1859–1880 (2005)
9. Semiconductor Industries Association, International Technology Roadmap for Semiconductors (2007), `http://www.itrs.net/`
10. Kim, J., Mccartney, M., Mai, K., Falsafi, B.: Modeling sram failure rates to enable fast, dense, low-power caches. In: IEEE Workshop on Silicon Errors in Logic (March 2009)
11. Paul, S., Cai, F., Zhang, X., Bhunia, S.: Reliability-driven ECC allocation for multiple bit error resilience in processor cache. IEEE Transactions on Computers 60(1), 20–34 (2011)
12. Wilkerson, C., Gao, H., Alameldeen, A., Chishti, Z., Khellah, M., Lu, S.L.: Trading off cache capacity for low-voltage operation. IEEE Micro 29(1), 96–103 (2009)
13. Sarangi, S., Greskamp, B., Teodorescu, R., Nakano, J., Tiwari, A., Torrellas, J.: Varius: A model of process variation and resulting timing errors for microarchitects. IEEE Transactions on Semiconductor Manufacturing 21(1), 3–13 (2008)