Adv. Radio Sci., 4, 281–285, 2006
www.adv-radio-sci.net/4/281/2006/

Advances in
Radio Science

Modelling of the parametric yield in decananometer SRAM-Arrays

Th. Fischer[1], **T. Nirschl**[2], **B. Lemaitre**[2], **and D. Schmitt-Landsiedel**[1]

[1]Lehrstuhl für Technische Elektronik, Technische Universität München, München, Germany
[2]Infineon Technologies AG, München, Germany

Abstract. In today's decananometer (90 nm, 65 nm, ...), CMOS technologies variations of device parameters play an ever more important role. Due to the demand for low leakage systems, supply voltage is decreased on one hand and the transistor threshold voltage is increased on the other hand. This reduces the overdrive voltage of the transistors and leads to decreasing read and write security margins in static memories (SRAM). In addition, smaller dimensions of the devices lead to increasing variations of the device parameters, thus mismatch effects increase. It can be shown that local variations of the transistor parameters limit the functionality of circuits stronger than variations on a global scale or hard defects.

We show a method to predict the yield for a large number of SRAM devices without time consuming Monte Carlo simulations in dependence of various parameters (Vdd, temperature, technology options, transistor dimensions, ...). This helps the designer to predict the yield for various system options and transistor dimensions, to choose the optimal solution for a specific product.

1 Introduction

With the scaling of modern VLSI circuits according to Moore's law and the ITRS (ITRS , 2005) new problems arise. The smaller dimensions of transistors in advanced CMOS technologies allow the circuits to consume smaller chip area and to operate at higher clock rates. But on the other hand the standby power consumption is rising due to higher leakage currents. Short channel effects increase the sub-threshold leakage and with the thinner gate oxides of scaled-down transistors the gate leakage is also increasing.

Therefore, today many systems operate at lower supply voltages and the threshold voltages of the transistors increase, resulting in a lower power dissipation. However, the functionality of many circuits depends on this decreasing overdrive voltage, i.e. $V_{DD} - V_{th}$. Hence the read and the write security margins of static memory devices (SRAM) are decreasing.

On the other side the variations of the device characteristics (e.g. transistor length and width, threshold voltage) are ever more increasing with proceeding technologies. Printing these small devices gets harder due to the subwavelength lithography used in todays semiconductor industry. Also the statistical variation of the dopant concentration under the gate of a MOSFET transistor increases with smaller dimensions.

With emerging technologies the share of static random access memory (SRAM) gets larger with every generation. Due to the large demand of memory, SRAM cells must be designed as small as possible, and the cell area is an important benchmark for the quality of a design shrink. The large number of SRAM cells on a System on a Chip (SOC) and the usage of minimum feature size transistors makes the SRAM cell an ideal device for the characterization of variation effects and a good vehicle for yield testing.

In the following section, local and global variations are described. In Sect. 3 the read and write margins of SRAMs are introduced. Sections 4 and 5 present a method to model the parametric yield using a worst case distance analysis. Section 6 shows some results of the yield analysis and is followed by a summary.

2 Local and global variations

Variations of transistor parameters occur on various scales, lot to lot, wafer to wafer, die to die and device to device. They can be divided into variations on a more global scale and a local scale. Lot to lot, wafer to wafer and die to die variation are of global nature. These variations are common

Correspondence to: Th. Fischer (thomas.fischer@tum.de)

Published by Copernicus GmbH on behalf of the URSI Landesausschuss in der Bundesrepublik Deutschland e.V.

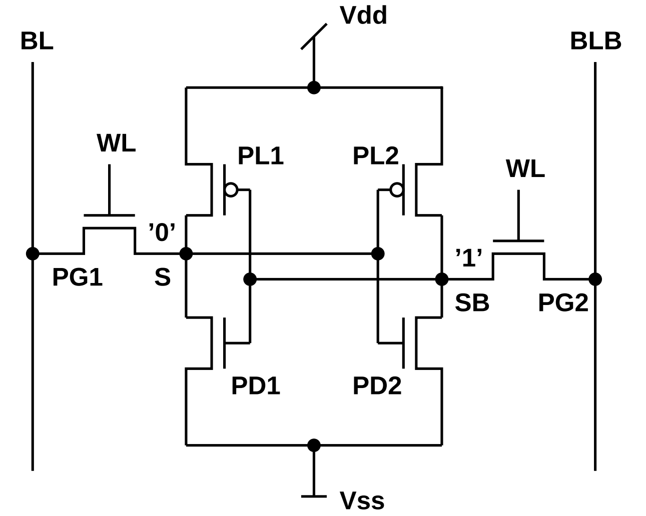


Fig. 1. A 6T SRAM core cell with the two pass gate n-FET transistors (PG1,2), the two pull down n-FET transistors (PD1,2) and the two p-FET pull up transistors (PL1,2).

to all devices of a die. An example for this type of variations is the oxide thickness t_{ox} or a variation of the width W and length L of the devices due to a global misalignment of the masks. Also the threshold voltage V_{th} of all the transistors on a die can be shifted by a constant value. These variations are of systematic nature. They can be improved during production with a tighter process control, and the impact of the global variations can be attenuated during design with layout techniques such as symmetry of important devices.

Local variations rom device to device are known as device mismatch. This are stochastic variations, such as the number of dopants under the gate area of a MOSFET device. These variations increase with smaller area of a device. Pelgrom et al. (1989) shows that the variation of the threshold voltage V_{th} obeys the following equation:

$$\sigma_{V_{th}} = \frac{A_{V_{th}}}{\sqrt{WL}} \tag{1}$$

where W is the width and L is the length of the transistor. $A_{V_{th}}$ is the matching constant for the threshold voltage. The variaton of the transconductance factor $\beta = \mu C_{ox} W/L$ is modeled in a similar way:

$$\frac{\sigma_\beta}{\beta} = \frac{A_\beta}{\sqrt{WL}} \tag{2}$$

In this case A_β, is the process related matching constant for the transconductance factor.

The only way a designer can improve the matching of the devices is to improve the area.

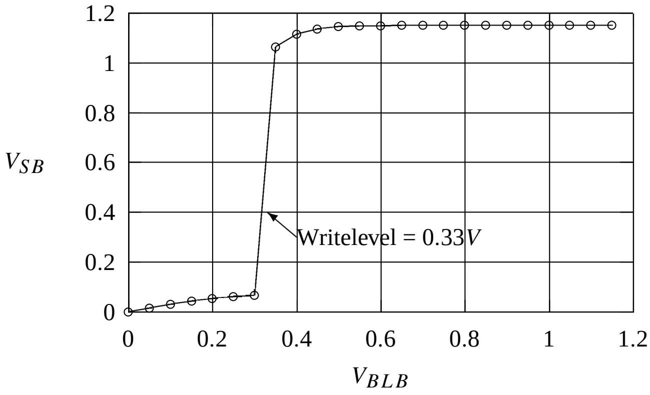


Fig. 2. Voltage at the storage node V_{SB} while lowering V_{BLB}. The core cell has a Write Level of 0.33 V.

3 Read and write margin of SRAM cells

Figure 1 shows a 6-transistor SRAM core cell. During read access the bitlines BL and BLB are pre-charged to a precharge level – usually V_{DD}. For further discussions we assume a '0' stored in the storage node S at the BL side. When opening the access devices PG1 and PG2, the voltage level on the storage node S at the bitline side rises due to the voltage-divider formed by the transistors PG1 and PD1. The cell ratio CR is defined as

$$CR = \frac{W_{PD}/L_{PD}}{W_{PG}/L_{PG}} \tag{3}$$

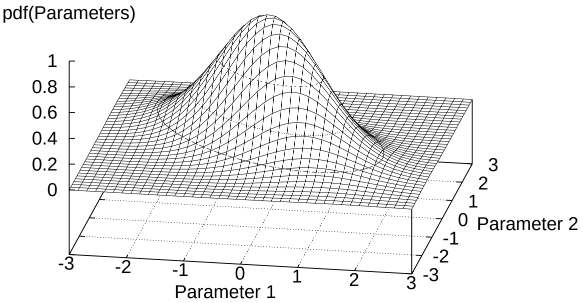
and sets the voltage V_S on the storage node. When this voltage V_S rises above the switching level (SWL) of the cross coupled inverters of the SRAM core cell, the cell looses its data while reading the cell. This is called destructive read. In this work, the static noise margin (Seevinck et al., 1987) is used to measure the stability of the SRAM core cell during read access. A destructive read occurs when the SNM drops to zero or below. Another performance measure of the SRAM core cell is the read current I_{read} through the transistors PG and PD.

When writing a '0' to the storage node SB on the BLB side of the core cell, the voltage on BLB is lowered to 0. Here the voltage V_{SB} on the storage node SB is defined by the voltage-divider formed by the transistors PL2 and PG2. Analog to the read case a pull-up ratio PR can be defined as

$$PR = \frac{W_{PU}/L_{PU}}{W_{PG}/L_{PG}}. \tag{4}$$

While lowering the voltage on BLB, the write level is defined as the value of the BLB when the cell flips (see Fig. 2). If the write level is larger than 0 the cell can be written.

The static noise margin, the write level and the read current are the three performances to define the functionality of a core cell in this work.



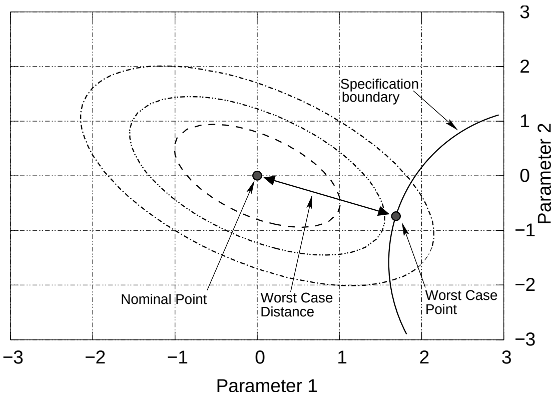


Fig. 3. Schematic probability density of two transistor parameters (top) and worst case distance and worst case point with specification boundary (bottom).

4 The worst case analysis

In order to define the yield of an SRAM core cell array the Worst Case Distance Analysis is used (Antreich et al., 1994). Simply spoken, the Worst Case Distance is a probability measure for the most likely violation of a performance specification. In Fig. 3 the parameter space of a circuit is symbolized. For example, the parameter space can be the local threshold voltage and mobility variations of all six transistors of an SRAM core cell. Here only two parameters are shown in the $x-y$-plane. The z-axes shows the probability density of the parameter space.

For every parameter-set a performance can be simulated with the according parameters as input for the circuit simulator. Now a specification for a circuit performance can be defined. On the top of Fig. 3 we see the top view of the probability density function. The solid line represents a specification boundary. This is given by all parameters sets that result in the same specified performance value. Everything right of the specification boundary violates the specification and results in an non-functional SRAM core cell.

The smallest distance between the nominal point, i.e. the parameter set with the highest probability, and the specification border is the worst case distance WCD. The point on the specification border that is closest to the nominal point has the highest probability that a parametric fault occurs.

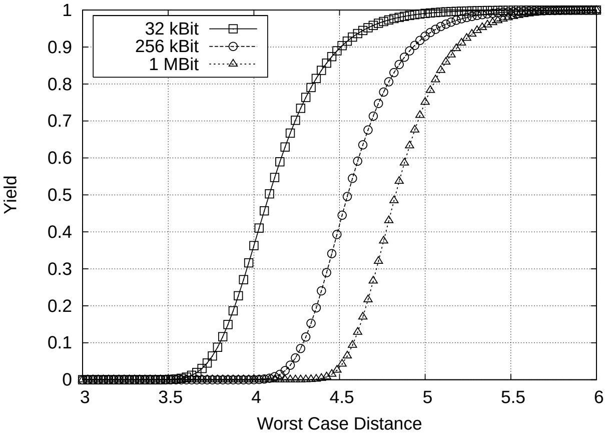


Fig. 4. Yield vs. Worst Case Distance of local variations for different SRAM core cell array sizes.

The WCD can be translated into a probability of a functional core cell array:

$$\text{Yield} = \left(\frac{1}{2}\left(1 + erf\left(\frac{\text{WCD}}{\sqrt{2}}\right)\right)\right)^N \quad (5)$$

where N is the number of core cells in an array. In order to achieve a yield of 90% for a 1MBit SRAM array a WCD of 5.2 (see Fig. 4) must be reached or redundancy must be used.

5 First global than local

In order to find the yield of an SRAM array, global and local variations must be simulated independently. First a set of global parameters must be found that represent the worst die in the production process — not including mismatch. Therefore only parameters that define global variations are considered for analysis, e.g. the oxide thickness t_{ox}, the global fluctuations of the n-FET and p-FET threshold voltage V_{thn} and V_{thp} and the length and width variations of the device x_L and x_W. For a given statistical variation of a process, the algorithm looks for the set of the global parameters that results in the worst performance of the core cell.

Once the parameter set for the worst die is found, an analysis of the local variations in an SRAM cell can be made. The local parameters considered in this work were the V_{th} variations and the mobility variations of the six transistors in the core cell. Hence 12 local parameters were taken into account.

In order to find the worst case distance, an optimization problem in a multidimensional space must be solved.

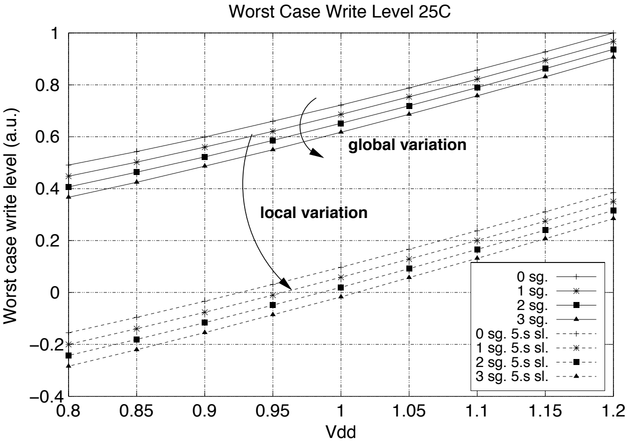


Fig. 5. Worst Case Write Level for global and local variations. Variations in sigma global (sg.) and sigma local (sl.).

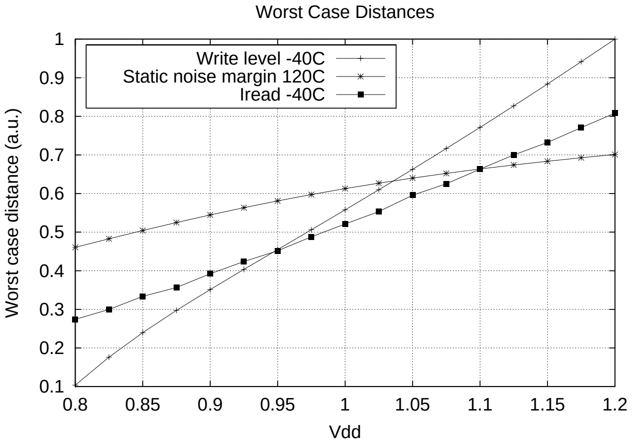


Fig. 6. Worst Case Distances for the write level, the static noise margin and the iread of a SRAM cell. All temperatures are the worst case temperatures.

For this task we used a tool called WiCkeD from Muneda (http://www.muneda.com). This tool uses an sequential quadratic programming (SQP) algorithm to find the Worst Case Distance (Antreich et al., 2000). An external TCL script was used to control the WicKeD program to run the worst case distance analysis first with just global variation, and then with local variations using the worst case parameters obtained by the global WCD analysis. A SPICE simulator simulated the performances of the SRAM core cell for a set of parameter values, and a Perl script wrote the simulation results back to WiCkeD for further evaluation.

Two modes of an analysis of the parametric yield are realized. First of all, a WCD can be found by defining specifications for each performance, like Write Level or Static Noise Margin. On the other hand it is possible to look for the worst performance in an array of a given size. For that purpose a WCD is chosen according to the SRAM array size. Then the parameter set for the worst performance with the given WCD is searched for.

6 Results and performance

For the simulation of the SRAM performances, the BSIM4 parameters of an 65 nm technology (Luo et al., 2002) were used. First of all, the influence of global and local variations on SRAM core cell performances were simulated. In Fig. 5 we see a normalized plot of the Write Level of an SRAM core cell for different supply voltages. The upper set of lines shows the relative worst case Write Level depending on global variations ranging from 0 to 3 sigma. With the worst case parameter-set found by the global simulation, a simulation for the worst case performance using just local variations such as the V_{th} and the mobility μ for the six core cell transistors was made. The local worst case distance WCD of 5.2 sigma was chosen to represent a SRAM core cell array with 1 MBit (see Fig. 4).

The local variation has a large influence on the functionality of the SRAM array due to two effects. First the large number of core cells that must be fulfill the specifications in order to get a SRAM array with 100% working cells. Second the smaller size of the transistors increase the local variations and with lower overdrive voltage result in more core cells failing the specification.

Although the local variations are more dominant, one can see form Fig. 5 that a tighter process control can result in circuits operating at a lower supply voltage.

To identify the performance that is the main yield detractor, the worst case distances of the write level, the static noise margin and the read current were simulated. The circuit is simulated with the temperature corners of the specification that results in the respective worst performance. This is -40°C for the write level and the read current and 120°C for the SNM. So the worst cases of the performances are comparable. It is shown in Fig. 6 that different performances limit the yield of the core cell. For high supply voltages the read stability (SNM) is the main yield detractor, whereas the write level limits the yield for supply voltages lower than $1.1V$, and the read current is the yield detractor for the low voltage operation.

The Worst Case Distance simulations are much faster than a comparable number of Monte Carlo simulations. To verify a yield of 5 sigma in a 95% confidence interval, 10 million Monte Carlo runs need to be simulated. These simulations take about 200 h of runtime on a single processor machine. The Worst Case Distance simulations need about 2 min of time on a comparable computer. Both, the Monte Carlo simulations and the WCD simulations can be divided into multiple threads and run on multiple processors. However, these figures show clearly that simulations of parametric yield can be simulated much faster with the Worst Case Distance analysis. Therefore multiple yield simulations with multiple parametric variations, such as temperatures, supply voltages and even different transistor sizes, can be done and used for design centering and verification considering global and local variations.

7 Conclusions

We presented a method to effectively simulate the yield and the worst case corners of several SRAM performances. This Worst Case Distance Analysis is by 6 orders of magnitudes faster than a Monte Carlo Analysis. So a six-sigma simulation can be done for several parameter alterations of a circuit. It is possible to find the main yield limiting performances for the worst case operating conditions. This gives the circuit designer a tool to simulate different transistor sizing options under the influence of global and local parameter variations.

References

ITRS, International Technology Roadmap for Semiconductors 2005 Edition,http://public.itrs.net/

Pelgrom, M., Duinmaije, A., and Welbers, A.: Matching Properties of MOS Transistors, IEEE J. Solid State Circuits, 24, 1433–1440, 1989.

Seevinck, E., List, F., and Lohstroh, J.: Static Noise Margin Analysis of MOS SRAM Cells, IEEE J. Solid State Circuits, 22, 525–536, 1987.

Antreich, K., Graeb, H., and Wieser, C.: Circuit Analysis and Optimization Driven by Worst Case Distances, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 13, 57–71, 1994.

Antreich, K., Eckmueller, J., Graeb, H., Pronath, M., Schenkel, F., Schwencker, R., and Zizala, S.: WiCkeD: Analog Circuit Synthesis Incorporating Mismatch, IEEE Conf. on Custom Integrated Circuits, 511–514, 2000.

Luo, Z., et. al.: High Performance and Low Power Transistors Integrated in 65 nm Bulk CMOS Technology, Proc. IEEE IEDM, 661–664, 2004.