TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Entwurfsautomatisierung

On the Sizing of Analog Integrated Circuits towards Lifetime Robustness

Xin Pan

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. sc. techn. Andreas Herkersdorf
Prüfer der Dissertation:	1.	Priv.-Doz. Dr.-Ing. Helmut Gräb
	2.	Univ.-Prof. Dr. rer. nat. Doris Schmitt-Landsiedel

Die Dissertation wurde am 31.10.2012 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 22.05.2013 angenommen.

A paperback version of this thesis was published by Verlag Dr. Hut, Munich, in 2013.
ISBN 978-3-8439-1198-6.

Acknowledgements

From March 2008 to October 2012, I worked as a research assistant at the analog CAD group of Institute for Electronic Design Automation, Technische Universitaet Muenchen. During this period , I had the opportunity to work with many nice people, colleagues and friends, with wonderful ideas and personalities. I would like to express my deep gratitude to them, without whom my doctoral research would not be successful.

Firstly I would like to thank Professor Ulf Schlichtmann for giving me the chance to work at the institute, as well as for his warm help during my Master study at the university. Thanks to the Master program, I had the opportunity to come to Germany and studied my favorite subjects, with contacts to many nice colleagues around several institutes already during that period.

My special thanks go to my doctoral research supervisor, PD Dr.-Ing. Helmut Graeb. He opened the gate of analog CAD to me with interesting topics and emerging physical effects of semiconductor technologies. I remember every moment we discussed ideas, checked formulas, shared progress and chatted for fun. He always motivated me to proceed in my research with warm encouragement and constructive feedbacks which I appreciate very much. I thank him also for the time and efforts in proofreading all my publications, thesis and presentation slides.

I would like to thank all of the colleagues at the institute, especially the members of the analog group, namely Dr. Daniel Mueller, Dr. Tobias Massier, Husni Habal, Michael Eick, Michael Pehl, Michael Zwerger and Aurélien Tchegho. The fruitful discussions during the study and research, valuable feedbacks on presentations and results, as well as the colorful events organized by them will be the most cherished piece of memories in my life.

Last but not least, I would like to express my deep gratitude to my parents Gang and Baozhen for their continuous and invaluable encouragement. Sincerely, I thank my wife Chang for her very kind understanding and solid supports all the time, and my lovely baby Keyun for bringing us the endless happiness.

Many thanks to you all.

Munich, October 2012

Xin Pan

Contents

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Analog Design . . . 3
  - 1.2.1 Typical Flow of Analog Integrated Circuit Design . . . 3
  - 1.2.2 Discussions and Challenges . . . 5
- 1.3 Contributions of this Thesis . . . 6
  - 1.3.1 Study on Joint Effects of Process Variations and Transistor Aging 6
  - 1.3.2 Design Flow for Lifetime Robustness Optimization . . . 7
  - 1.3.3 Analytical Modeling for Aged Yield Prediction . . . 7
- 1.4 Previous Publications . . . 8
- 1.5 Organization of this Thesis . . . 8
- 1.6 Summary . . . 8

2 Reliability Issues 11
- 2.1 Process Variations . . . 11
- 2.2 Reliability . . . 12
  - 2.2.1 Reliability Function $R(t)$ and Failure Rate $z(t)$ . . . 12
  - 2.2.2 Negative Bias Temperature Instability . . . 15
  - 2.2.3 Hot Carrier Injection . . . 18
  - 2.2.4 Time-Dependent Dielectric Breakdown . . . 19
  - 2.2.5 Electromigration . . . 19
- 2.3 State of the Art . . . 20
  - 2.3.1 Reliability Simulation . . . 20
  - 2.3.2 Solutions towards Transistor Aging . . . 21
  - 2.3.3 Design Centering considering Process Variations . . . 23
    - 2.3.3.1 Statistical Methods . . . 23
    - 2.3.3.2 Deterministic Methods . . . 25
  - 2.3.4 Joint Effects of Transistor Aging and Process Variations . . . 27
- 2.4 Summary . . . 29

3 Problem Formulation 31
- 3.1 Age and Lifetime . . . 31

3.2 Parameters . . . 32
3.2.1 Design Parameters . . . 32
3.2.2 Statistical Parameters with Aging . . . 33
3.2.3 Operating Parameters . . . 35
3.3 Performances with Aging . . . 36
3.4 Fresh Yield and Aged Yield . . . 37
3.4.1 Definition . . . 37
3.4.2 Statistical Analysis Method . . . 40
3.5 Sizing Rules with Aging . . . 43
3.6 Summary . . . 45

4 Aged Yield Optimization with Fresh and Aged Sizing Rules 47
4.1 Worst-Case Distance . . . 47
4.1.1 Yield Analysis and Worst-Case Analysis . . . 47
4.1.2 Yield Estimation Based on Worst-Case Distance . . . 50
4.1.3 Problem Formulation towards Worst-Case Distance Based Yield Estimation . . . 54
4.1.4 Solution using Lagrangian Functions . . . 56
4.2 Aged Worse-Case Distance and Aged Yield . . . 59
4.3 Design Flow . . . 61
4.3.1 Simulation Flow of the Aged Circuit . . . 61
4.3.2 Fresh and Aged Sizing Rules of a Circuit . . . 65
4.3.3 Circuit Layout Area Estimation . . . 65
4.3.4 Optimization of Fresh Circuit with Fresh and Aged Sizing Rules Checking and Maximum Area Constraints . . . 66
4.3.5 Aged Yield Analysis . . . 70
4.4 Summary . . . 71

5 Aged Yield Prediction 73
5.1 Aged Worst-Case Distance Prediction Model . . . 75
5.1.1 Idea . . . 75
5.1.2 Linear Performance Model at t_1 . . . 75
5.1.3 Mapping from t_0 to t_1 . . . 77
5.1.4 Prediction of $\beta_{w,U}(t_1)$ . . . 77
5.1.5 Second Order Sensitivity Term . . . 79
5.2 Algorithm for the Aged Yield Prediction . . . 80
5.3 Summary . . . 81

6 Experimental Results 83
6.1 Miller Operational Amplifier . . . 83

6.1.1 Circuit Topology . . . 83
6.1.2 Circuit Performances and their Specifications . . . 84
6.1.3 Results on Aged Yield Optimization . . . 87
6.1.4 Results on Aged Yield Prediction . . . 90
6.2 Folded Cascode Operational Amplifier . . . 92
6.2.1 Results on Aged Yield Optimization . . . 92
6.2.2 Results on Aged Yield Prediction . . . 94
6.3 Summary . . . 96

7 Conclusion **97**

Bibliography **99**

Lists **111**
List of Figures . . . 111
List of Tables . . . 113

Abstract in German **115**

Chapter 1

Introduction

1.1 Motivation

Entering the 2010's, a huge progress in the electrical engineering and information technology changes our daily life in various ways. For example, the traditional cell phone has been gradually replaced by the smartphones and multi-function cell phones [wikb], which are capable of running various applications based on their mobile operating platforms. They integrate together various functions such as digital camera and video recorder, GPS receiver, accelerometer, wireless internet connections and so on. The spread of the tablet personal computer (Tablet PC) [wikc] also gives people a totally different view about how a computer can be used and played on one single touchscreen which serves as both input and output devices . Another example is in the automotive industry [eet]. A typical modern car responds to driver's commands and environmental conditions in a way which is much smarter, faster and safer than ever before. More and more functions of the car are assisted automatically by microcontrollers in real time, such as driving, controlling, safety functions, navigation and entertainment systems, etc.

One of the key enabler of all these advances is the continuous development and improvement in the Integrated Circuits (IC) industry. After the first practical IC was invented simultaneously by Jack Kilby at Texas Instruments [Kil] and Robert Norton Noyce at Intel [Noy] in 1959, the idea of manufacturing various circuit components on a small piece of semiconductor material was intensively explored and further developed. From 1960's until today, the number of transistors in an integrated circuit follows the Moore's law, which states that such number doubles every two years [Moo75] (although initially it was stated as every year in [Moo65]). Thanks to the numerous innovations since then (such as the invention of the Complementary

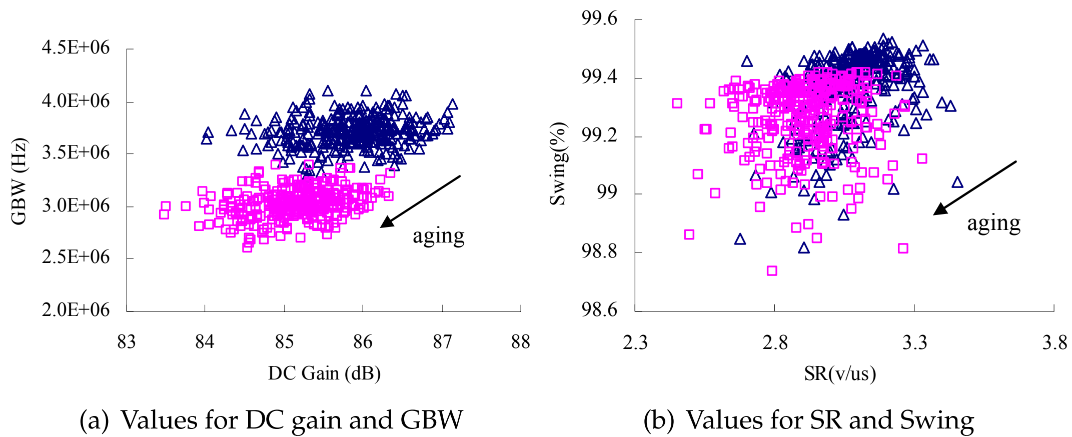
Metal-Oxide Semiconductor (CMOS) process by Frank Wanlass in 1963 [Wan], the invention of the excimer laser photolithography by K. Jain at IBM in 1982 [JWL82], and so on), the capability of integrating an increasing number of transistors on a single chip is greatly enhanced. The trend described by the Moore's law has been followed quite well by now and is expected to be valid in the near future. [Sch97]

Today, what a complex system did in the past has been integrated onto a single chip, consisting of several billions of components. Such chip can provide much more functionality than ever before, combining logic functions, memory, analog and mixed-signal applications. Such improvement means increased circuit complexity and enhanced circuit performance features, computational capability, data storage and the power consumption. With the help of the continuous improvement and progress in photolithography, the corresponding minimum feature sizes in chip manufacturing process has shrunk from 500 nanometer in 1990's to 45 nanometers and below in 2010's, as summarized and predicted by the International Technology Roadmap for Semiconductors (ITRS) Report 2009 [I.T].

However, despite the bright future seen by the above mentioned integration, many side effects and challenges arise due to the continuous shrinking of the semiconductor technology. The ITRS Report 2009 summarizes the challenges faced in the semiconductor industry in the near future and in the long term for logic device, memory device, RF, analog and mixed-signal circuitry, as well as for the new materials and manufacturing process [I.T].

Among the challenges arising from the development of the integrated circuits, the effects of manufacturing process variations and operational lifetime circuit reliability are becoming significant. Different from many other challenges, these effects can be and should be considered early during design phase by the designers, thus the designed circuits are well tolerant of such known effects. Especially for those safety-critical application areas, such as automotive and aviation industry, safety controller, or computational capability-sensitive fields, such as high performance computers, where deviations from specifications are not acceptable.

An example of the joint effects on a typical analog circuit block is illustrated in Figure 1.1, where 300 Monte-Carlo simulations are run on a fresh and 5-year-old Miller operational amplifier with a 180nm industrial technology. Values of DC Gain and Gain-Bandwidth Product (GBW) are shown in Figure 1.1(a), and values of slew rate (SR) and output voltage swing (Swing) are shown in Figure 1.1(b). The clouds of performance distributions, as shown in the figures, are the result of manufacturing process variations. They deviate from their nominal values due to the imperfectness during the manufacturing process. The shifts of the performance distributions in 5 years, on the other hand, result from drifts of transistor parameters, such as V_{th}, due



(a) Values for DC gain and GBW (b) Values for SR and Swing

Figure 1.1: Shift of the performance distributions from 300 Monte-Carlo simulation samples on a fresh (triangles) and 5-year-old (squares) Miller operational amplifier.

to negative-bias-temperature-instability and hot carrier injection during circuit operations. As can be clearly seen in the figures, both performance distributions move towards negative directions. Certain samples of the circuits thus may fall out of the possible performance specifications during operational time, resulting in an early wear-out, or in other words, a shorter lifetime than expected.

As a result, the tolerance design of integrated circuits considering manufacturing process variations and lifetime circuit reliability have been the major concern for the integrated circuit designers as well as for the manufacturers since the last decades until today. The physical roots and behaviors of these effects will be detailed in the next chapter.

1.2 Analog Design

1.2.1 Typical Flow of Analog Integrated Circuit Design

In the context of this thesis, the focus is on the methodology of design and analysis of the analog part of the integrated circuits. A typical analog integrated circuit design flow, consisting of several steps and loops [Bak08], can be summarized in Figure 1.2.

The input of the flow are the specifications. They describe the functionalities of the circuit, as well as the information from the manufacturing technology needed for the

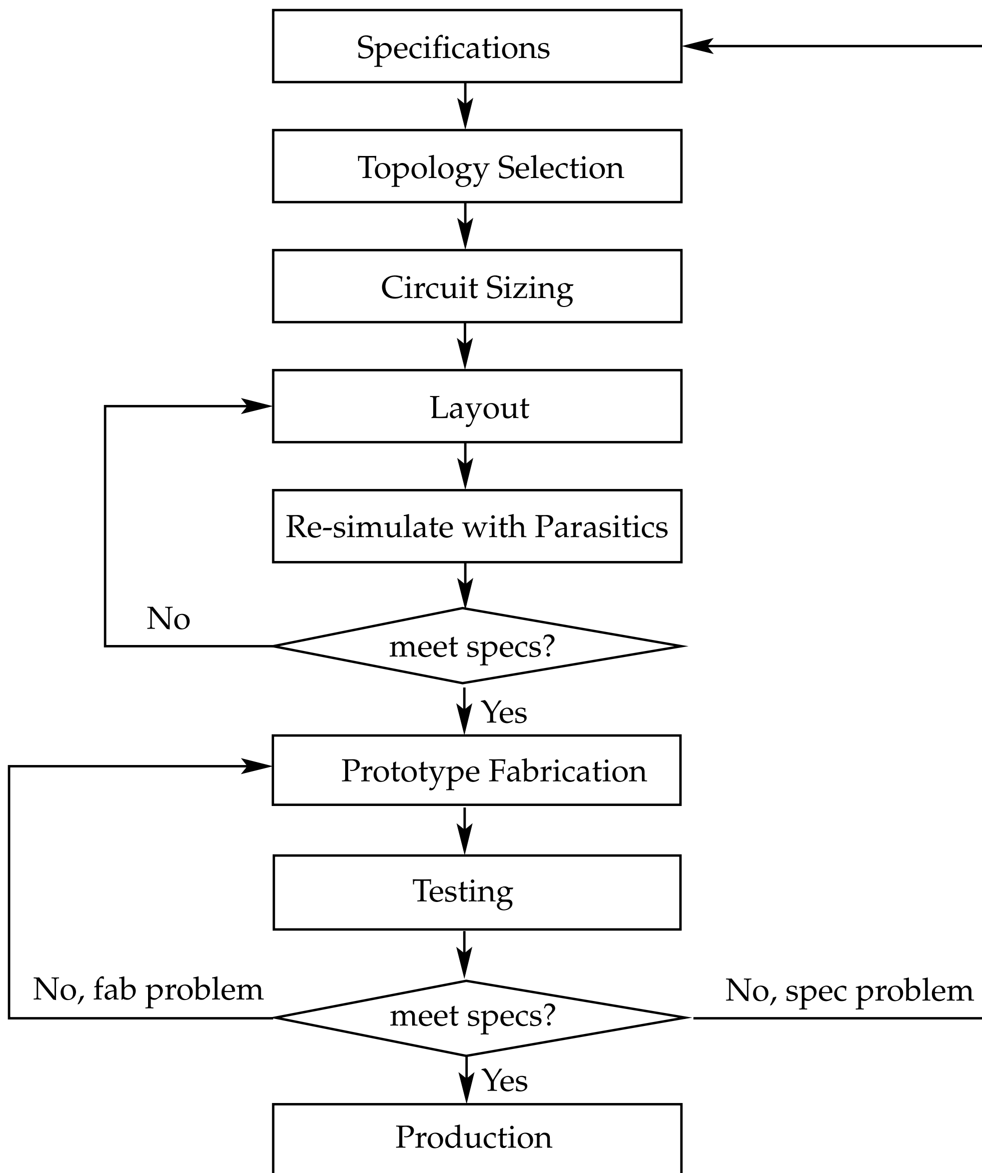


Figure 1.2: Typical design flow of analog integrated circuits [Bak08]

design. The functionalities specify the operational range of the circuit such as temperature, supply voltage, and the output requirements such as gain, power, speed of the circuit block. They are the targets that the produced circuit must meet. The information from the manufacturing technology on the other hand include manufacturing process statistics, technology constraints like minimal size and space in physical dimensions.

The next step is the topology selection, which remains one of the most creative task for a circuit designer who has to select the appropriate device types and connections to achieve the specifications. This selection is mostly based on the experience of the designer.

Then, for the selected circuit topology, the device parameters, such as transistor dimensions, values of resistors and capacitors, and so on, have to be tuned properly in order to meet the specifications and to increase the robustness over variations in the manufacturing process and the operational environment. This step is called circuit sizing. This step is performed in a looped manner such that the designer must check the simulation results after they make adjustments to certain device parameters.

The steps until now are also referred to as the frontend of the analog design process. Then, the next steps belong to the backend part.

If all of the pre-defined specifications are fulfilled, now it comes to the step of layout, generally performed by layout designers. The placement and interconnections of each device on the chip over different layers are determined at this step. Once the layout is done, parasitic parameters such as strap capacitance and leakage can be extracted and calculated. Re-simulations with these parasitics are necessary. This step is repeated until all specification are met considering layout parasitics.

The next tasks are moved onto fabrication. At first, prototype chips are fabricated and tested. Then any encountered fab problem is fixed. If at this stage the performance of fabricated chip cannot meet the requirements and it is not a fab problem, a redesign from the first step has to be performed, which is obvious a huge waste of time and investment.

Finally, when the chip meets all the specifications, it is ready for production.

1.2.2 Discussions and Challenges

One observation from the above typical analog integrated circuit design flow is that the designed circuit performances are subject to many influential factors, especially the influences due to the uncertainties during circuit operation in real-time, as well as the imperfectness during the manufacturing process.

Although beyond designers' control, those influential factors must be well considered by designers during the design step. The circuit must meet the specifications under the maximal tolerance region of uncertainties from operational conditions and process variations, otherwise several redesign loops are needed, as can be seen in

Figure 1.2. Such redesign loops should be kept as few as possible, since they result in an increase of overall costs and decrease of the time to market.

One typical solution for designers to meet those requirements is to separate their design/sizing process into several steps, such as nominal design and design centering. During the nominal design step, no tolerance region is considered. It is mainly used for the architecture investigation, served as the starting point for the design centering as well. During design centering, the tolerance region of process variations and operational conditions are considered. Certain mathematical models are built up for different uncertainty sources. The designers then refine their design with consideration of the uncertainties by the help of those mathematical models, in order to make sure that their circuit can meet all the specifications under all circumstances.

Another observation from the analog design flow is that, in contrast to the digital part, the above analog design flow is mostly done manually. The analog design automation is generally available only for the circuit simulation step.

From the design of the circuit to the circuit layout, most of the steps in the analog circuit design flow still require experience from designers and layout engineers. As the circuit complexity grows and many challenges arise as discussed in Section 1.1, it is pointed out by numerous studies that the analog parts of the chip design are most frequently at fault when chips fail at first silicon [BC10]. It means a huge re-design cost, if the initial design cannot meet specifications considering possible side effects and challenges. Thus new methodologies for analog design automation are needed considering effects such as process variations and lifetime parameter degradations.

1.3 Contributions of this Thesis

1.3.1 Study on Joint Effects of Process Variations and Transistor Aging

This thesis studies deeply the joint effects of manufacturing process variations and transistor aging. The state-of-the-art methods for design centering considering process variations, and solutions towards transistor aging are studied thoroughly. The physical modeling and behavior as well as the impact of various transistor aging issues are focused. The general problem of the joint effects is formulated analytically as optimization problems.

1.3.2 Design Flow for Lifetime Robustness Optimization

This thesis extends the formulations and applications of the so-called worst-case distance, which is a measure of the design robustness over process variations and operating conditions, into reliability modeling and optimization considering transistor aging over lifetime. The aged worst-case distance in lifetime can be used to study the aged yield value.

A new design flow is proposed to optimize the lifetime robustness of analog circuits, by optimizing the fresh circuit with the checking of both fresh and aged sizing rules, as well as maximum layout area constraints, to achieve x-sigma robustness in circuit's lifetime. Then the lifetime robustness of the circuit is analyzed by the evaluation of aged worst-case distance values.

By applying the design flow repeatedly with different maximum area constraints, the trade-off between circuit's lifetime robustness and the price we pay in terms of the circuit layout area can be obtained. Circuit designers can choose from different product reliability categories with an acceptable area overhead.

1.3.3 Analytical Modeling for Aged Yield Prediction

This thesis proposes a modeling and prediction framework to predict the aged worst-case distance value and the corresponding lifetime robustness of analog circuits. The proposed method is based on the sensitivity analysis of transistor parameters over aging, as well as the sensitivity analysis of the circuit robustness over transistor parameters.

It does not involve either analytical formulation of circuit performance or Monte-Carlo simulations. In comparison to the aged yield analysis based on the geometrical yield modeling, the proposed method is more efficient in obtaining the aged worst-case distance values.

Using the proposed method, circuit designers can obtain quickly an overview of the lifetime robustness of their design, since the fresh worst-case distance is already available for a fresh-optimal design. Certain weakness in the lifetime robustness of their design can be obtained early and quickly, thus reducing the redesign cost.

1.4 Previous Publications

During the past four years, parts of the work presented here were published in [GP09], [PG09], [PG10b], [PG10c], [PG10a], [PG11a], [PG11b] and [PG12]. A two-step reliability optimization flow involving a fresh yield optimization step for the fresh circuit and a lifetime yield optimization step for the aged circuit was detailed in [GP09] and [PG09]. Its software demonstration was presented in [PG10a]. To speed up the analysis of the aged yield value of the circuit, a linear approximation model was introduced in [PG10b], while in [PG10c] several improvements were detailed. The layout area cost for the reliable design was presented in [PG11a]. In [PG11b] the detailed trade-off between circuit reliability and layout area cost was analyzed. An improved version with study into each transistor area and circuit performance was published in [PG12].

1.5 Organization of this Thesis

The rest of the thesis is organized as follows. Chapter 2 discusses in detail of the reliability issues of the modern analog integrated circuit design process. Chapter 3 gives the problem formulation of the work presented in this thesis. Special focus is on the formulation of both fresh and aged yield on different design spaces, as well as the fresh and aged sizing constraints. The analysis and requirements concerning statistical analysis methods are discussed in detail. Chapter 4 studies the problem of robustness optimization by fresh yield optimization with consideration of both fresh and aged sizing rules. The fresh circuit is over-designed such that it is tolerant of both process variations and transistor aging. Then Chapter 5 proposes an analytical prediction model based on sensitivity analysis to approximate the aged worst-case distance value and its corresponding aged yield. The model can be used to predict the age of the circuit as well, providing the acceptable aged yield value as the input. The experimental results on different circuitries using industrial models are given in Chapter 6. Finally Chapter 7 concludes the thesis.

1.6 Summary

The continuous scaling of semiconductor technology into nanometer scale contributes to the higher chip densities, circuit performances, lower cost per transistor, as well as several challenges and side effects, which will limit the product yield value

after manufacturing and in circuit lifetime. Among those hazards, most influential problems arise from manufacturing process variations and transistor degradation related lifetime circuit reliability.

The thesis concentrates on the sizing methodology solutions to the joint effects of process variations and transistor aging. New modeling and prediction framework will be introduced in the thesis.

Chapter 2

Reliability Issues

This chapter presents in detail the reliability issues studied in the thesis. The reliability issues in general relate with the uncertainties of the produced circuits in operating time, in comparison to the figure of merit specified during design time. Section 2.1 covers the manufacture process induced variations and the resulting uncertainties of the manufactured circuits. Section 2.2 introduces the important degradation effects occurred in operating time. Section 2.3 introduces and discusses about the current solutions in solving manufacturing process variations and transistor aging problems.

2.1 Process Variations

The modern semiconductor manufacturing normally consists of series of processing steps. From now on we focus on the CMOS technology as it is used in most Very Large Scale Integrated (VLSI) or Ultra Large Scale Integrated (ULSI) circuit chips [Bak08]. Typically those processing steps are performed on ultrapure, defect-free slices of silicon wafers, and photolithography is used repeatedly to build up various features on different locations through multiple layers on the surface of the wafer.

The variations induced during the manufacturing process can be both systematic and random [Nas08]. The systematic variations, or intra-die variations, refer to those variations occurring repeatedly over many chips or wafers, i.e., at system level. Examples of the systematic variations can be wafer-level variations due to layout-induced strain, optical-proximity correction [Sah10], the rapid ramp-rate of the lamp thermal annealing process [ea06], etc. The random variations or inter-die variations, on the other hand, refer to the fluctuations which happen in a statistical manner during the manufacturing process such as thermal oxidation, doping process, etc. Examples of

random variations can be random discrete doping, line-edge roughness, line-width roughness, interface roughness [Sah10], etc. They contribute to the variations of each transistor's threshold voltage or oxide thickness, and so on.

In comparison to the systematic variations which can be addressed either by making changes to the design or by improvements in the manufacturing process, the random variations can only be tolerated if the initial design has enough margins built by the designers. In other words, the designers have to consider during the design phase the worst case scenario that may happen during the manufacturing process to ensure that the circuit can work properly under process variations.

2.2 Reliability

2.2.1 Reliability Function $R(t)$ and Failure Rate $z(t)$

In traditional reliability engineering, *Reliability Function* and *Failure Rate* are two very important indicators of the device reliability properties. The study presented in this thesis is closely linked to the evaluation and approximation of the reliability function and failure rate of analog integrated circuits. While detailed discussion will be presented in later chapters, here some basic introduction and definitions regarding these reliability engineering terms are given.

The term *Reliability* is defined as the probability that a device will function without failure over a specified time period or amount of usage, according to the IEEE Standard Dictionary of Electrical and Electronic Terms [RI97].

Here, the term *amount of usage* refers to those kinds of one-shot items, such as electronic fuses, safety matches, etc., the usage of which can be divided into two phases: a non-active phase and an active phase. Since analog circuits mainly operate continuously, we focus our discussion only on the continuous operation devices hereafter, i.e., the term *a specified time period* is of interest here.

The reliability thus can be defined as follows. Assume the lifetime of a device is a random variable, denoted by X, and its cumulative distribution function $F(t)$ corresponds to the probability that X will not exceed a certain t, i.e., $F(t) = \text{prob}(X \leq t)$ Then, the reliability of the device, denoted by $R(t)$, is

$$R(t) = \text{prob}(X > t) \quad (2.1)$$
$$= 1 - F(t) \quad (2.2)$$

$R(t)$ is the so-called *Reliability function*, while $F(t)$ is the so-called *lifetime distribution function* [BJ77]. The above definition comes from the fact that the reliability of a device at time t is also the probability that the lifetime of the device will exceed t. In other words, $1 - R(t)$ equals the value of the lifetime distribution function at t. Three observations from (2.2) can be made:

1. when $t = 0$, $R(0) = 1$;
2. when $t \to \infty$, $\lim_{t\to\infty} R(t) = 0$;
3. $R(t)$ must be an non-increasing function of time t.

The first observation implies an important assumption that, at $t = 0$, all of the devices are just manufactured and all of them can work properly. At this time, no aging effect happens, and no transistor parameter drifts due to reliability issues. The second observation can be stated also as all of the devices have their maximum lifetime, beyond which they will not work properly any more. And the last observation comes from the definition of $R(t)$.

The failure rate $z(t)$, on the other hand, comes from such a probability evaluation. Considering a small time interval between t and $t + dt$, the product $z(t)dt$ is thus the probability that a device is failed during this time interval dt, given the condition that it works properly at least until t:

$$\begin{aligned} z(t)dt &= \text{prob}(t < X < t + dt | X > t) \\ &= \frac{\text{prob}(t < X < t + dt)}{\text{prob}(X > t)} \\ &= \frac{F(t + dt) - F(t)}{R(t)} \end{aligned} \tag{2.3}$$

$z(t)$ can be obtained if (2.3) is divided by dt:

$$z(t) = \frac{F(t + dt) - F(t)}{dt} \cdot \frac{1}{R(t)} = \frac{f(t)}{R(t)} \tag{2.4}$$

where $f(t)$ is the lifetime probability density function, defined as

$$f(t) = \frac{dF(t)}{dt} \tag{2.5}$$

$$= -\frac{dR(t)}{dt} \tag{2.6}$$

The failure rate $z(t)$ is also known as hazard rate, or hazard.

The relationship between $R(t)$ and $z(t)$ can be obtained from (2.4). Since

$$z(t) = \frac{f(t)}{R(t)} = \frac{-dR(t)/dt}{R(t)}, \tag{2.7}$$

we can get $R(t)$ by integration:

$$R(t) = \exp\left[-\int_0^t z(\xi)d\xi\right] \tag{2.8}$$

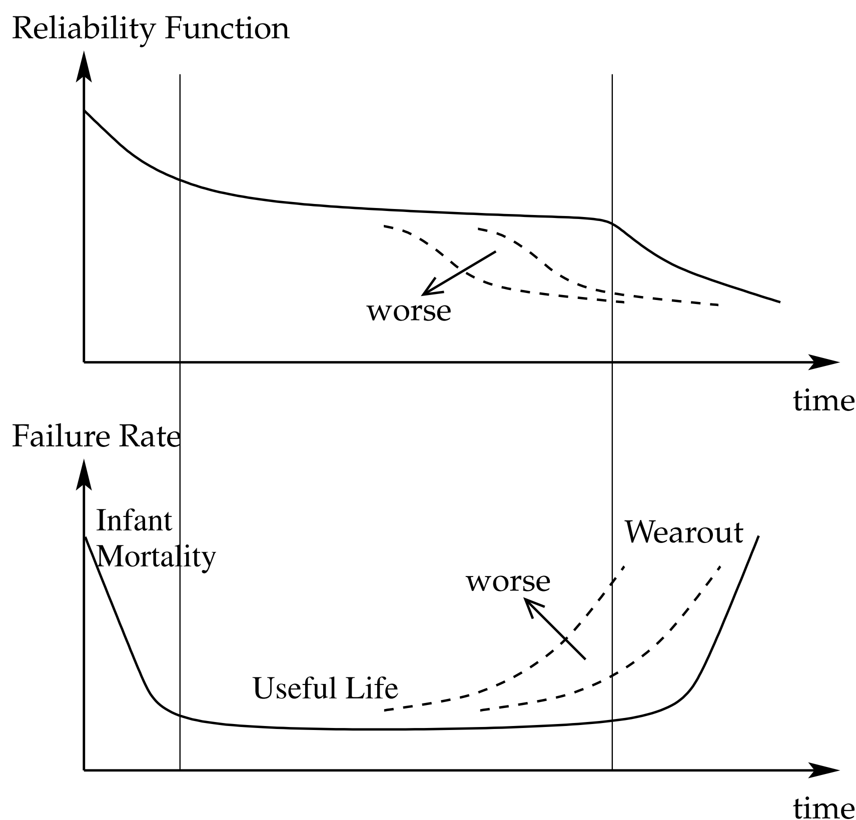


Figure 2.1: The Bathtub Curve with effects of the device increasing wearout degradations.

The typical reliability curve is the so-called *bathtub curve* [BJ77], [KKW03], [Hjo80], as shown in Figure 2.1. The name *bathtub* comes from the shape of the failure rate curve $z(t)$ in the lower part of Figure 2.1.

Three typical regions can be identified in Figure 2.1.

- The first is the "infant mortality" period, during which the devices may fail due to initial weakness or defects. The failure rate of this period often drops quickly, until reaching a relatively constant level.
- Now it is the second period when the devices are in their useful normal operating life. In this period the failure rate is approximately constant and very small. It is also called intrinsic failure period. As the time proceeds, the devices gradually degrade due to various aging effects.
- Then the system enters the last period, the wearout period. The failure rate of this period is increasing, and the whole system gradually reaches the end of its useful lifetime.

An example function concerning $R(t)$ and $z(t)$ is from exponential distribution, which is useful when approximating $R(t)$. In this case, the lifetime distribution function $F(t)$ is $1 - e^{-\lambda t}$, where λ is a positive constant. According to (2.2) and (2.7), $R(t)$ and $z(t)$ can be expressed as

$$R(t) = e^{-\lambda t} \tag{2.9}$$
$$z(t) = \lambda \tag{2.10}$$

where the failure rate $z(t)$ remains constant during the useful lifetime of the product.

Also shown in Figure 2.1 are the effects which may worsen the device reliability due to increasing aging effects, as can be seen on the dotted lines. Such degradation may happen early during the device normal lifetime, causing the failure rate to increase even during the designed useful lifetime of the devices. As introduced in the following, such problem is getting worse as the semiconductor technology continuously scales.

Some of the most important degradation effects on transistors and on-chip interconnects are reviewed in the following sections. Their impacts on the transistor parameters or on the interconnects are discussed. For a more complete discussion, please refer to [HTH+85], [SB03], [AKVM07], [WRK+07], [WSH00].

2.2.2 Negative Bias Temperature Instability

The physical behavior of Negative Bias Temperature Instability (NBTI) on a PMOS transistor is shown in Figure 2.2. As the name indicates, NBTI manifests itself when the PMOS transistor is "negative" biased, i.e., $V_{gs} < 0$. It is commonly accepted that NBTI is the result of hole-assisted breaking of Si-H bonds at Si/SiO_2 inter-

face [AKVM07] when a PMOS is negative biased using the Reaction-Diffusion (R-D) model:

$$\frac{dN_{IT}}{dt} = k_F(N_0 - N_{IT}) - k_R N_H(0) N_{IT} \tag{2.11}$$

where N_{IT} is the fraction of Si-H bonds at the Si/SiO_2 interface which breaks at time t, N_0 is the initial number of all Si-H bonds, and k_F is the dissociation rate constant. The second term in (2.11) describes the annealing process of the released H atoms. $N_H(0)$ is the H concentration at the interface.

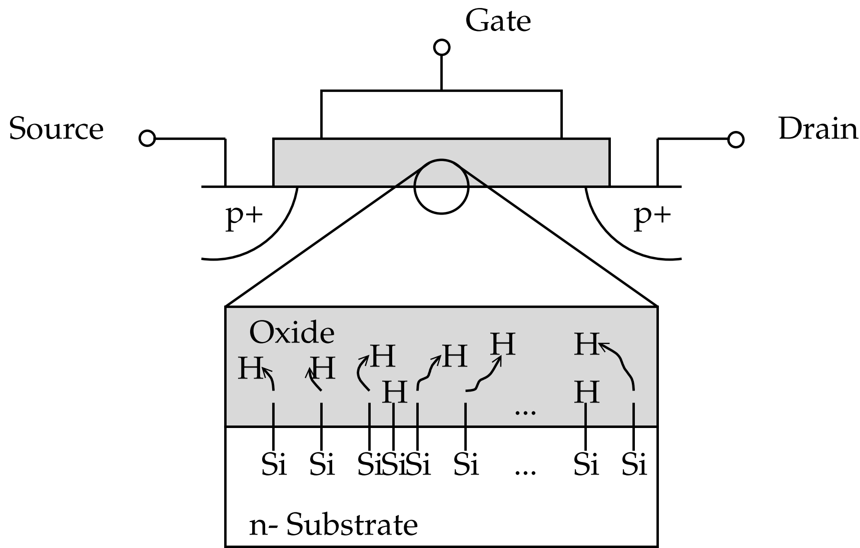


Figure 2.2: Effects of NBTI

NBTI is getting more serious as technology scales, since the vertical oxide field is continuously increasing to enhance transistor performance. Thus a hole in the channel can be easily captured and a two-electron Si-H covalent bond at the Si/SiO_2 interface can be weakened by it. The weakened Si-H bonds break easily at certain high temperature. Atomic H's are released in short time, then they convert to and diffuse as molecular H_2 in long time (>100 s) [AKVM07].

NBTI effect will degrade certain transistor parameters, such as threshold voltage, drain current, transconductance, etc. Threshold voltage degradation due to NBTI is given by [YQD+09]

$$\triangle V_{th} = A \left(\frac{V_{gs}}{t_{ox}}\right)^{\alpha} \exp\left(-\frac{E_a}{kT}\right) t^n \tag{2.12}$$

where K is Boltzmann's constant, A is a process related prefactor, E_a is the activation energy, α denotes voltage acceleration factor, $n = 1/4$ for atomic H in short time, and $n = 1/6$ for molecular H_2 in long time as discussed above.

A well known effect of NBTI on PMOS transistor is its partial recovery, or annealing, when the stress is removed [CCL+03], [RMY03]. Several studies on the modeling of this dynamic behavior and its application in the design of digital circuits are presented in [VWC06], [LWH+07]. For SRAM cell, the impact of fast-recovering NBTI degradation is studied in [DHGSL10]. But for analog circuits, the NBTI recovery is not obvious [JRSR05]. The reason for this is the presence of the constant DC biasing voltage in the most of analog circuits, which leads to a continuous stress voltage applied on the transistors in analog circuits. Such continuous stress voltage is not depend on the input signals. As a result, NBTI recovery or annealing is a minor effect for analog circuits and will be ignored in the rest of this thesis.

The intrinsic variations of NBTI effects are studied in [Rau02]. The expression of variation in $\triangle V_{th}$ shift is

$$\sigma(\triangle V_{th}) = \sqrt{\frac{K t_{ox} \mu(\triangle V_{th})}{A_G}} \tag{2.13}$$

where t_{ox} is effective gate oxide thickness, A_G is its area and K is an empirical constant. As tested by authors in [FAH+08] and [FAH+09], for the transistor parameters V_{th} and I_d, their probability density functions follow a Gaussian distribution pre and post NBTI stress.

It is pointed out in [SB03] that, NBTI should not exhibit any gate length dependence, since it does not depend on lateral electric fields. But NBTI is sometimes enhanced with reduced gate length, which is not well understood yet. The closeness of the source and drain maybe one of the reasons for that.

The introduction of new dielectric material, such as high-κ gate dielectrics (with high dielectric constant κ compared to silicon dioxide), is one of several strategies developed to allow further shrinking recently [wika]. But at the same time, a so-called Positive Bias Temperature Instability (PBTI) effect on a NMOS transistor occurs as the transistor degrades over time if the NMOS transistor is biased positively.

Recently a reliability assessment of voltage controlled oscillators in 32nm high-κ, metal gate technology is presented in [CFSL10] with aging behavior assessment due to NBTI on PMOS transistors and PBTI on NMOS transistors. A detailed study into the impact of analog circuit operations is presented later in [CMFSL11b]. Another study of NBTI and PBTI effects on 6T SRAM memory cell is presented in [DGSL09], showing the significant impact of process variations, NBTI and PBTI on future technologies with new material.

2.2.3 Hot Carrier Injection

Figure 2.3 shows the simplified physical effects of Hot Carrier Injection (HCI) on an NMOS transistor. HCI refers to the injection of channel carriers from the conducting channel under the gate into the gate dielectric. In contrast to NBTI, which happens uniformly in the channel, HCI mainly happens near the drain area where the lateral electric field is high and the channel carriers gain enough kinetic energy during the acceleration along the channel. Hot channel carriers may hit an atom in the substrate, breaking an electron-hole pair or a Si-H bond, and introducing interface traps and a substrate current.

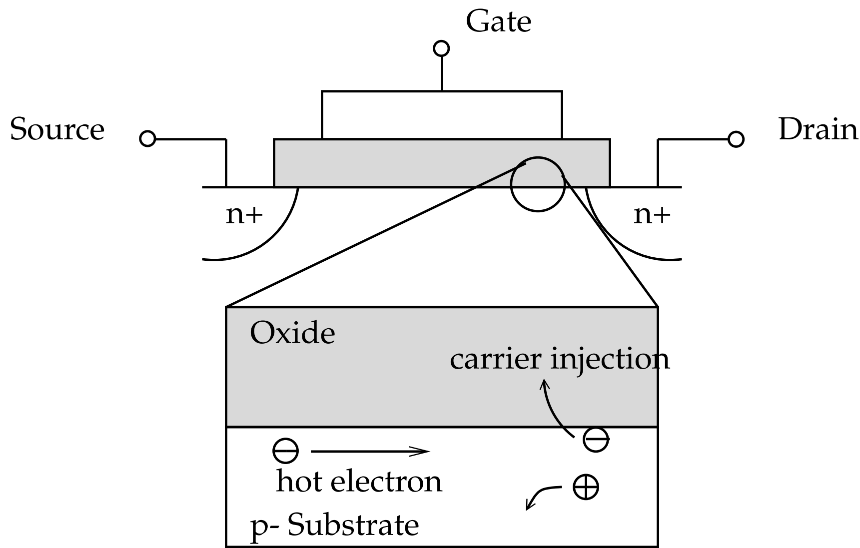


Figure 2.3: Effects of HCI

Traditional modeling method of HCI is by analyzing the substrate current I_{sub} [HTH+85]. The correlation is due to the fact that both hot-carriers and substrate current are driven by a common factor-the maximum channel electric field E_m at the drain end. Some recent research [WRK+07] point out that, as technology scales, I_{sub} will be dominated by various leakage components such as gate leakage, junction current, etc. Authors in [WRK+07] proposed the following reaction-diffusion based model for the degraded parameter ΔV_{th} due to HCI as:

$$\Delta V_{th} = \frac{q}{C_{ox}} K_2 \sqrt{Q_i} \exp\left(\frac{E_{ox}}{E_{o2}}\right) \exp\left(-\frac{\psi_{it}}{q\lambda E_m}\right) t^{n'} \tag{2.14}$$

where Q_i is the inversion charge, ψ_{it} is the trap generation energy and the time exponential constant n' is 0.45.

2.2.4 Time-Dependent Dielectric Breakdown

Time-Dependent Dielectric Breakdown (TDDB) is a reliability issue of the transistor gate oxide. As the technology scales, the thinner gate oxide and the stronger electric fields across the gate oxide can damage the oxide in such a way that the transistor gate current increases, resulting in a totally loss of the isolating property of the gate oxide [WSH00].

There are two types of the dielectric breakdown: soft break down (SBD) and hard break down (HBD). Depending on the number of positions where an increased local gate current occurs, SBD manifests itself as an increase of the leakage current. When the number of such positions and the resulting random traps inside the oxide reache a certain limit, HBD occurs such that the oxide isolating property is completely lost and a percolating path through the oxide will short the gate to the substrate, resulting in a transistor failure [GDWM+08]. The time to HBD can be modeled by a Weibull distribution.

As pointed out in [AWS02], the breakdown of oxides stressed at operating voltages (1.0V-1.5V) can "never be" hard. In addition, authors in [AVK08] show that as supply voltage reduces, the transistor can maintain functional under several SBD paths in the oxide. The positions of SBD paths in the oxide have significant influence here.

2.2.5 Electromigration

Electromigration problem is the reliability issue of the on-chip interconnects [TR07]. In modern technologies, the on-chip interconnects are very thin and narrow. Such a small cross section area of the interconnects will increase the current density that flows through it, which means a movement of a huge amount of electrons. The electron movements then can interact with the metal ions in the interconnects and replace them. As a result, "voids" and "hillocks" are formed in the interconnects. The former, a vacancy area of metal ions, can cause open circuit, or in other words extremely large resistance in the interconnects, corresponding to a failure event, while the latter, locally accumulated metal ions, can cause short circuit between neighboring interconnects, resulting in a malfunction of the circuit.

Electromigration is a reliability effect that must be taken care of during layout phase. Certain interconnects must be widened where current will be high in the operation. Some special layout techniques, such as Slotted Wires, can be applied as well [Lie06].

2.3 State of the Art

2.3.1 Reliability Simulation

Starting from the early 1990's, microelectronic system reliability problems, such as HCI, TDDB, raised due to the rapid advances of fabrication technologies and the emerging VLSI circuits at that time. Several reliability simulators based on software programs were proposed in academia as well as in industry to help the designers gain more insights into their design quality.

- Sheu *et al.* from the University of Southern California, proposed the simulator RELY [SHL89], which simulated the HCI effects based on the substrate current model.
- Leblebici *et al.* from the University of Illinois at Urbana-Champaign proposed a simulation framework considering the dynamic behaviors of HCI, by solving of a set of differential equations at t to obtain the interface trap densities and thus the transistor damage at that time [LK89].
- Hu *et al.* from the University of California at Berkeley proposed the reliability simulation tool BERT [Hu92], which enclosed several modules for different aging effects.
- From industry side, Texas Instruments proposed HOTRON [AHY87] for HCI effects simulation. Philips (later known as NXP) proposed PRESS [LWM+93] for HCI effects simulation.

Entering early 2000's, with the ever shrinking of the device feature size and the emerging of new aging effects such as NBTI, the reliability modeling and simulation again attracted the attention from various communities. [LMM06] presented the recent available EDA tools to simulate the HCI, NBTI and Electromigration effects.

- One of the major commercial tool RelXpert from Cadence Design Systems based on BERT was presented in [LMM06]. The general workflow of RelXpert is shown in Figure 2.4. The prebert and postbert are the internal processors during the aging simulation. The detailed aging simulation using RelXpert is presented in Section 4.3.1.
- The implementation of HCI simulation in another commercial simulator Eldo from Mentor Graphics was described in [KFHR01], where the new `.AGE` command calls repetitive simulations to obtain an accurate prediction of the circuit degradation with dynamic operating conditions. The workflow of such repetitive simulations is shown in Figure 2.5. Inside such a flow, the target time point

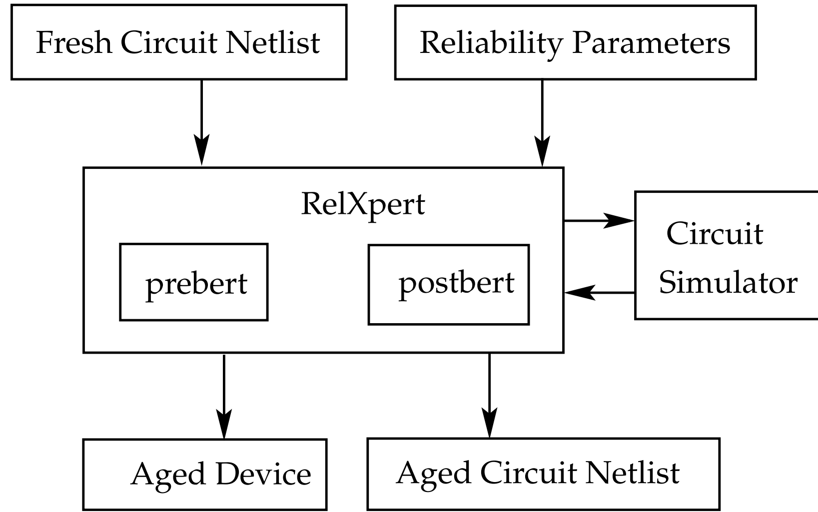


Figure 2.4: General workflow of RelXpert

T_{age} is divided into n smaller intervals T_l. The circuit is simulated at the end of each time intervals, such that the gradual change of bias conditions as a result of the transistor degradation can be simulated.

- The simulator ARET from Georgia Institute of Technology was presented in [XCS+03]. It can handle HCI and Electromigration simulations. For Electromigration effects, the probability of certain post-fabrication defects on interconnects are obtained based on statistical models.
- Li *et al.* from the University of Maryland introduced another reliability simulator MaCRO in [LQH+06]. MaCRO can simulate HCI, NBTI and TDDB by substituting the degradation-sensitive transistors with failure-equivalent circuit models, such that a large number of circuit simulations on small time intervals can be avoided.

2.3.2 Solutions towards Transistor Aging

There are several methods in literature trying to solve the transistor aging issue. They includes initial over-design [KKAR06], smart clock tree signaling [CGRP09], in-situ monitory circuitry [Die07], [QS08], as well as the using of chopper stabilization and autozeroing [MFCSL11].

- Authors in [KKAR06] propose a gate sizing algorithms for digital circuits to initially over-design the circuits. They first calculate the V_{th} degradation for each transistor assuming signal probability for the gate inputs. Then they size the

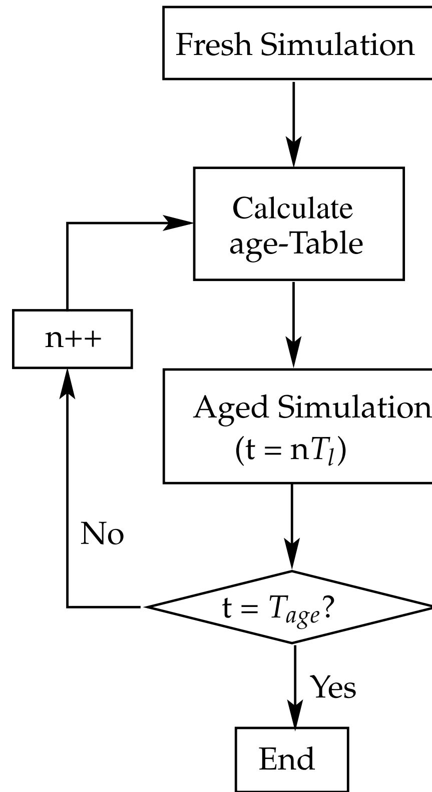


Figure 2.5: Repetitive simulation workflow of ELDO

gates assuming the intended lifetime and the calculated V_{th} degradation, thus achieving a degradation-aware gate sizing. As pointed out by various communities, over-design is straightforward to account for reliability issues. The over-design solutions rely on efficient algorithms to minimize the area overhead while achieving the expected product lifetime reliability.

- Authors in [CGRP09] propose design techniques with low overhead to overcome the NBTI induced skew degradation of clock tree, using a so-called Gating with Both Logic Value (GBLV) scheme. They observe that the PMOS transistors in clock buffers experience alternating stress and recovery stages of NBTI during switches of the clock signal in every cycle. The PMOS transistors in gated clock trees, on the other hand, do not experience such alternating cycles, since that part of the clock tree is shut down by the clock-gating. They generate an auxiliary signal AUX alternating between low and high values, thus balancing NBTI degradations among various clock buffers.
- Authors in [Die07] review the idea of applying additional monitors to the circuits and additional knobs to countermeasure the degradation and other side effects.

The solution can be in both circuit level down to hardware and system level up to intelligent software algorithms to control the circuit behaviors.

- Authors in [QS08] propose an in-situ monitor circuitry to track the effects of NBTI and mitigate the degradation in real time using an adaptive body biasing scheme by forward-biasing the PMOS transistors under stress. Applying the output voltage of the monitor circuitry directly to the body of the PMOS transistors under stress, the tolerance of ΔV_{th} increases in comparison to a PMOS transistor with body connected to Vdd as in normal cases. The deploying of such monitor circuitry, however, turns out to be another trade-off between the additional layout area and the measurement accuracy, since it is impossible to deploy the monitor for each PMOS transistor under stress. In practice one monitor circuitry is allocated for a group of neighboring transistors to reduce area overhead and the influence of local process variations.
- Authors in [MFCSL11] apply chopper stabilization and autozeroing to reduce the effects of transistor aging on circuit level. The methods were originally developed to reduce the offset and low frequency noise. By applying chopper stabilization, the input low frequency noise can be shifted to high frequencies which locate outside the baseband, and the input differential pair are stressed equally resulting in a symmetrical degradation of the transistor pair. By autozeroing technique, on the other hand, the total stress time of the input pair is reduced by one half, since the amplifier operates in close and open loop in an alternative manner.

None of these solutions considers also the manufacturing process variations. Their methods rely only on the nominal value of circuit parameters, which cannot ensure a robust design over process variations.

2.3.3 Design Centering considering Process Variations

On the other hand, considering process variation effects only, design methodologies towards a robust design tolerant of such process variations have been widely studied during the last 30 years. In these so-called design centering problems, an optimal set of circuit parameters are assigned to optimize the yield for an assumed statistical distribution of process variations. The approaches of design centering can be classified into two categories: statistical methods and deterministic methods.

2.3.3.1 Statistical Methods

The feature of the **statistical methods** is the statistical yield analysis using Monte-Carlo simulations [HH64], [Sch66]. The necessary information about the circuit is

collected through those simulations. Based on the Monte-Carlo analysis result of the yield value, the yield can be optimized with the formulations of its gradient and Hessian matrix with respect to the statistical parameters. A variety of techniques and methods concerning the efficiency of Monte-Carlo analysis as well as the statistical yield enhancements are proposed.

Importance sampling [SSG97] is a technique using a different sampling distribution from the statistical parameter distribution to estimate the yield with improved estimation quality. It is applied widely in several methods such as [HLT83], [STPW76], [SP81] and [SR85].

- Authors in [HLT83] propose a stratified sampling method, where the Monte-Carlo simulations are performed in several disjoint subregions of the original parameter perturbation region. The total sample size is reduced by emphasizing the samples in the region where the performance specifications are met. This method is similar to a regionalization method proposed by authors in [STPW76].
- Authors in [SP81] propose a parameter sampling method, where the information in a single run of Monte-Carlo simulation is reused to derive several yield estimations before being updated.
- Authors in [SR85] propose a control variate technique consisting basically two Monte-Carlo experiments. First a control run is done consisting a small number of samples, used to estimate the yield difference between the main circuit and its simulation-cheaper shadow model. Then an auxiliary run with a larger sample size is done to estimate more accurately the yield of the shadow model. The yield of the main circuit can thus be obtained by the results of these two Monte-Carlo analysis.

Another category of the statistical approaches is the statistical experiment-based, such as response surface method. The basic idea is to build up a quadratic model of the circuit performance, and fit the model parameters via a number of samples on the response surface of the circuit. Such model then can replace the original circuit in simulations for a fast yield estimation and enhancement.

- Authors in [YKHT87] propose an average mean-squared error criterion to select an optimal set of circuit simulations in order to derive an accurate performance model.
- Authors in [PH93] develop a statistical regression procedure to estimate the response function of the circuit performances, where the higher order terms are added selectively to improve the accuracy. The yield is then maximized by pseudo objective function substitution method (POSM).

- More recently, authors in [LFJG09] propose a new stop criterion during the evolutionary computation based yield optimization to reduce the number of iterations. They monitor both the average improvement in the whole population of samples and the improvement in the best objective function value. The former part is important especially at the beginning of the algorithm to avoid wrong detection, while the latter part is important especially at the final stage of the algorithm to better locate the local optimal.
- Authors in [LFG10] further reduce the computational effort in each iteration by allocating the computing budget to each candidate in the population in an optimized manner. They identify those critical candidate solutions through an ordinal optimization problem, allocating enough number of samples to the Monte-Carlo simulation of these solutions, in comparison to the few samples allocated to non-critical solutions.

The advantages of the above-mentioned statistical methods are the yield estimation accuracy in comparison to the deterministic methods discussed below. The main drawbacks of the statistical methods include the high simulation efforts, which are reduced for the deterministic methods discussed below.

2.3.3.2 Deterministic Methods

The **deterministic methods**, as its name implies, optimize the yield by approximating and maximizing the acceptance regions, or by building up and maximizing the robustness measures, in a deterministic manner, i.e., based on sensitivities calculation in stead of a number of random samples as in the statistical methods. Then design centering is performed such that either the center of the approximated acceptance region is found, or the robustness measures are maximized.

The acceptance region is defined either on the performance space or parameter space, where the part of the circuit realization after manufacturing process can meet all performance specifications. The exact definition and formulation is detailed later in Chapter 3. Authors in [BGT81] propose the ellipsoidal method, where the yield is maximized by maximizing the volume of the ellipsoid that is inscribed by the acceptance region. Since the acceptance region itself is nonlinear in most cases, its shape is too difficult to determine. So several other papers make simplification for the shape of the acceptance region.

- Authors in [SPV99] use advanced first-order second moment (AFOSM) method to approximate the yield, where the acceptance region is approximated by a polyhedral. The yield is maximized by finding the maximum-volume norm body

contained in the approximated polytope. The authors also propose a unified framework for different design centering task such as tolerance design, worst-case disign, process design, etc., by selecting appropriate norms.

- Authors in [WVO97] replace the single ellipsoidal approximation of the acceptance region by piecewise second-order functions, so-called piecewise ellipsoidal approximation (PEA). The second-order derivative of the constraint is from the boarder region, i.e., the ellipsoid that matches the constraint region. Then, this information is inserted into the second-order Taylor series expansion in the neighborhood of the nominal value. They show this mixed construction is accurate for yield optimization problem.
- Authors in [PSV01] approximate the acceptance region by a general polytope. The yield is then optimized using convex programming approach with an estimation of the yield gradient.
- Authors in [DH77] approximate the acceptance region by a simplex, the number of which is extended in every iteration during yield optimization. The yield is then optimized by finding of the center of the largest hypersphere inscribed into the convex hull of all approximating simplex.
- Authors in [AMHH99] improve the speed of convergence of the ellipsoidal technique by using double-sided ellipsoidal section. The double-sided ellipsoidal is bounded by two hyperplane, the first of which is built up by linearization of the acceptance region boundary at one boundary point, the second of which is found by determining a boundary point at which the gradient of the boundary of the acceptance region is opposite to that of the first hyperplane.

The other type of deterministic optimization methods is building up and maximizing certain robustness measures.

- Authors in [AGW94] propose the formulation of the worst-case distance, which is defined to be the distance between a performance specification and the mean value of that performance in terms of a number of standard deviations. The standard deviation of a performance is formulated by the attributes of statistical parameters which have underlying statistical distribution during manufacturing process. The analysis and optimization of worst-case distances thus are equivalent to the analysis and optimization of the circuit robustness over process variations. A sequential quadratic programming approach is proposed in [Sch03] to solve that optimization formulation. This thesis further extends the idea of the worst-case distance into the time domain. The methodologies of analysis and optimization considering the aged worst-case distance after transistor aging are proposed. The first- and second-order sensitivities of the worst-case distance over

time are derived for the first time, enabling a quick prediction of the aged worst-case distance based on Taylor expansion.

- Authors in [KD95] use a linearized performance penalty (LPP), which is the performance model linearized over the mean value of the statistical parameters. The evaluation of such model requires only one circuit simulation without using an iterative optimization algorithm, with a trade-off over the accuracy.
- Authors in [DK95] optimize the worst-case performance to increase the total yield, where the performance is built up by a response surface model. It is not exactly a design centering approach, but a method to find a design with predefined worst-case performance, i.e., the worst-case robustness.
- Authors in [AS94] and [DG98] make use of the capability indices C_p and C_{pk}, which originate from process control. C_p measures how "narrow" the performance distribution is (the variability part), while C_{pk} measures the distance between the mean value and the most critical performance specifications (the centering part). Their methods are based on new target functions, combining the above two indices, such that the variability can be minimized and design can be centered. The method in [AS94] builds up response surface models for the performances, while the method in [DG98] makes symbolic equations for the performances.

2.3.4 Joint Effects of Transistor Aging and Process Variations

It is only since very recent years that the joint effects of process variations and lifetime parameter degradations are studied [AKPR07]. A various of solutions are proposed in literature. They differ in the type of reliability effects considered and the type of circuits studied.

For digital circuits, NBTI-aware statistical timing analysis considering process variations are proposed in [VOXW09], [VOX09], [WRY+08] and [LSZ+09].

- Authors in [VOXW09] build up a gate-level delay fall-out model by propagating the device parameter fall-out model due to NBTI and process variations into the gate delay model. To study the joint effects on the circuit level with multiple gate stages, they use HSPICE based Monte Carlo simulations. They consider in addition the intrinsic variations of NBTI process in [VOX09]. A sizing methodology considering the joint effects is not covered in their works.
- Authors in [WRY+08] propose a statistical prediction methodology considering process variations and transistor aging due to NBTI. They study the joint effects on gate level delay by applying the transistor level aging model into a process

variation-aware gate delay model. Then they are able to model the timing behavior of a single path considering the joint effects. No sizing solution is proposed in their work either.

- Authors in [LSZ+09] build up an NBTI-aware statistical gate delay model using the stochastic collocation method. They apply their model also into the circuit level statistical timing analysis considering various working conditions of the circuit in runtime. Then they propose a sensitivity analysis framework based on their NBTI-aware circuit level statistical timing analysis, such that the critical gates can be identified and optimized during circuit sizing.

All of those methods rely on the analytical expression of performance features such as delay time, which is suitable for digital circuits but difficult in analog domain.

For analog circuits, various methodologies on the investigation and mitigation of the joint effects are proposed in [MG09], [MG10], [MG11], [MDJG12] and [CMFSL11a].

- Authors in [MG09] use Monte-Carlo simulation loop to obtain the degraded performance values for each fresh random sample at every lifetime point. Then the most appropriate distribution function at each time is fitted, thus a failure distribution throughout the lifetime can be found. It results in a high simulation effort and difficulty for further optimization.
- They improve their method in [MG10] using a response surface model to speed up the simulations, where certain numbers of random samples are still required to obtain the degraded distribution information. They verify that an initial overdesign can improve the lifetime robustness of the circuits. However, a quantified solution is not available from their work to guide the circuit sizing process. The temporal stochastic reliability effects are considered in addition in [MG11] using a similar methodology.
- In [MDJG12], the authors further speed up the simulation on large analog and mixed-signal systems by partitioning of the large system into smaller manageable subblocks. They use fast function extraction symbolic regression method to cope with the high number of dimensions and the nonlinear circuit behavior. An active learning sample selection algorithm is proposed to select optimal model training samples and to limit the amount of expensive aging simulations. No sizing solution is considered either.
- Authors in [CMFSL11a] propose another technique to suppress the effects of aging and process variations on analog circuits. Firstly a Burn-In phase is applied where the asymmetric open-loop stress conditions are switched into symmetric stress to control the BTI effect in saturation. The symmetric stress is generated by switching the asymmetric input stress with a 10Hz clocking frequency. Secondly

a Calibration phase is applied where a selective asymmetric stress is applied to transistors to compensate the offsets caused by process variations. The proposed technique allows smaller device dimensions be used in the design, since offsets can be calibrated after manufacturing.

2.4 Summary

The reliability issues from manufacturing process variations and transistor aging are discussed in detail. These have been the major concern for both circuit design and chip manufacturing communities for decades, since these will result in yield loss and extra redesign costs.

Most of the previous research consider these reliability problems separately. Although there are proposals in solutions towards transistor aging or process variations alone, it is only since recent years that the studies on the joint effects appear. The state-of-the-art studies on the joint effects concentrate on digital circuits, where device parameter variations and aging can be propagated into gate level or circuit level performance formulations. For analog counterparts, the studies are still limited and no sizing solutions are available.

This thesis will study the joint effects of manufacturing process variations and transistor degradation related lifetime circuit reliability in detail, with proposal of new models and new design methodologies for analog circuits.

Chapter 3

Problem Formulation

This chapter formulates the problem studied in this thesis and gives formal definition of terms used throughout the thesis.

3.1 Age and Lifetime

In this section, the differentiation between two terms which are used throughout the thesis, *age* and *lifetime*, is discussed.

Literally, *age* is defined as "length of time that a person or organism has been alive; length of time that an object has existed", while *lifetime* is defined as "span of a person's life, time during which a person is alive; period of time during which something functions or exists". So for a single person or an object, the value of age is always less or equal to the value of lifetime, since the lifetime refers to the whole length of the functioning period of that person or object.

Similarly, in this thesis, age and lifetime are defined as follows.

First, age, denoted by t, is any point of interest on the time axis. Especially, t_0 corresponds to the time when the circuit is just manufactured without any transistor aging. It can be called as fresh circuit.

Second, given a minimal acceptable yield value, Y_{min}, the lifetime, denoted by T_{life}, is the time when the aged yield value $Y(T_{life})$ of a circuit products drops to Y_{min}. In other words, at T_{life}, we have

$$Y(T_{life}) = Y_{min} \tag{3.1}$$

The choosing of Y_{min} will influence the lifetime T_{life} of a circuit products, since the aged yield is a decreasing function over time. If the predefined acceptable Y_{min} drops, the product's lifetime will be longer.

Note that the value of T_{life} can be smaller than, equal to or bigger than the value of t, since T_{life} needs a predefined Y_{min} as an input criteria. In our study t is chosen for any point of interest without a direct indication of the value of T_{life}.

Sometimes people say "lifetime yield", which has the same meaning as "aged yield", i.e., the yield value of an aged circuit. To avoid any misunderstanding, the term "aged yield" is used throughout the thesis.

3.2 Parameters

Parameters of a circuit include all of the contributing factors which influence the behavior of that circuit. These factors can be fixed values, or random variables. They can be from the circuit itself, or from the operating environment. They can drift from their nominal values over time, or remain to be the same amount after manufacturing process.

The circuit parameters can be classified into three categories:

- Design parameters, represented by a vector $\mathbf{d} \in \mathbb{R}^{n_\mathbf{d}}$
- Statistical parameters, represented by a vector $\mathbf{s} \in \mathbb{R}^{n_\mathbf{s}}$
- Operating parameters, represented by a vector $\boldsymbol{\theta} \in \mathbb{R}^{n_{\boldsymbol{\theta}}}$

In addition, if time-dependent parameter drifts are considered, as discussed in Section 2.2, some of the parameters will be a function of age t. Detailed definition and discussion are as follows.

3.2.1 Design Parameters

The *design parameters* $\mathbf{d} = [d_1, d_2, \ldots, d_{n_\mathbf{d}}]^T \in \mathbb{R}^{n_\mathbf{d}}$ correspond to the circuit parameters that the designer can choose during the design phase in order to obtain an "optimal" design. The examples of design parameters in CMOS circuits are transistor widths and lengths, nominal values of capacitors and resistors.

For each of these design parameters, there are correspondingly lower and upper bounds. Usually a lower bound is defined by the manufacturing technology, minimal

grids, for example, while an upper bound may arise from the limit of the maximal available on-chip area. These boundary values can be combined as vectors: $\mathbf{d}_L$ for the lower bounds and $\mathbf{d}_U$ for the upper bounds. Thus a *design parameter space* $\mathcal{D}$ is formed, bounded by an $n_{\mathbf{d}}$-dimensional hypercube:

$$\mathcal{D} = \{\mathbf{d} | \mathbf{d}_L \leq \mathbf{d} \leq \mathbf{d}_U\} \tag{3.2}$$

In Equation (3.2) and the rest of the thesis, the vector inequality is defined as follows. Assume two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n_{\mathbf{x}}}$,

$$\mathbf{x} \leq \mathbf{y} \Leftrightarrow \underset{i=1,\ldots,n_{\mathbf{x}}}{\forall} x_i \leq y_i \tag{3.3}$$

Since either the transistor dimensions or the capacitor and resistor values will not change after manufacturing process, we accept the fact that, design parameters $\mathbf{d}$ will not drift over time. They can only be changed during the design phase, before the manufacturing process starts. Note that for those manual layout designers, the widths and lengths of on-chip interconnects are also designable and may suffer from time-dependent reliability problem, such as Electromigration (EM). But these effects are beyond the scope of this thesis. For more complete discussion of on-chip interconnects reliability problem please refer to [Bla69], [TR07].

3.2.2 Statistical Parameters with Aging

Corresponding to the uncertainty and imperfectness of the manufacturing process, the *statistical parameters* $\mathbf{s} = [s_1, s_2, \ldots, s_{n_{\mathbf{s}}}]^T \in \mathbb{R}^{n_{\mathbf{s}}}$ model the variations during the manufacturing, as introduced in Section 2.1. Such variations can be captured usually by a statistical distribution. In most cases, the probability density functions (pdf) of the statistical parameter distributions are considered.

The types of the statistical distributions vary for different statistical parameters. For example, Normal (Gaussian) distribution for the threshold voltage V_{th}, lognormal distribution for the oxide thickness t_{ox}, etc [Ml95]. These distributions can be transformed into Gaussian distribution as shown in [Esh92]. So without loss of generally, the Gaussian distribution are assumed for the statistical parameters throughout the thesis.

For one statistical parameter s_i, $i = 1, ..., n_{\mathbf{s}}$, it follows Gaussian distribution with mean value $s_{i,0}$ and standard deviation σ_i. Such distribution can be denoted as

$$s_i \sim \mathcal{N}(s_{i,0}, \sigma_i^2) \tag{3.4}$$

The probability density function of s_i is given by

$$\mathrm{pdf}(s_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \cdot \exp\left(-\frac{(s_i - s_{i,0})^2}{2\sigma_i^2}\right) \tag{3.5}$$

For a vector $\mathbf{s}$, the n_s-dimensional Gaussian distribution with mean vector $\mathbf{s}_0$ and covariance matrix $\mathbf{C}$, denoted by $\mathbf{s} \sim \mathcal{N}(\mathbf{s}_0, \mathbf{C})$, has the probability density function as follows:

$$\mathrm{pdf}(\mathbf{s}) = \frac{1}{\sqrt{2\pi}^{n_\mathrm{s}} \cdot \sqrt{\det \mathbf{C}}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{s} - \mathbf{s}_0)^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{s} - \mathbf{s}_0)\right) \tag{3.6}$$

The level contours of the $\mathrm{pdf}(\mathbf{s})$ are ellipsoids:

$$(\mathbf{s} - \mathbf{s}_0)^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{s} - \mathbf{s}_0) \equiv \beta^2(\mathbf{s}) \tag{3.7}$$

where the covariance matrix $\mathbf{C}$ is defined by

$$\mathbf{C} = \mathbf{\Sigma} \cdot \mathbf{R} \cdot \mathbf{\Sigma} \tag{3.8}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{1,2} & \cdots & \sigma_1\sigma_{n_\mathrm{s}}\rho_{1,n_\mathrm{s}} \\ \sigma_2\sigma_1\rho_{2,1} & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_{n_\mathrm{s}-1}\sigma_{n_\mathrm{s}}\rho_{n_\mathrm{s}-1,n_\mathrm{s}} \\ \sigma_{n_\mathrm{s}}\sigma_1\rho_{n_\mathrm{s},1} & \cdots & \sigma_{n_\mathrm{s}}\sigma_{n_\mathrm{s}-1}\rho_{n_\mathrm{s},n_\mathrm{s}-1} & \sigma_{n_\mathrm{s}}^2 \end{bmatrix} \tag{3.9}$$

The matrix $\mathbf{\Sigma}$ has all of the non-negative standard deviations σ_i for every component of vector $\mathbf{s}$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{n_\mathrm{s}} \end{bmatrix} \tag{3.10}$$

The matrix $\mathbf{R}$ has all of the correlations $\rho_{i,j}$ between the i-th and the j-th component of vector $\mathbf{s}$

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n_\mathrm{s}} \\ \rho_{2,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{n_\mathrm{s}-1,n_\mathrm{s}} \\ \rho_{n_\mathrm{s},1} & \cdots & \rho_{n_\mathrm{s},n_\mathrm{s}-1} & 1 \end{bmatrix}, \tag{3.11}$$

where

$$\rho_{i,j} = \rho_{j,i} \tag{3.12}$$
$$-1 \leq \rho_{i,j} \leq +1 \tag{3.13}$$

When the transistor aging effects are taken into consideration, certain statistical parameters, such as V_{th}, will shift their values over time, as introduced in Section 2.2. Thus the vector of statistical parameter can be denoted as a function of age t with aged mean vector $\mathbf{s}_0(t)$ and aged covariance matrix $\mathbf{C}(t)$ as:

$$\mathbf{s}(t) \sim \mathcal{N}(\mathbf{s}_0(t), \mathbf{C}(t)) \tag{3.14}$$

whose probability density function at that time is

$$\mathrm{pdf}(\mathbf{s}(t)) = \frac{1}{\sqrt{2\pi}^{n_\mathbf{s}} \cdot \sqrt{\det \mathbf{C}(t)}} \cdot \exp\left(-\frac{1}{2} \cdot (\mathbf{s}(t) - \mathbf{s}_0(t))^T \cdot \mathbf{C}(t)^{-1} \cdot (\mathbf{s}(t) - \mathbf{s}_0(t))\right), \tag{3.15}$$

with the level contours as:

$$(\mathbf{s}(t) - \mathbf{s}_0(t))^T \cdot \mathbf{C}(t)^{-1} \cdot (\mathbf{s}(t) - \mathbf{s}_0(t)) \equiv \beta^2(t) \tag{3.16}$$

3.2.3 Operating Parameters

The *operating parameters* $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_{n_\boldsymbol{\theta}}]^T \in \mathbb{R}^{n_\theta}$ are used to model the influence of the circuit operating conditions. For instance, supply voltage of the circuit, temperature, capacitive load. These parameters are mostly variable during circuit operations. Such variations are different from that of statistical parameters, since their behavior cannot be modeled by statistical distributions. In other words, these parameters will "fluctuate" during circuit operations, but will neither have any statistical behavior, nor "degrade" monotonically over time.

The operating parameters are usually bounded by lower and upper bounds, $\boldsymbol{\theta}_L$ and $\boldsymbol{\theta}_U$ respectively. Such boundaries specify the maximal operating range of the circuit under which the circuit must work properly, for instance, an operational temperature between $-40°$C and $120°$C. Thus an *operating range* Θ is formed, bounded by an $n_\boldsymbol{\theta}$-dimensional hypercube:

$$\Theta = \{\boldsymbol{\theta} | \boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_U\} \tag{3.17}$$

3.3 Performances with Aging

The *circuit performance* corresponds to the behavior of the circuit. For a typical analog circuit, such as an operational amplifier, the performances can be gain, slew rate, phase margin, common mode rejection ratio, etc. In practical design, performances are the output of a numerical circuit simulation. This is in contrary to the circuit parameters, which are usually the inputs of a numerical circuit simulation.

When no time-dependent transistor aging effects are considered, the performance vector $\mathbf{f} = [f_1, f_2, \ldots, f_{n_\mathrm{f}}]^T \in \mathbb{R}^{n_\mathrm{f}}$ results from a mapping as follows:

$$\mathbf{d}, \mathbf{s}, \boldsymbol{\theta} \mapsto \mathbf{f}(\mathbf{d}, \mathbf{s}, \boldsymbol{\theta}), \tag{3.18}$$

where all of the circuit parameters, $\mathbf{d}$, $\mathbf{s}$ and $\boldsymbol{\theta}$ are considered as the contributing factors of the performance $\mathbf{f}$, and $\mathbf{f}$ is called "evaluated" at $(\mathbf{d}, \mathbf{s}, \boldsymbol{\theta})$.

Thus the *performance space* $\mathcal{F}$ is the set of all possible values of the performance $\mathbf{f}$ resulting from the mapping of any possible value of parameter $(\mathbf{d}, \mathbf{s}, \boldsymbol{\theta})$ in (3.18):

$$\mathcal{F} = \{\mathbf{f} | \mathbf{f} \textit{ evaluated at } (\mathbf{d}, \mathbf{s}, \boldsymbol{\theta})\} \tag{3.19}$$

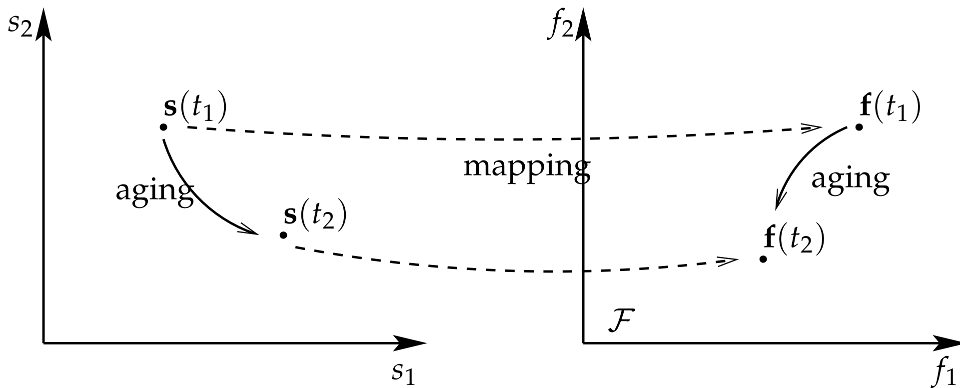


Figure 3.1: Mapping from the statistical parameter space onto the performance space $\mathcal{F}$ considering parameter aging

When the transistor aging effects are taken into consideration, at age t, the performance vector $\mathbf{f}(t) = [f_1(t), f_2(t), \ldots, f_{n_\mathrm{f}}(t)]^T$ is evaluated at the vector $(\mathbf{d}, \mathbf{s}(t), \boldsymbol{\theta})$ as:

$$\mathbf{d}, \mathbf{s}(t), \boldsymbol{\theta} \mapsto \mathbf{f}(t)(\mathbf{d}, \mathbf{s}(t), \boldsymbol{\theta}) \tag{3.20}$$

Figure 3.1 shows qualitatively the aging effects during the mapping from a two-dimensional statistical parameter space onto the two-dimensional performance space

$\mathcal{F}$. As can be seen, at t_1, the statistical parameter vector $\mathbf{s}(t_1)$ maps to the performance vector $\mathbf{f}(t_1)$, while at t_2, after parameter aging, $\mathbf{s}(t_2)$ maps to $\mathbf{f}(t_2)$. The aging-induced performance degradation thus can change the behavior of the circuit. The changed performance value may make it not work properly. In Section 3.4 we will see how this performance degradation influences the aged yield of the circuit.

3.4 Fresh Yield and Aged Yield

3.4.1 Definition

Before introducing the yield, the definitions about performance specification, as well as the acceptance region on the performance space $\mathcal{F}$ and the statistical parameter space will be discussed first.

Each element f_i of $\mathbf{f}$ has a certain lower bound and/or upper bound, denoted as $f_{i,L}$ and/or $f_{i,U}$, for example, lower bound of slew rate, lower and upper bounds of phase margin of an operational amplifier. These boundary values are called *performance specifications*. If no specification is given for a performance f_i, its missing lower or upper bound can be denoted as $f_{i,L} \to -\infty$ or $f_{i,U} \to +\infty$. All performance specifications thus can be formulated as

$$f_i \leq f_{i,U} \wedge f_{i,L} \leq f_i, i = 1, \ldots, n_{\mathbf{f}} \tag{3.21}$$

The performance *acceptance region* $\mathcal{A}_f$ then is the part of $\mathcal{F}$ that satisfies all performance specifications:

$$\begin{aligned} \mathcal{A}_f &= \{\mathbf{f} | f_i \leq f_{i,U} \wedge f_{i,L} \leq f_i, i = 1, \ldots, n_{\mathbf{f}}\} && (3.22) \\ &= \{\mathbf{f} | \mathbf{f}_L \leq \mathbf{f} \leq \mathbf{f}_U\} && (3.23) \end{aligned}$$

We can also define an acceptance region $\mathcal{A}_s$ on the statistical parameter space, according to the mapping in (3.18), as

$$\begin{aligned} \mathcal{A}_s &= \{\mathbf{s} | \underset{\boldsymbol{\theta} \in \Theta}{\forall} \mathbf{f}_L \leq \mathbf{f}(\mathbf{d}, \mathbf{s}, \boldsymbol{\theta}) \leq \mathbf{f}_U\} && (3.24) \\ &= \{\mathbf{s} | \underset{\boldsymbol{\theta} \in \Theta}{\forall} \mathbf{f}(\mathbf{d}, \mathbf{s}, \boldsymbol{\theta}) \in \mathcal{A}_f\} && (3.25) \end{aligned}$$

Figure 3.2 shows the acceptance region (marked in grey) on the statistical parameter space and the performance space. On the statistical parameter space, the ellipsoids correspond to the level contours of the pdf of the Gaussian distributed two-dimensional statistical parameter. In contrast to $\mathcal{A}_f$, which is a tolerance region in

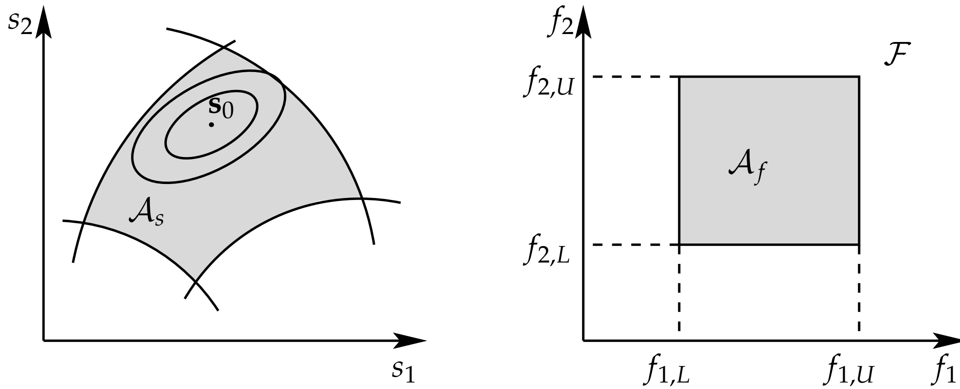


Figure 3.2: Acceptance region (marked in grey) on the statistical parameter space (left) and the performance space (right)

box shape according to the performance specifications, the region $\mathcal{A}_s$ on the statistical parameter space is usually non-linear.

Consider the performance specifications at t_0 right after manufacturing process, the **Yield** of the circuit products, denoted by Y, or more specifically, the parametric yield value, is the percentage of the circuit products after manufacture that can satisfy all of the performance specifications, as formulated in (3.21). In other words, it corresponds to the percentage of the circuits whose performances fall into the performance acceptance region $\mathcal{A}_f$, which is defined in (3.22). It is equivalent to the following probability:

$$Y = \text{prob}\{ \underset{\boldsymbol{\theta}\in\Theta}{\forall} \mathbf{f}_L \leq \mathbf{f}(\mathbf{d}, \mathbf{s}, \boldsymbol{\theta}) \leq \mathbf{f}_U \} \quad (3.26)$$

$$= \text{prob}\{ \underset{\boldsymbol{\theta}\in\Theta}{\forall} \mathbf{f} \in \mathcal{A}_f \} \quad (3.27)$$

At this moment, no transistor aging has yet happened. The resulting parametric yield value is the **Fresh Yield** of the circuits. From the definition it can be observed that, the fresh parametric yield value considers manufacturing process variations induced performance variations and the accordingly specification violations.

Consider the mapping as shown in Figure 3.1, the fresh parametric yield can also be defined on the statistical parameter space using the definition of $\mathcal{A}_s$ in (3.24) as

$$Y = \text{prob}\{ \underset{\boldsymbol{\theta}\in\Theta}{\forall} \mathbf{s} \in \mathcal{A}_s \}, \quad (3.28)$$

which is equivalent of the integration of pdf of $\mathbf{s}$ in $\mathcal{A}_s$:

$$Y = \int_{\mathbf{s}\in\mathcal{A}_s} \cdots \int \mathrm{pdf}(\mathbf{s})\mathrm{d}\mathbf{s} \tag{3.29}$$

An important observation from Figure 3.2 concerning the yield definition is that, the region $\mathcal{A}_s$ on the statistical parameter space cannot be calculated explicitly from performance specifications as shown in (3.24). In other words, whether or not a circuit statistical parameter vector lies inside the region $\mathcal{A}_s$ can only be checked on the performance space by means of circuit simulations.

When the transistor aging effects are taken into consideration, circuit performances degrade over time, as introduced in (3.20), but both of the performance acceptance region $\mathcal{A}_f$ and statistical parameter acceptance region $\mathcal{A}_s$ do not change over time, since the performance specifications are fixed boundary requirements.

Thus the performance acceptance region $\mathcal{A}_f$ at age t is

$$\mathcal{A}_f = \{\mathbf{f}(t)\ |f_i(t) \leq f_{i,U} \wedge f_{i,L} \leq f_i(t), i = 1, \ldots, n_{\mathbf{f}}\} \tag{3.30}$$

$$= \{\mathbf{f}(t)\ |\mathbf{f}_L \leq \mathbf{f}(t) \leq \mathbf{f}_U\} \tag{3.31}$$

and statistical parameter acceptance region $\mathcal{A}_s$ at age t is

$$\mathcal{A}_s = \{\mathbf{s}(t)|\ \underset{\boldsymbol{\theta}\in\Theta}{\forall}\ \mathbf{f}_L \leq \mathbf{f}(\mathbf{d}, \mathbf{s}(t), \boldsymbol{\theta}) \leq \mathbf{f}_U\} \tag{3.32}$$

$$= \{\mathbf{s}(t)|\ \underset{\boldsymbol{\theta}\in\Theta}{\forall}\ \mathbf{f}(\mathbf{d}, \mathbf{s}(t), \boldsymbol{\theta}) \in \mathcal{A}_f\} \tag{3.33}$$

Figure 3.3 shows the acceptance region (marked in grey) on the statistical parameter space and the performance space, with consideration of transistor aging. On the statistical parameter space, the two groups of ellipsoids correspond to the level contours of the pdf of the Gaussian distributed two-dimensional statistical parameter before (at t_0) and after (at t_1) transistor aging.

Considering transistor aging on top of the manufacturing process variations, the **Aged Yield** $Y(t)$ at t corresponds to the percentage of the circuits which still can satisfy all of the performance specifications. Using (3.14), (3.15) and (3.20), $Y(t)$ can be formulated accordingly as

$$Y(t) = \mathrm{prob}\{\underset{\boldsymbol{\theta}\in\Theta}{\forall}\ \mathbf{f}(t) \in \mathcal{A}_f\} \tag{3.34}$$

$$= \mathrm{prob}\{\underset{\boldsymbol{\theta}\in\Theta}{\forall}\ \mathbf{s}(t) \in \mathcal{A}_s\} \tag{3.35}$$

$$= \int_{\mathbf{s}(t)\in\mathcal{A}_s} \cdots \int \mathrm{pdf}(\mathbf{s}(t))\mathrm{d}\mathbf{s}(t) \tag{3.36}$$

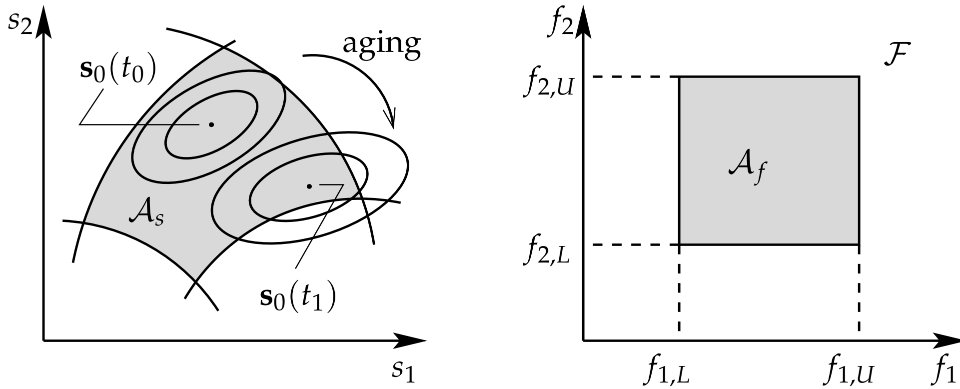


Figure 3.3: Acceptance region (marked in grey) on the statistical parameter space (left) and the performance space (right) with transistor aging

As can be seen in Figure 3.3, since the original distribution around $\mathbf{s}_0(t_0)$ will shift to a new distribution with new mean vector $\mathbf{s}_0(t_1)$ and covariance matrix $\mathbf{C}(t_1)$, a certain percentage of the fresh circuits which satisfies the specification at t_0 will fall out of the acceptance region at t_1, causing an aged yield loss.

It is important to note that, at t_1, after parameter aging happens, the mean vector of the aged statistical parameter, $\mathbf{s}_0(t_1)$, may still satisfy all of the performance specifications. But it cannot ensure that the whole distributions of the statistical parameter at t_1 remain inside the statistical parameter acceptance region $\mathcal{A}_s$ at t_1. The problem may get even worse if the spread of the aging process itself is taken into account. In that case, the distribution of certain statistical parameters may spread even more widely than that of the fresh circuit, resulting in more serious loss of the aged yield value.

3.4.2 Statistical Analysis Method

As mentioned in the last section, the non-linear feature of the acceptance region $\mathcal{A}_s$ on the statistical parameter space makes the calculation of the yield according to the integration as in (3.29) a difficult task.

For integrated circuit designers, the traditional method of the circuit yield estimation is the statistical analysis using Monte-Carlo simulations. The information regarding process variations are provided by the manufacturer in terms of statistical distributions of certain modelcard parameter values. In this case, those values of modelcard parameters are the results of the calculation of certain statistical parameters, in stead of a fixed value as in nominal cases.

Then, Monte-Carlo method calculates the circuit parametric yield value by random sampling inside the statistical parameter space according to the provided distributional information. The performance value corresponding to each sample is checked over the predefined performance specifications. Since the output of each sample is either 1 (pass) or 0 (fail), such an experiment is actually a Bernoulli Trial [GS97].

If the real theoretical yield is denoted by Y, the experimental yield is denoted by $\hat{Y}$, the total number of samples is denoted by N, then $\hat{Y}$ is obtained by

$$\hat{Y} = \frac{\Sigma_{i=1}^{N} X_i}{N} \tag{3.37}$$

where X_i is either 1 or 0, corresponding to the output of each Bernoulli Trial sample.

According to the law of large numbers [GS97], $\hat{Y}$ and Y follows:

$$P(|\hat{Y} - Y| \geq \epsilon) \rightarrow 0, \text{ when } N \rightarrow \infty \tag{3.38}$$

where ϵ is any positive real numbers. (3.38) indicates that, theoretically only when the sample size N approaches infinity then the approximated yield $\hat{Y}$ is equal to the theoretical yield Y. This is the condition that of course is not applicable in practice.

In practice, the reliability of such estimation is determined through a confidence level C_L, as well as confidence intervals $\pm\epsilon$. As introduced in [DS98], [Li10], [Pha06], [Gra07], according to Chebyshev inequality, $\hat{Y}$ and Y follows:

$$P(|\hat{Y} - Y| \geq \epsilon) \leq \frac{\sigma_Y^2}{\epsilon^2} \tag{3.39}$$

where σ_Y^2 is the variance of $\hat{Y}$. From probability theory $\hat{Y}$ has expectation Y and variance $Y(1-Y)/N$. If we define a confidence level

$$C_L = \text{prob}\{|Y - \hat{Y}| < \epsilon\}, \tag{3.40}$$

then from (3.39), C_L can be derived as

$$C_L \geq 1 - \frac{Y \cdot (1-Y)}{N\epsilon^2} \tag{3.41}$$

To meet a certain confidence level C_L, the number of samples N must satisfy

$$N \leq \frac{Y \cdot (1-Y)}{(1-C_L)\epsilon^2} \tag{3.42}$$

The confidence interval $\pm\epsilon$ for a given sample size N and confidence level C_L satisfies

$$\epsilon \leq \sqrt{\frac{Y \cdot (1 - Y)}{N \cdot (1 - C_L)}} \tag{3.43}$$

The confidence level C_L tells how often the theoretical yield Y lies within the confidence interval $\pm\epsilon$ of the estimated yield $\hat{Y}$. For example, for a typical value of C_L of 95% and with ϵ equals 5%, we can be sure that, with a probability of 95%, the theoretical yield Y lies within the region $[\hat{Y} - 0.05, \hat{Y} + 0.05]$.

The relationship among the sample size N, the confidence level C_L, as well as the confidence interval $\pm\epsilon$ is clear from (3.41)-(3.43). For a given C_L, a narrower confidence interval requires a larger sample size N to ensure. For a given confidence interval, to reach a higher confidence level, a larger N is also needed. Last but not least, since the product $Y \cdot (1 - Y)$ has a maximal value at $Y = 0.5$, the confidence interval is also a function of the value of Y. Around 50% yield, the confidence interval itself becomes wider, indicating that a larger N is needed in order to achieve the same C_L for other yield values.

Due to the requirements on the sample size N discussed above, for complicated circuits the Monte-Carlo method is limited by its simulation costs. The alternative method for parametric yield evaluation will be presented in detail in Chapter 4.

The determination of the aged yield value $Y(t)$ can also be carried out by Monte-Carlo simulations on the aged circuits. In this case, the aged transistor modelcard with aged model parameters as well as variation information are needed. Using Monte-Carlo simulation, the approximated aged parametric yield value $\hat{Y}(t)$ at t is given by

$$\hat{Y}(t) = \frac{\Sigma_{i=1}^{N} X(t)_i}{N(t)} \tag{3.44}$$

where where $X(t)_i$ is either 1 or 0, corresponding to the output of each Bernoulli Trial sample at t, and $N(t)$ corresponds to the satisfied samples and total sample size at t respectively.

To avoid the simulation overhead as introduced above, the formulation of the aged yield analysis based on geometric approximation of the aged robustness measures will be presented later in Chapter 4.

One advantage of the statistical analysis method like Monte-Carlo simulation is that, the analysis accuracy depends neither on the number of parameters of the circuit, nor the linearity of the performances. This makes Monte-Carlo simulation applicable to a variety of circuits with easy setup and accurate estimation results, provided the sample size is large enough.

3.5 Sizing Rules with Aging

As shown in [SEGA99], [GZEA01], [MGS08], [Mas10], sizing rules of the analog circuits are constraints that must be satisfied during circuit sizing. They include, for example, geometry constraints (e.g., transistor width, length, area) and electrical constraints (e.g., transistor gate-source voltage v_{gs}, drain-source voltage v_{ds}). They are used to ensure the proper function of the circuits, for example, preventing the transistors from entering inappropriate operation regions, or limiting the voltage difference of v_{ds} in a transistor pair to a certain value, etc.

Figure 3.4 shows a most simple current mirror made up a pair of NMOS transistors. Table 3.1 below lists a complete set of sizing rules for this current mirror.

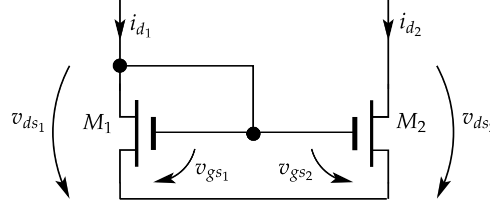


Figure 3.4: Example: a simple current mirror

Table 3.1: Example: sizing rules for a simple current mirror.

	Geometrical	Electrical
Function	$L_2 = L_1$ (3.45) $W_2 = K \cdot W_1$ (3.46)	$v_{gs_{1,2}} - V_{th} \geq 0$ (3.47) $v_{ds_{1,2}} \geq 0$ (3.48) $v_{ds_{1,2}} - (v_{gs_{1,2}} - V_{th_{1,2}}) \geq V_{sat_{min}}$ (3.49) $\lvert v_{ds_2} - v_{ds_1} \rvert \leq \Delta V_{ds_{max}}$ (3.50)
Robustness	$W_1 \geq W_{min}$ (3.51) $L_1 \geq L_{min}$ (3.52) $W_1 \cdot L_1 \geq A_{min}$ (3.53)	$v_{gs_{1,2}} - V_{th_{1,2}} \geq V_{gs_{min}}$ (3.54)

As can be seen from Table 3.1, the sizing rules can be classified into several categories:

- Geometrical and Electrical:
  Geometrical sizing rules refer to constraints of transistor geometries, such as width, length, area. They are the requirements concerning the relationship of geometrical parameters between a pair or a group of transistors, for example, transistors making up of a differential stage, a current mirror, or a level shifter. Electrical sizing rules refer to the constraints of voltages and currents between transistor nodes, such as gate-to-source voltage v_{gs}, drain-to-source current i_{ds}, etc.
- Function and Robustness:
  A less obvious category in contrast to the above is based on the difference between circuit Function and Robustness. Sizing rules concerning function are the critical requirements to ensure the defined function of the transistor and transistor blocks. For example, v_{gs} must be greater than the threshold voltage in order to operate the single transistor (3.47), the saturation condition (3.49) must be met for the transistors in a current mirror to ensure the function of that current mirror. Robustness sizing rules, on the other hand, give the additional requirements for circuit's proper working under variations of operating conditions and process manufacturing parameters. Take again the example of v_{gs}. For robustness it must be greater than the threshold voltage at least the amount of $V_{gs_{min}}$ (3.54).
- Equality and Inequality:
  The difference between equality and inequality sizing rules comes from the constraint formulation itself. The equality sizing rules appear for transistor geometries. They ensure the dependency of design parameters **d** among different transistors, through the relationship of equality, which can be summarized as

  $$\mathbf{c}(\mathbf{d}) = \mathbf{0} \tag{3.55}$$

  Each equality sizing rules reduces the dimension of the design parameter space $\mathcal{D}$ in (3.2) by one, since in this case one design parameter is just the duplicate of the other. Such design space shrink can speed up the sizing process. Inequality sizing rules concern the relationship between either electrical or geometrical parameters. For example, the saturation conditions for transistors, or the minimum values for transistor width or length. These inequality constraints further limit the available range for each design parameter. Without loss of generality, the inequality sizing rules can be summarized as

  $$\mathbf{c}(\mathbf{d}) \geq \mathbf{0} \tag{3.56}$$

  The resulting space of design parameter is now limited to *feasible design parameter space* $\mathcal{D}'$:

  $$\mathcal{D}' = \{\mathbf{d} \in \mathcal{D} | \mathbf{c}(\mathbf{d}) \geq \mathbf{0}\} \tag{3.57}$$

which is the part of design parameter space $\mathcal{D}$ in (3.2) where all inequality sizing rules are satisfied. The corresponding performance space $\mathcal{F}$ from (3.19) becomes the *feasible performance space* $\mathcal{F}'$:

$$\mathcal{F}' = \{\mathbf{f}|\mathbf{f}\ \textit{evaluated at}\ (\mathbf{d}, \mathbf{s}, \boldsymbol{\theta}) \wedge \mathbf{c}(\mathbf{d}) \geq \mathbf{0}\} \tag{3.58}$$

As pointed out in [Mas09], a feasible circuit design does not necessarily meet all performance specifications. The feasible performance space $\mathcal{F}'$ is not the same as the performance acceptance region $\mathcal{A}_f$ in (3.22). Considering both feasibility and performance specification, the *feasible acceptance region* $\mathcal{A}_f{}'$ can be defined as

$$\mathcal{A}'_f = \mathcal{A}_f \cap \mathcal{F}' = \{\mathbf{f}|\mathbf{f}\ \textit{evaluated at}\ (\mathbf{d}, \mathbf{s}, \boldsymbol{\theta}) \wedge \mathbf{f}_L \leq \mathbf{f} \leq \mathbf{f}_U \wedge \mathbf{c}(\mathbf{d}) \geq \mathbf{0}\} \tag{3.59}$$

When transistor parameters degrade over operating time, some electrical sizing rules are influenced. As can be observed from Table 3.1, for example, the electrical sizing rules involving the value of transistor threshold voltage, such as (3.47), (3.49), will change their left-hand-side values over time, due to the drift of V_{th}. These aging-sensitive constraints might be violated during lifetime operation, or even if they are still fulfilled, their distances to the boundary values might decrease, i.e., they might more likely be violated after aging in comparison to the fresh circuit.

As a result, the feasible design parameter space $\mathcal{D}'$ changes over time. At age t the aged feasible design parameter space can be denoted as $\mathcal{D}'(t)$. In the proposed reliability optimization methodology, both fresh and aged sizing rules must be considered and checked.

3.6 Summary

This chapter formulates the basic problem of fresh and aged yield analysis considering process variations and transistor aging. Different types of circuit parameters and performances are defined for future discussion. The statistical analysis method of the fresh and the aged yield values is introduced. The requirements to meet certain confidence level on the sample size of the statistical yield analysis are formulated in detail. As can be seen, a large number of random samples are needed for the statistical analysis method. Though accurate and easy to perform by circuit designers, the statistical yield analysis method needs a huge simulation costs. Further more, the relationship concerning circuit age and lifetime is discussed. The circuit robustness and the total area spent in layout can be two major concern for future circuit design for reliability.

Chapter 4

Aged Yield Optimization with Fresh and Aged Sizing Rules

This chapter proposes the complete sizing flow towards aged yield optimization with fresh and aged sizing rules checking. As already discussed in section 3.4.2, the analysis of fresh and aged yield of the circuit using statistical method is limited by its high simulation costs. Instead, the methodology presented in this chapter is based on the approximation of the acceptance region using so-called worst-case distance.

4.1 Worst-Case Distance

The worst-case distance of a circuit is used to model the circuit robustness and the resulting yield value, considering the manufacturing process variations and various operating conditions. The idea of the worst-case distance was first proposed by Antreich *et al.* in [AGW94]. Its complete descriptions and applications in state-of-the-art design centering methodologies can be found in [Sch03] and [Gra07].

In this section, the idea of the worst-case distance is briefly introduced. It is refined later in the next section into the aged worst-case distance and the corresponding aged yield approximation considering transistor parameter aging.

4.1.1 Yield Analysis and Worst-Case Analysis

At first, two different types of tasks relating with process variations and various operating conditions are explained.

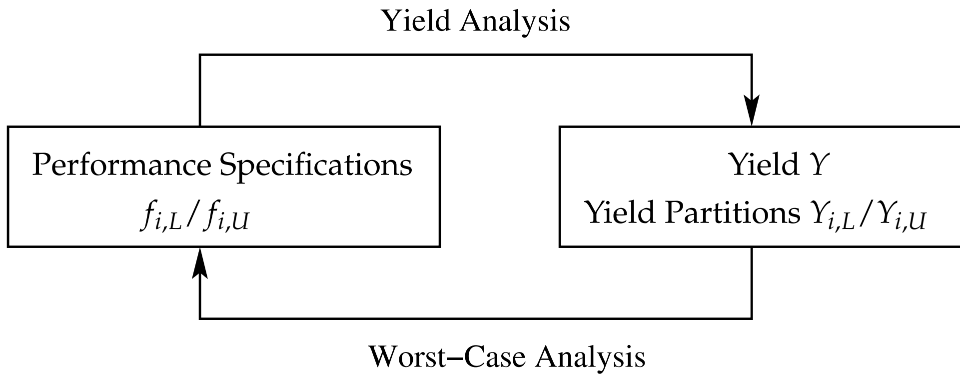


Figure 4.1: The relationship between yield analysis and worst-case analysis.

- The task of *yield analysis* is to compute the parametric yield value (3.26) given the performance specification (3.21) as input. It computes how many percents of the products can satisfy the given performance specification, i.e., what is the parametric yield value considering manufacturing process variations and various operating conditions.
- The task of *worst-case analysis*, on the other hand, is to compute the performance specification that must be accepted, in order to achieve a given parametric yield value. It is the inverse of the yield analysis task. In the worst-case analysis, the design robustness requirements considering manufacturing process variations and various operating conditions, i.e., the parametric yield, are given as input, and the goal is to obtain the worst performance boundary values that have to be accepted to achieve such requirements.

As can be seen, the tasks of yield analysis and worst-case analysis manifest themselves by switching their respective inputs and outputs. Such relationship between the task of yield analysis and worst-case analysis is shown in Figure 4.1. In the context of this thesis, the task of yield analysis is of interest. In this case, the circuits performance specifications (3.21) in the performance space $\mathcal{F}$ (3.19) are given as fixed values, and the goal is to find the corresponding parametric yield values, i.e., the upper half mapping of the Figure 4.1. This mapping will be refined later as we discuss about the worst-case parameters and worst-case distances.

As shown in Figure 3.2 and defined in (3.26) and (3.28), the parametric yield can be computed by two approaches, either in the performance space $\mathcal{F}$ or in the statistical parameter space.

The advantage of the first approach is that the performance specifications (3.21) in the performance space $\mathcal{F}$ are explicitly known. The evaluation of the parametric yield

value in this case is simply done by the Monte-Carlo method (3.37). The accuracy of this method only depends on the sample size N, though, which means a large simulation costs.

If we want to compute the parametric yield by the second approach, we must obtain the performance boundary values (3.21) in the statistical parameter space, which is not known explicitly.

An example is shown in Figure 4.2. Here, one performance upper bound $f_{i,U}$, shown as slashed curve, is considered for the ith performance in the two-dimensional statistical parameter space. Ellipsoids correspond to the level contours of the two-dimensional Gaussian distributed statistical parameters (3.6), while dashed lines correspond to the level contours of the performance f_i in this statistical parameter space.

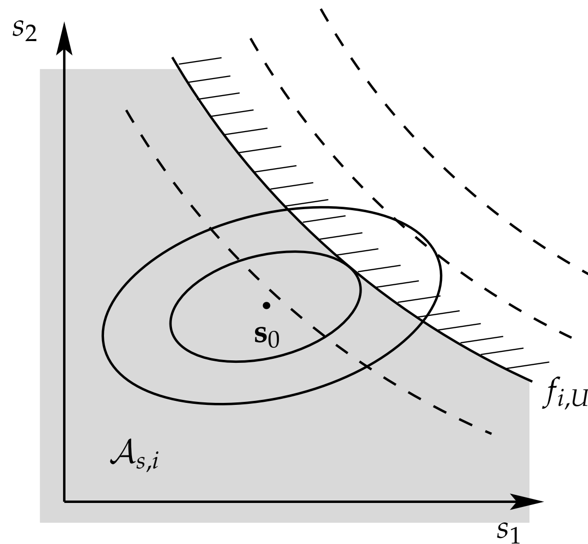


Figure 4.2: Partial acceptance region $\mathcal{A}_{s,i}$ (in grey) in the two-dimensional statistical parameter space with one performance specification $f_{i,U}$ (slashed curve).

As can be seen from Figure 4.2, all of the statistical parameters at the lower-left of the slashed $f_{i,U}$ curve fulfill the specification. Shown in grey, this is the partial acceptance region $\mathcal{A}_{s,i}$, the boarder of which is $f_{i,U}$ in the statistical parameter space. The yield partition in this case is

$$Y_{i,U} = \int_{\mathbf{s}\in\mathcal{A}_{s,i}} \cdots \int \mathrm{pdf}(\mathbf{s})\mathrm{d}\mathbf{s} \tag{4.1}$$

4.1.2 Yield Estimation Based on Worst-Case Distance

Now the yield partition $Y_{i,U}$ in Figure 4.2 will be approximated using worst-case distance.

First, notice in Figure 4.2 that there exists a certain ellipsoid, which is one of the level contours of the Gaussian distributed statistical parameters, that just touches the performance boundary $f_{i,U}$ at one point. We call this point the *worst-case parameter* $\mathbf{s}_{i,w}$, see Figure 4.3. It has such a property that, among all of the parameter points along the performance boundary $f_{i,U}$ in the statistical parameter space, $\mathbf{s}_{i,w}$ has the highest possibility of occurrence. Remember each ellipsoid of the level contours of the Gaussian distributed statistical parameters corresponds to the same probability of occurrence, and the smaller the ellipsoid is, the bigger probability of occurrence it corresponds to.

For the Gaussian distributed statistical parameter $\mathbf{s}$, the level contour corresponding to $\mathbf{s}_{i,w}$ can be formulated from (3.7) as

$$\beta_{i,w}^2 = (\mathbf{s}_{i,w} - \mathbf{s}_0)^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{s}_{i,w} - \mathbf{s}_0) \tag{4.2}$$

and $\beta_{i,w}$ is called *worst-case distance* between the mean value and the boundary of the ith performance feature f_i.

Second, the performance boundary $f_{i,U}$ in the statistical parameter space is linearized at the worst-case parameter point $\mathbf{s}_{i,w}$. The linearization error in this case is minimum, since the linear approximation is accurate enough at the point with highest probability of occurrence. The linearized performance model in the statistical parameter space from $\mathbf{s}_{i,w}$ is

$$\bar{f}_i = f(\mathbf{s}_{i,w}) + \nabla_{\mathbf{s}} f(\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{i,w}} \cdot (\mathbf{s} - \mathbf{s}_{i,w}) \tag{4.3}$$

$$= f_{i,U} + \nabla_{\mathbf{s}} f(\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{i,w}} \cdot (\mathbf{s} - \mathbf{s}_{i,w}), \tag{4.4}$$

where Nabla operator $\nabla_{\mathbf{s}}$ is defined as

$$\nabla_{\mathbf{s}} = \left[\frac{\partial}{\partial s_1}, \frac{\partial}{\partial s_2}, \cdots, \frac{\partial}{\partial s_{n_s}} \right]^T \tag{4.5}$$

and for the linear performance model at the worst-case point, its sensitivity over statistical parameter $\mathbf{s}$ is a constant, which is denoted as $\mathbf{g}_i$:

$$\nabla_{\mathbf{s}} f(\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{i,w}} \equiv \mathbf{g}_i^T = const. \tag{4.6}$$

as can be seen in Figure 4.3.

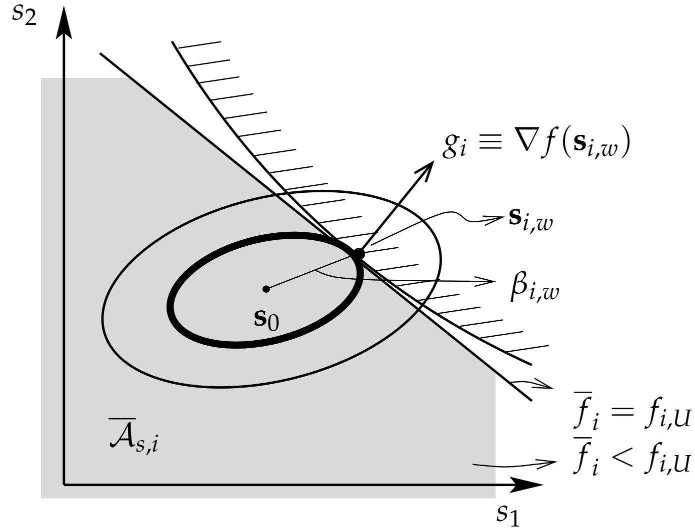


Figure 4.3: The idea of the worst-case distance β_w. $\overline{\mathcal{A}}_s$ is the linear bounded approximated acceptance region.

With the linear performance model at the worst-case point, the partial acceptance region $\mathcal{A}_{s,i}$ now can be approximated by this linear bounded single-side acceptance region $\overline{\mathcal{A}}_{s,i}$:

$$\begin{aligned} \overline{\mathcal{A}}_{s,i} &= \{\mathbf{s}|\mathbf{g}_i^T \cdot (\mathbf{s} - \mathbf{s}_{i,w}) \le 0\} \quad (4.7) \\ &= \{\mathbf{s}|\overline{f}_i \le f_{i,U}\} \quad (4.8) \end{aligned}$$

which corresponds to the grey area in Figure 4.3. Note that the gap between this grey area $\overline{\mathcal{A}}_{s,i}$ and the original performance boundary curve is the error from this linear approximation. Since the linear performance model is accurate at the worst-case point $\mathbf{s}_{i,w}$, the linear approximation error is thus kept minimum.

Based on the approximated acceptance region $\overline{\mathcal{A}}_{s,i}$, (4.1) becomes

$$Y_{i,U} \approx \int \cdots \int_{\mathbf{s} \in \overline{\mathcal{A}}_{s,i}} \mathrm{pdf}(\mathbf{s}) \mathrm{d}\mathbf{s} \quad (4.9)$$

Now we formulate (4.9) in the performance space and relate it with worst-case distance. With Gaussian distributed statistical parameters $\mathbf{s} \sim \mathcal{N}(\mathbf{s}_0, \mathbf{C})$ and (4.3), the linear performance model $\overline{f}_i$ in (4.3) can be transformed into Gaussian distributed performance as

$$\overline{f}_i \sim \mathcal{N}(\overline{f}_{i,0}, \sigma^2_{\overline{f}_i}), \quad (4.10)$$

where

$$\overline{f}_{i,0} = f_{i,U} + \mathbf{g}^T \cdot (\mathbf{s}_0 - \mathbf{s}_{i,w}) \tag{4.11}$$

$$\sigma^2_{\overline{f}_i} = \mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g} \tag{4.12}$$

The pdf of the Gaussian distributed $\overline{f}_i$ thus is

$$\mathrm{pdf}_{\overline{f}_i}(\overline{f}_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{\overline{f}_i}} \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{\overline{f}_i - \overline{f}_{i,0}}{\sigma_{\overline{f}_i}}\right)^2} \tag{4.13}$$

So (4.9) in the performance space is

$$Y_{i,U} \approx \int_{-\infty}^{f_{i,U}} \mathrm{pdf}_{\overline{f}_i}(\overline{f}_i) \mathrm{d}\overline{f}_i \tag{4.14}$$

By inserting (4.13) into (4.14), we have

$$Y_{i,U} \approx \int_{-\infty}^{\frac{f_{i,U} - \overline{f}_{i,0}}{\sigma_{\overline{f}_i}}} \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{1}{2}\xi^2} \mathrm{d}\xi \tag{4.15}$$

The upper limit of the integration (4.15) is equivalent to $\beta_{i,w}$. This can be shortly proved as follows. A complete discussion of $\beta_{i,w}$ concerning performance lower/upper bounds and whether or not the mean value of the statistical parameter falls into the statistical acceptance region will be presented in the next section.

From (4.11) and (4.12), we have

$$\frac{f_{i,U} - \overline{f}_{i,0}}{\sigma_{\overline{f}_i}} = -\frac{\mathbf{g}^T \cdot (\mathbf{s}_0 - \mathbf{s}_{i,w})}{\sqrt{\mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g}}} \tag{4.16}$$

Since the level contour $\beta^2_{i,w}$ (4.2) touches the performance boundary at $\mathbf{s}_{i,w}$, its orthogonal is parallel to $\mathbf{g}$:

$$\mathbf{C}^{-1} \cdot (\mathbf{s}_{i,w} - \mathbf{s}_0) = \lambda \cdot \mathbf{g} \tag{4.17}$$

The following equation can be obtained

$$(\mathbf{s}_{i,w} - \mathbf{s}_0) = \frac{\beta_{i,w} \cdot \mathbf{C} \cdot \mathbf{g}}{\sqrt{\mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g}}} \tag{4.18}$$

so that the right side of (4.16) is equal to $\beta_{i,w}$, and we have

$$\frac{f_{i,U} - \overline{f}_{i,0}}{\sigma_{\overline{f}_i}} = \beta_{i,w} \tag{4.19}$$

$Y_{i,U}$ can now be expressed using worst-case distance as

$$Y_{i,U} \approx \int_{-\infty}^{\beta_{i,w}} \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{1}{2}\xi^2} \mathrm{d}\xi \tag{4.20}$$

From (4.20) the yield value can be obtained immediately for different worst-case distance β_w values. The correspondence between them is shown in Figure 4.4 and Table 4.1 below.

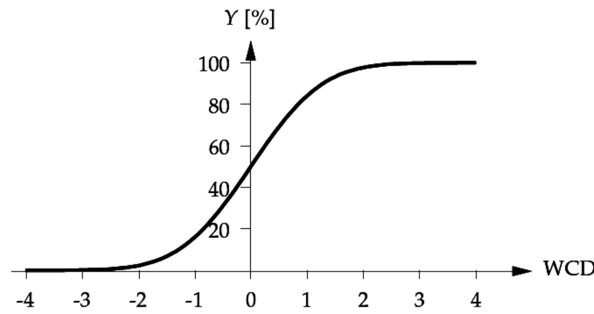


Figure 4.4: Yield values over different worst-case distances.

Table 4.1: Worst-case distances and the corresponding yield value.

β_w	-4	-3	-2	-1	0	1	2	3	4
Yield [%]	0.01	0.13	2.28	15.87	50.00	84.13	97.73	99.87	99.99

The meaning of the worst-case distance can be interpreted by (4.19). For a performance feature, β_w is the distance between the mean value and the worst-case value of that performance. It is measured in the unit "performance standard deviation σ". $\beta_w = 3$ means the worst-case performance value is three times of performance standard deviation away from the mean value of that performance, and it is commonly understood as "β_w-sigma" design.

As can be seen from (4.20) and Table 4.1, the bigger the value of β_w is, the higher the value of yield it corresponds to. One advantage of using worst-case distance as

a measurement of the circuit robustness is that, when the yield value approaches 0% or 100%, the tiny difference between yield value almost disappear, as can be seen in Figure 4.4, whereas the worst-case distance values can still discriminate themselves in such situations. For example, when the worst-case distance equals four, it corresponds to a yield value of 99.99968%, and when the worst-case distance equals five, it represents a yield of 99.99999%. This advantage is very important in the algorithm design for yield optimization/design centering.

Another observation from the above discussion is that, the worst-case distance β_w will be negative, if the mean value of the performance violates the specification. A detailed formulation concerning the negative worst-case distance will be presented later in the next section. Last but not least, for different performance standard deviations, a same distance between the mean value and the worst-case value of the performance corresponds to different worst-case distances and thus different yields.

The overall yield of the circuit can be estimated by a Monte-Carlo analysis on the piecewise linear acceptance region of all performance specifications with no additional simulation cost. The total yield Y is bounded by:

$$1 - \sum_i (1 - Y_i) \lesssim Y \lesssim \min_i Y_i \tag{4.21}$$

Since a smaller worst-case distance during transistor aging leads to more significant yield loss, it is important in the new design flow to analyze and optimize the worst-case distances and the corresponding yield values for the fresh circuit. Those reliability-sensitive worst-case distances should be increased in order to be more robust over transistor aging.

4.1.3 Problem Formulation towards Worst-Case Distance Based Yield Estimation

In practice, neither the worst-case parameter $\mathbf{s}_{i,w}$ nor the worst-case distance is known a priori. Thus formulations towards mathematical optimization solutions are needed. The basic idea can be derived from the above discussion, i.e., the worst-case parameter $\mathbf{s}_{i,w}$ has the highest probability of occurrence among all statistical parameters on the performance specification border [Gra07].

For one of the performance f_i, whose mean value satisfies the specification $f_i \leq f_{i,U}$, to locate the worst-case parameter $\mathbf{s}_{i,w}$ is equivalent to find the statistical parameter vector with highest probability of occurrence among all parameters which are outside or on the boarder of the acceptance region $\mathcal{A}_{s,U,i}$:

$f_i \leq f_{i,U}$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U,i}$:

$$\max_{\mathbf{s}} \text{pdf}_N(\mathbf{s}), \quad \text{s.t. } \mathbf{s} \notin \mathcal{A}_{s,U,i} \tag{4.22}$$

To further extend the formulation in (4.22), we first consider the variations of operating parameters such as operating temperature and supply voltage of the circuit [GWA93]. A statistical parameter vector $\mathbf{s}$ is within the partial acceptance region, $\mathbf{s} \in \mathcal{A}_{s,U,i}$, for a specification $f_i \leq f_{i,U}$ means that, the maximal performance value over operating parameter variations still satisfies that specification, and vice versa, i.e., once the maximal performance value over operating parameter variations violates the specification $f_i \leq f_{i,U}$, then it can be concluded that $\mathbf{s} \notin \mathcal{A}_{s,U,i}$:

$f_i \leq f_{i,U}$:

$$\mathbf{s} \in \mathcal{A}_{s,U,i} \quad \Leftrightarrow \quad \max_{\boldsymbol{\theta} \in \Theta} f_i(\mathbf{s}, \boldsymbol{\theta}) \leq f_{i,U} \tag{4.23}$$

$$\mathbf{s} \notin \mathcal{A}_{s,U,i} \quad \Leftrightarrow \quad \max_{\boldsymbol{\theta} \in \Theta} f_i(\mathbf{s}, \boldsymbol{\theta}) > f_{i,U} \tag{4.24}$$

Another development from (4.22) is that, to maximize the probability density function is equivalent to minimize β, i.e., minimize the size of the equidensity contour according to (3.7).

Thus, considering $n_{\mathbf{f}}$ performance specifications, (4.22) can be reformulated as

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s}), \quad \text{s.t.} \quad \max_{\boldsymbol{\theta} \in \Theta} f_i(\mathbf{s}, \boldsymbol{\theta}) \geq f_{i,U}, \quad i = 1, \ldots, n_{\mathbf{f}}. \tag{4.25}$$

The other cases corresponding to the specification $f_i \geq f_{i,L}$ and whether or not the mean value satisfies the specification can be formulated accordingly as:

$f_i \leq f_{i,U}$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,U}$:

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s}), \quad \text{s.t.} \quad \max_{\boldsymbol{\theta} \in \Theta} f_i(\mathbf{s}, \boldsymbol{\theta}) \leq f_{i,U}, \quad i = 1, \ldots, n_{\mathbf{f}}. \tag{4.26}$$

$f_i \geq f_{i,L}$ and $\mathbf{s}_0 \in \mathcal{A}_{s,L}$:

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s}), \quad \text{s.t.} \quad \min_{\boldsymbol{\theta} \in \Theta} f_i(\mathbf{s}, \boldsymbol{\theta}) \leq f_{i,L}, \quad i = 1, \ldots, n_{\mathbf{f}}. \tag{4.27}$$

$f_i \geq f_{i,L}$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,L}$:

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s}), \quad \text{s.t.} \quad \min_{\boldsymbol{\theta} \in \Theta} f_i(\mathbf{s}, \boldsymbol{\theta}) \geq f_{i,L}, \quad i = 1, \ldots, n_{\mathbf{f}}. \tag{4.28}$$

The solutions of (4.25)-(4.28) give the worst-case distance $\beta_{i,w}$ for each performance specification, and can be used in the yield approximation as introduced in Section 4.1.2.

4.1.4 Solution using Lagrangian Functions

In the following, the analytical form of the worst-case distance β_w is developed using Lagrangian function with its first-order optimality condition [Gra07]. Such formulation is needed in the discussion about the gradient of the worst-case distance later on. The performance index i is left out for simplicity in the discussion below.

Inside (4.25) and (4.26), the evaluation of the constraint requires the solving of the following maximization problem first:

$$\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}), \;\; \text{s.t.} \;\; \boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_U \tag{4.29}$$

The Lagrangian function of (4.29) can be written as

$$\mathcal{L}_U(\boldsymbol{\theta}, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U) = -f(\boldsymbol{\theta}) - \boldsymbol{\lambda}_L^T \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_L) - \boldsymbol{\lambda}_U^T \cdot (\boldsymbol{\theta}_U - \boldsymbol{\theta}) \tag{4.30}$$

Similarly, for the inner optimization problems of (4.27) and (4.28), the Lagrangian function can be written as

$$\mathcal{L}_L(\boldsymbol{\theta}, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U) = f(\boldsymbol{\theta}) - \boldsymbol{\lambda}_L^T \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_L) - \boldsymbol{\lambda}_U^T \cdot (\boldsymbol{\theta}_U - \boldsymbol{\theta}) \tag{4.31}$$

By applying (4.30) and (4.31), the Lagrangian functions of (4.25)-(4.28) can be written as:

$f \leq f_U$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U}$:

$$\mathcal{L}(\mathbf{s}, \boldsymbol{\theta}, \lambda, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U) = \beta^2(\mathbf{s}) + \lambda \cdot (f_U + \mathcal{L}_U(\boldsymbol{\theta}, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U)) \tag{4.32}$$

$f \leq f_U$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,U}$:

$$\mathcal{L}(\mathbf{s}, \boldsymbol{\theta}, \lambda, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U) = \beta^2(\mathbf{s}) - \lambda \cdot (f_U + \mathcal{L}_U(\boldsymbol{\theta}, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U)) \tag{4.33}$$

$f \geq f_L$ and $\mathbf{s}_0 \in \mathcal{A}_{s,L}$:

$$\mathcal{L}(\mathbf{s}, \boldsymbol{\theta}, \lambda, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U) = \beta^2(\mathbf{s}) - \lambda \cdot (f_L - \mathcal{L}_L(\boldsymbol{\theta}, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U)) \tag{4.34}$$

$f \geq f_L$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,L}$:

$$\mathcal{L}(\mathbf{s}, \boldsymbol{\theta}, \lambda, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U) = \beta^2(\mathbf{s}) + \lambda \cdot (f_L - \mathcal{L}_L(\boldsymbol{\theta}, \boldsymbol{\lambda}_L, \boldsymbol{\lambda}_U)) \tag{4.35}$$

Applying Karush-Kuhn-Tucker conditions [Fle87], [NW00] for one of the above Lagrangian functions (4.32), and with (4.2), the solution $\mathbf{s}_{w,U}$ can be obtained:

$f \leq f_U$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U}$:

$$\nabla \mathcal{L}(\mathbf{s}_{w,U}) = 2\mathbf{C}^{-1} \cdot (\mathbf{s}_{w,U} - \mathbf{s}_0) + \lambda_{w,U} \cdot \nabla \mathcal{L}_U(\mathbf{s}_{w,U}) \equiv 0 \tag{4.36}$$

where

$$\nabla \mathcal{L}_U(\mathbf{s}_{w,U}) = -\nabla f(\mathbf{s}_{w,U}) \tag{4.37}$$

from the Lagrangian function (4.30).

The formulation of (4.36) an (4.37) can lead to the representation of the worst-case parameter in the case $f \leq f_U$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U}$:

$$(\mathbf{s}_{w,U} - \mathbf{s}_0) = \frac{\lambda_{w,U}}{2} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U}) \tag{4.38}$$

The worst-case distance in (4.2) can then be represented as

$$\beta_{w,U}^2 = \frac{\lambda_{w,U}^2}{4} \cdot \nabla f(\mathbf{s}_{w,U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U}) \tag{4.39}$$

which gives

$$\frac{\lambda_{w,U}}{2} = \frac{\beta_{w,U}}{\sqrt{\nabla f(\mathbf{s}_{w,U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U})}} \tag{4.40}$$

The worst-case parameter can be represented by applying (4.40) into (4.38):

$$(\mathbf{s}_{w,U} - \mathbf{s}_0) = \frac{\beta_{w,U}}{\sqrt{\nabla f(\mathbf{s}_{w,U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U})}} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U}) \tag{4.41}$$

For the cases $f \leq f_U$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,U}$, the worst-case parameter is

$$(\mathbf{s}_{w,U} - \mathbf{s}_0) = -\frac{\beta_{w,U}}{\sqrt{\nabla f(\mathbf{s}_{w,U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U})}} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,U}) \tag{4.42}$$

The other two cases corresponding to (4.34) and (4.35) can be derived in a similar way:

$f \geq f_L$ and $\mathbf{s}_0 \in \mathcal{A}_{s,L}$:

$$(\mathbf{s}_{w,L} - \mathbf{s}_0) = -\frac{\beta_{w,L}}{\sqrt{\nabla f(\mathbf{s}_{w,L})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L})}} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L}) \tag{4.43}$$

$f \geq f_L$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,L}$:

$$(\mathbf{s}_{w,L} - \mathbf{s}_0) = \frac{\beta_{w,L}}{\sqrt{\nabla f(\mathbf{s}_{w,L})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L})}} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L}) \tag{4.44}$$

To obtain an analytical expression for the worst-case distance $\beta_{w,L/U}$, a linear performance model can be built up at the worst-case parameter vector $\mathbf{s}_{w,L/U}$ in the statistical parameter space as:

$$\overline{f}_w(\mathbf{s}) = f_{L/U} + \nabla f(\mathbf{s}_{w,L/U})^T \cdot (\mathbf{s} - \mathbf{s}_{w,L/U}) \tag{4.45}$$

At the nominal parameter vector $\mathbf{s}_0$, (4.45) gives

$$\nabla f(\mathbf{s}_{w,L/U})^T \cdot (\mathbf{s}_{w,L/U} - \mathbf{s}_0) = f_{L/U} - \overline{f}_w(\mathbf{s}_0) \tag{4.46}$$

Inserting (4.42) and (4.43) into (4.46) gives:

For the cases $f \leq f_U$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0 \in \mathcal{A}_{s,L}$:

$$\beta_{w,L/U} = \frac{\overline{f}_w(\mathbf{s}_0) - f_{L/U}}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.47}$$

$$= \frac{\nabla f(\mathbf{s}_{w,L/U})^T \cdot (\mathbf{s}_0 - \mathbf{s}_{w,L/U})}{\sigma_{\overline{f}_w}} \tag{4.48}$$

Inserting (4.41) and (4.44) into (4.46) gives:

For the cases $f \leq f_U$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,L}$:

$$\beta_{w,L/U} = \frac{f_{L/U} - \overline{f}_w(\mathbf{s}_0)}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.49}$$

$$= \frac{\nabla f(\mathbf{s}_{w,L/U})^T \cdot (\mathbf{s}_{w,L/U} - \mathbf{s}_0)}{\sigma_{\overline{f}_w}} \tag{4.50}$$

It is clear from (4.47)-(4.50) that, if the mean value of the statistical parameter and performance violates the specification, the worst-case distance is negative.

4.2 Aged Worse-Case Distance and Aged Yield

When transistor parameters degrade over time, the statistical parameters shift their distributions in the statistical parameter space as introduced in Figure 3.3. Using the similar modeling method as introduced above, the worst-case distance during transistor aging with respect to one performance specification can be illustrated in Figure 4.5. For simplicity, the performance index i is left out in the discussion hereafter.

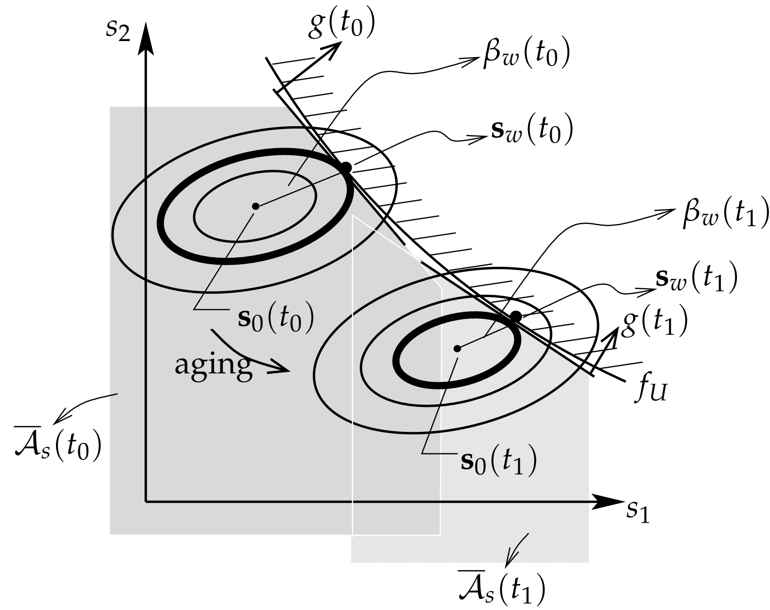


Figure 4.5: The degradation of worst-case distance from t_0 to t_1 with respect to one performance specification f_U (slashed curve).

In Figure 4.5, again, one performance upper bound f_U, shown as slashed curve, is considered in the two-dimensional statistical parameter space. The two groups of ellipsoids correspond to the level contours of the two-dimensional Gaussian distributed statistical parameters at t_0 and t_1 respectively. During parameter aging from t_0 to t_1, both of the mean and the worst-case value of the statistical parameters change, as shown in Figure 4.5.

To approximate the aged yield using aged worst-case distance value, a linear performance model is built up at t_0 and t_1 respectively, from the corresponding worst-case point $\mathbf{s}_w(t_0)$ and $\mathbf{s}_w(t_1)$. The sensitivity of the linear performance model over statistical parameters is $g(t_0)$ and $g(t_1)$ respectively. The linear bounded approximated acceptance region is now $\overline{\mathcal{A}}_s(t_0)$ and $\overline{\mathcal{A}}_s(t_1)$. Note the linear approximation error at

the different points on the performance boundary in this case, which lead to the slight difference in the two approximated acceptance regions.

At time t_1, the linearized performance model from $\mathbf{s}_w(t_1)$ is

$$\overline{f}(t_1) = f_U + \mathbf{g}^T(t_1) \cdot (\mathbf{s}(t_1) - \mathbf{s}_w(t_1)) \, , \tag{4.51}$$

and the corresponding linear bounded approximated acceptance region at t_1 is

$$\begin{aligned} \overline{\mathcal{A}}_s(t_1) &= \{\mathbf{s}(t_1) | \mathbf{g}^T(t_1) \cdot (\mathbf{s}(t_1) - \mathbf{s}_w(t_1)) \le 0\} & (4.52)\\ &= \{\mathbf{s}(t_1) | \overline{f}(t_1) \le f_U\} & (4.53) \end{aligned}$$

With Gaussian distributed statistical parameters at t_1: $\mathbf{s}(t_1) \sim \mathcal{N}(\mathbf{s}_0(t_1), \mathbf{C}(t_1))$, the level contour corresponding to $\mathbf{s}_w(t_1)$ is

$$\beta_w^2(t_1) = (\mathbf{s}_w(t_1) - \mathbf{s}_0(t_1))^T \cdot \mathbf{C}^{-1}(t_1) \cdot (\mathbf{s}_w(t_1) - \mathbf{s}_0(t_1)), \tag{4.54}$$

and the linear performance model can be transformed into Gaussian distributed performance as

$$\overline{f}(t_1) \sim \mathcal{N}(\overline{f}_0(t_1), \sigma_{\overline{f}}^2(t_1)), \tag{4.55}$$

where

$$\begin{aligned} \overline{f}_0(t_1) &= f_U + \mathbf{g}^T(t_1) \cdot (\mathbf{s}_0(t_1) - \mathbf{s}_w(t_1)) & (4.56)\\ \sigma_{\overline{f}}^2(t_1) &= \mathbf{g}^T(t_1) \cdot \mathbf{C}(t_1) \cdot \mathbf{g}(t_1) & (4.57) \end{aligned}$$

The aged yield value at t_1 can be approximated as

$$Y_U(t_1) \approx \int_{-\infty}^{\frac{f_U - \overline{f}_0(t_1)}{\sigma_{\overline{f}}(t_1)}} \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{1}{2}\xi^2} \mathrm{d}\xi, \tag{4.58}$$

where the upper limit of the integration (4.58) can be reformulated, similar to (4.16)-(4.19), using (4.56), (4.57) and (4.54) as

$$\begin{aligned} \frac{f_U - \overline{f}_0(t_1)}{\sigma_{\overline{f}}(t_1)} &= -\frac{\mathbf{g}^T(t_1) \cdot (\mathbf{s}_0(t_1) - \mathbf{s}_w(t_1))}{\sqrt{\mathbf{g}^T(t_1) \cdot \mathbf{C}(t_1) \cdot \mathbf{g}(t_1)}} & (4.59)\\ &= \beta_w(t_1) & (4.60) \end{aligned}$$

The aged yield at t_1, $Y_U(t_1)$, can be expressed using aged worst-case distance as

$$Y_U(t_1) \approx \int_{-\infty}^{\beta_w(t_1)} \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{1}{2}\xi^2} \mathrm{d}\xi \tag{4.61}$$

A complete formulation of the aged worst-case distance concerning performance lower/upper bounds and whether or not the mean value of the aged statistical parameter falls into the statistical acceptance region is shown in Section 4.3.5.

4.3 Design Flow

As transistor parameters degrade over time, the aged worst-case distance values decrease, resulting in an increasing yield loss over time. In the proposed yield analysis and optimization flow, the fresh circuit is analyzed and optimized for x-sigma robustness, both fresh and aged sizing rules as well as the maximum layout area constraint are checked. For those aging-sensitive performances, the corresponding worst-case distances will be increased for the fresh circuit, ensuring a more robust design over process variations and transistor aging.

One assumption of the proposed fresh circuit optimization flow is that, for analog circuits, annealing from NBTI is a minor effect due to the constant existence of dc biasing voltages [JRSR05]. Thus the aged yield decreases over time with the monotonic aging of $\beta_{i,w}(t)$. The robustness of the circuits during the lifetime can be ensured by optimizing fresh circuits with enough design margins.

The core part of the proposed reliability optimization flow is shown in Figure 4.6. The input of the flow is the fresh circuit netlist with initial design parameters. Then, the step of the fresh circuit optimization with one maximum area constraint A_{max} is performed. This step involves the checking of sizing rules for both fresh and aged circuits. The output of this step is the optimal design parameters for the circuit. Then, the aged yield can be analyzed, provided that the aged circuit netlist is obtained. The aged circuit netlist is obtained using AgeMOS model from RelXpert as explained later.

Running the flow repeatedly with increasing maximum area value A_{max} as a constraint, the designer can obtain the trade-off between approximated layout area and circuit reliability, and choose a reasonable area with its acceptable aged yield value.

In the following, Section 4.3.1 gives in detail the simulation flow of the aged circuit and how the aged performances are obtained. Section 4.3.2 explains the checking of fresh and aged sizing rules. Section 4.3.3 discusses the approximation of the circuit layout area. Section 4.3.4 and Section 4.3.5 detail the problem formulations as a numerical optimization problem for the fresh circuit optimization and aged yield analysis.

4.3.1 Simulation Flow of the Aged Circuit

This Section briefly introduces the workflow concerning the generation and simulation of the aged circuit.

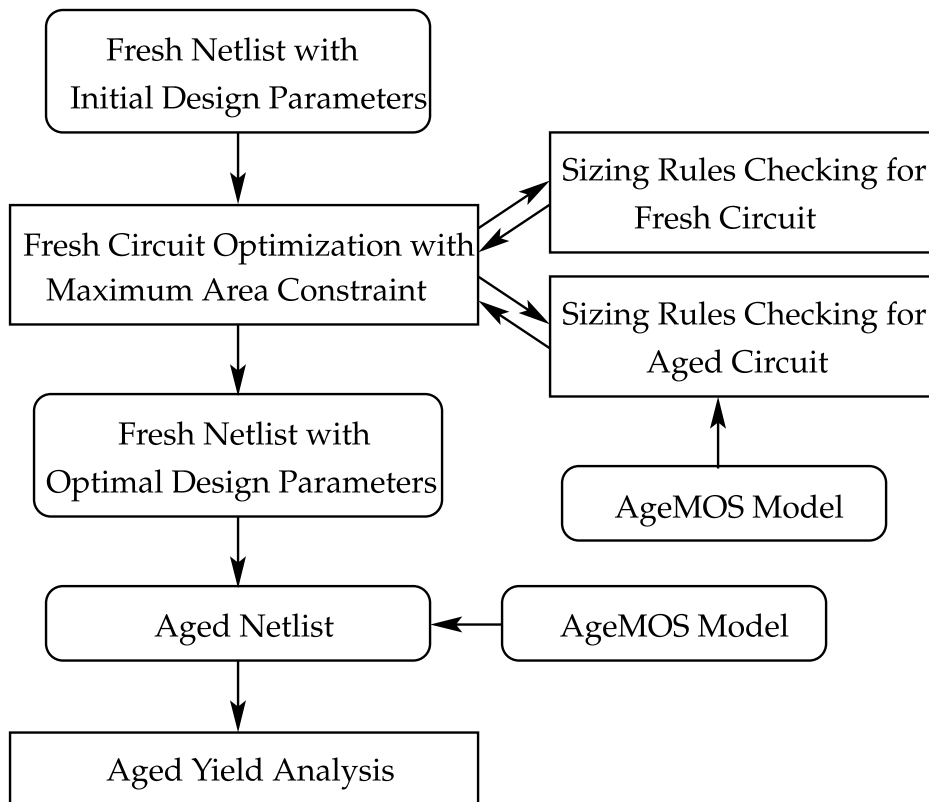


Figure 4.6: Core of the proposed reliability optimization flow with one maximum area constraint.

The complete workflow is shown in Figure 4.7. The round-corner boxes indicate the input and the output of the flow. The tool RelXpert from Cadence [cad] is used for transistor aging calculation and aged circuit netlist generation. The methodology presented in this thesis however does not depend on the specific type of aging calculation tools. Any commercial or academic tools, which can process the circuit aging behavior, calculate and produce the aged netlist, can be adopted as well.

The flow takes two inputs: the fresh circuit netlist with RelXpert parameters, as well as the fresh modelcards with AGEMOS parameters. In the fresh circuit netlist, the circuit designers can specify intended circuit age, temperature, aging modules needed (i.e., NBTI, HCI, etc.), as well as other control statements for RelXpert tool, such as accuracy. The AGEMOS parameters added into fresh modelcards, on the other hand, are special parameters needed by RelXpert to calculate transistor aging using AGEMOS method.

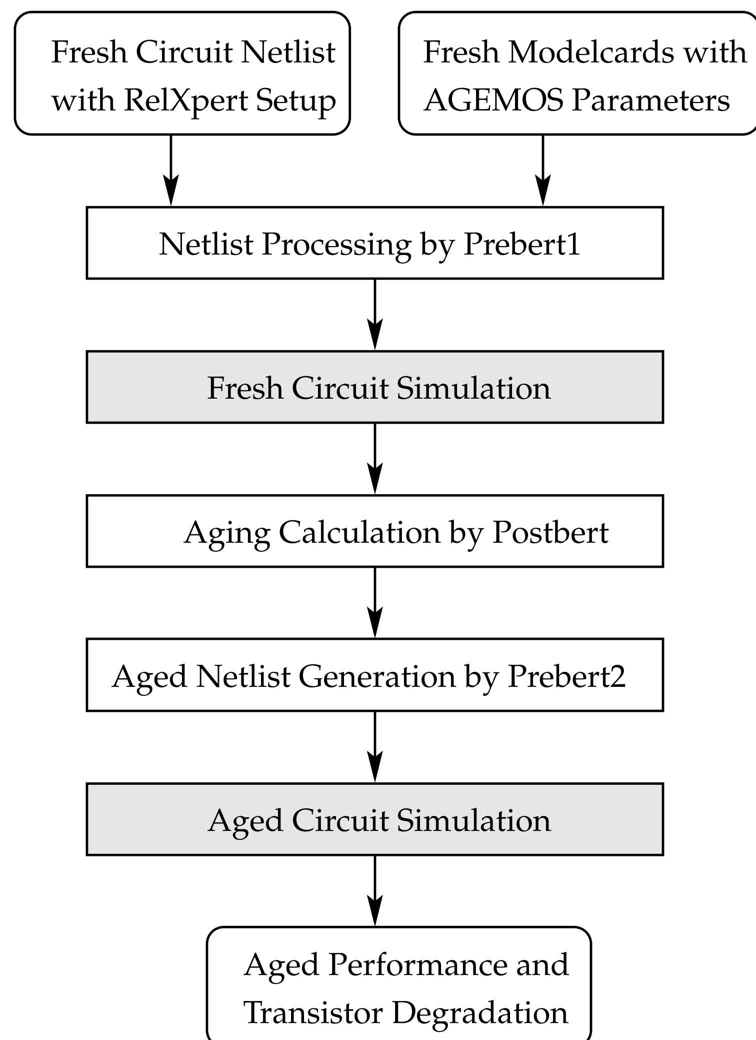


Figure 4.7: The generation and simulation flow of the aged circuit. The two grey boxes indicate the steps where fresh and aged sizing rules are checked respectively.

AGEMOS method is one of the methods used to generate the degraded modelcard parameters. In contrast to other methods generating the "aged" model files by interpolation or regression based on the existing aged model files at several intervals in time domain, the AGEMOS method calculates the degraded modelcard parameters, such as `vth0`, using AGEMOS model parameters directly from their fresh values.

Generally, for a modelcard parameter P, the generation of degraded modelcard parameters is based on such a function [cad]:

$$\Delta P = f(P0, sign, age, d1, d2, n1, n2, s) \tag{4.62}$$

where ΔP is the change value of the modelcard parameter P, P0 is its fresh value, sign is the direction of this parameter changing, age is the degradation age value; and d1, d2, n1, n2, and s are AGEMOS parameters. These AGEMOS parameters are provided by the fabrication foundry.

In comparison to the "aged" model files generation based on several existing interval values, the AGEMOS method is more accurate and efficient. There is no need to add any additional aged modelcard inside circuit netlist.

With the above two inputs, the tool `prebert1` first processes the fresh circuit netlist, by adding statements for measuring the transistor node voltages and currents. `prebert1` stores in addition the device elements information in a temporary file to be read later.

After this step, a simulator such as Spectre, simulates the fresh circuit with those additional measurements, obtaining transistor node voltage and currents which are needed for aging calculation. At this step, the fresh sizing rules of the circuit can be checked. The results can be obtained at no extra cost since only a DC simulation is needed.

Then, the tool `postbert` calculates the aging for each transistor. It reads the device information, model parameters, and simulation options from the temporary file stored by `prebert1`. The aging calculation also needs the transistor node voltages and currents information from the fresh circuit simulation output. At this step, `postbert` generates the aged modelcard for each transistor based on their respective node voltages and currents, i.e., the stress levels. From the output of `postbert`, the ranking of transistors based on their degree of degradation is generated and ready for designers to analyze the critical part of the circuit. It also produces the lifetime T_{life} of each transistor, if the acceptable percentage of degradation is specified as input in the fresh circuit netlist.

To obtain the aged circuit behavior, another step of `prebert2` is performed. `prebert2` generates the aged circuit netlist, by replacing the fresh modelcard by individually aged modelcard for each transistor. For example, for the same type of PMOS transistors that share the same modelcard in the fresh circuit netlist, they now have individual aged modelcards based on their aging behavior calculated by `postbert`.

Then, aged circuit simulation can be performed on the produced aged circuit netlist. The circuit simulator itself does not see any difference between the fresh and the aged circuit netlist, since each ageable modelcard parameter has been obtained already and replaced by `postbert` in the last step. At this step, the aged sizing rules of the circuit can be checked, because it is an aged circuit simulation. Again, these results can be obtained at no extra cost since only a DC simulation is needed.

Finally as output of the flow, the aged circuit performance values are obtained, as indicated in the round-corner box at the bottom of Figure 4.7.

4.3.2 Fresh and Aged Sizing Rules of a Circuit

As discussed in Section 3.5, sizing rules are either geometrical or electrical constraints checked during sizing process, used to ensure the function and robustness of analog integrated circuits.

Some of the electrical constraints are sensitive over aging. For example, to ensure a PMOS transistor working in saturation, v_{ds} must be greater than $v_{gs} - V_{th}$ (in terms of their absolute values). A drift of V_{th} over time thus may violate such constraints. In a simple method as shown in [PG09], some of the sizing rules for the fresh circuit will not be fulfilled after the step of aged yield optimization carried out on the aged circuit. Which means, even if the fresh yield has happened to be high after the step of aged yield optimization, the resulting circuit is very sensitive to the process variations at fresh time.

Considering such sizing rules for both fresh and aged circuits, we apply the fresh and aged sizing rules checking during fresh circuit optimization process, which will ensure the function and robustness of both fresh and aged circuits.

4.3.3 Circuit Layout Area Estimation

The price we pay for a more robust circuit is the additional chip area. The dependency of the process variations on the channel size of a transistor in circuit layout is known as Pelgrom's model, as introduced in [PDW89]:

$$\sigma^2(\Delta P) = \frac{A_P^2}{WL} + S_P^2 D_x^2 \tag{4.63}$$

where P can be certain transistor parameters such as V_{th}, and A_P is the area proportionality constant for parameter P. S_P describes the variation of parameter P with the spacing between devices, denoted by D_x.

Pelgrom's model (4.63) states that the variation of transistor parameters is inversely related with the transistor channel size ($W \times L$), and decreases with the increasing device spacing. So from the design point of view, to minimize the influence of process variations, the designers should design their transistors as "big" as possible, and lay them out as far as possible between each other. This fact is of course limited by the maximal available on-chip area. Thus the investigation of the area penalty over the achieved robustness is of great interest here.

In this thesis, the area of a transistor is approximated by the product of the channel width and the channel length. The influence of the device spacing on parameter variations is ignored, i.e., the second term in (4.63). Since the purpose here is to compare the different design robustness and its additional area penalty, the difference in layout style can be ignored, especially for the transistors which do not change their size too much. In the thesis, for the compensating capacitor in the netlist, it is transformed into the corresponding area by a constant. In this way, the difference between different sizing results can be compared in a unified manner.

4.3.4 Optimization of Fresh Circuit with Fresh and Aged Sizing Rules Checking and Maximum Area Constraints

For a deeper look into the core part of the flow, at the beginning, a fresh netlist with initial design parameters serves as input of the flow. Then, the step of fresh circuit optimization with constraints checking is performed. It can be formulated as

$$\max_{\mathbf{d}} \left\{ \alpha_i(\mathbf{d}) \cdot \beta_{i,w}(\mathbf{d}) \right\}, i = 1, \ldots, n_{\mathbf{f}} \text{ s.t. } \begin{cases} c(\mathbf{d}) \geq 0 \\ area \leq A_{max} \\ c(\mathbf{s})|_{t=t_0} \geq 0 \\ c(\mathbf{s})|_{t=t_1} \geq 0 \end{cases} \tag{4.64}$$

where

$$\alpha_i(\mathbf{d}) = \begin{cases} +1, & \mathbf{s}_0(t_0) \in \mathcal{A}_s \\ -1, & \mathbf{s}_0(t_0) \notin \mathcal{A}_s \end{cases} \tag{4.65}$$

All of the constraints concerning design parameters such as transistor width or length are included in $c(\mathbf{d}) \geq 0$. $area \leq A_{max}$ is the maximum area constraint. Fresh (at t_0) and aged (at t_1) sizing rules relating with ageable statistical parameters are included in $c(\mathbf{s})|_{t=t_0} \geq 0$ and $c(\mathbf{s})|_{t=t_1} \geq 0$ respectively.

The sign $\alpha_i(\mathbf{d})$ for the i'th specification indicates whether the mean value $\mathbf{s}_0(t_0)$ satisfies the specification. If not, with the negative $\alpha_i(\mathbf{d})$, the distance to the performance boundary will be decreased until it reaches zero (the mean value now is on

the boundary), then with the positive $\alpha_i(\mathbf{d})$, the distance is maximized again to increase the robustness.

After the fresh circuit optimization step, the optimal design parameters for the fresh circuit can be obtained. The total circuit area after layout can be approximated which indicates the price the designer has to afford for the achieved robustness.

The maximization problem in (4.64) is a multiple-objective optimization problem. Since each $\alpha_i(\mathbf{d}) \cdot \beta_{i,w}(\mathbf{d})$ as a function of $\mathbf{d}$ represents one partition of the overall yield, (4.64) will increase $\alpha_i(\mathbf{d}) \cdot \beta_{i,w}(\mathbf{d})$ as much as possible.

To solve (4.64), the gradient of the partial worst-case distance $\beta_{i,w}$ over design parameter vector $\mathbf{d}$ is needed. The development of the gradients is presented below. The performance index i is left out for simplicity in the following development process.

To consider the effect of a fluctuation of design parameter vector $\mathbf{d}$ on the performance, a first-order sensitivity term around an arbitrary design point $\mathbf{d}_\epsilon$,

$$\nabla f(\mathbf{d}_\epsilon)^T \cdot (\mathbf{d} - \mathbf{d}_\epsilon) \tag{4.66}$$

is added to the linear performance model (4.45), and the performance boundaries $f_{L/U}$ is extended into

$$f_{L/U} - \nabla f(\mathbf{d}_\epsilon)^T \cdot (\mathbf{d} - \mathbf{d}_\epsilon) \tag{4.67}$$

Applying (4.67) into (4.47) and (4.49) gives [Gra07]:

For the cases $f \leq f_U$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0 \in \mathcal{A}_{s,L}$:

$$\beta_{w,L/U} = \frac{\nabla f(\mathbf{s}_{w,L/U})^T \cdot (\mathbf{s}_0 - \mathbf{s}_{w,L/U}) + \nabla f(\mathbf{d}_\epsilon)^T \cdot (\mathbf{d} - \mathbf{d}_\epsilon)}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.68}$$

For the cases $f \leq f_U$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,L}$:

$$\beta_{w,L/U} = \frac{\nabla f(\mathbf{s}_{w,L/U})^T \cdot (\mathbf{s}_{w,L/U} - \mathbf{s}_0) - \nabla f(\mathbf{d}_\epsilon)^T \cdot (\mathbf{d} - \mathbf{d}_\epsilon)}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.69}$$

The gradients of the worst-case distance over design parameter vector $\mathbf{d}$, as well as statistical parameter vector $\mathbf{s}$ can be deduced from (4.68) and (4.69) as [Gra07]:

For the cases $f \leq f_U$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0 \in \mathcal{A}_{s,L}$:

$$\nabla \beta_{w,L/U}(\mathbf{d}) = \frac{\nabla f(\mathbf{d}_\epsilon)}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.70}$$

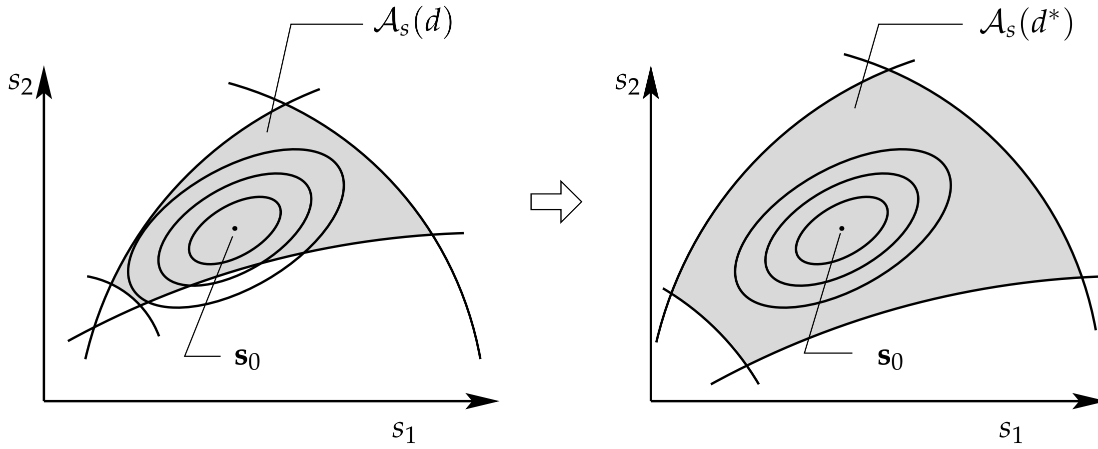


Figure 4.8: Illustration of yield optimization through the change of design parameters.

$$\nabla \beta_{w,L/U}(\mathbf{s}_0) = \frac{\nabla f(\mathbf{s}_{w,L/U})}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.71}$$

For the cases $f \leq f_U$ and $\mathbf{s}_0 \in \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0 \notin \mathcal{A}_{s,L}$:

$$\nabla \beta_{w,L/U}(\mathbf{d}) = -\frac{\nabla f(\mathbf{d}_\epsilon)}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.72}$$

$$\nabla \beta_{w,L/U}(\mathbf{s}_0) = -\frac{\nabla f(\mathbf{s}_{w,L/U})}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})}} \tag{4.73}$$

The sensitivity expressions of the worst-case distance over design parameter vectors in (4.70) and (4.72) explain how a change in the design parameter vector can alter the value of the worst-case distance. The changes in the design parameter contribute to the shift of the performance boundary values according to (4.67). Shifting the performance boundary further away from the nominal statistical parameter vector $\mathbf{s}_0$ can increase the worst-case distance according to (4.70) and (4.72), thus increasing the yield for the performance specification. The shape of the acceptance region is thus changed, such that the volume of the statistical parameter distributions inside the acceptance region is maximized. The effect of the change of design parameters on the yield optimization is illustrated in Figure 4.8.

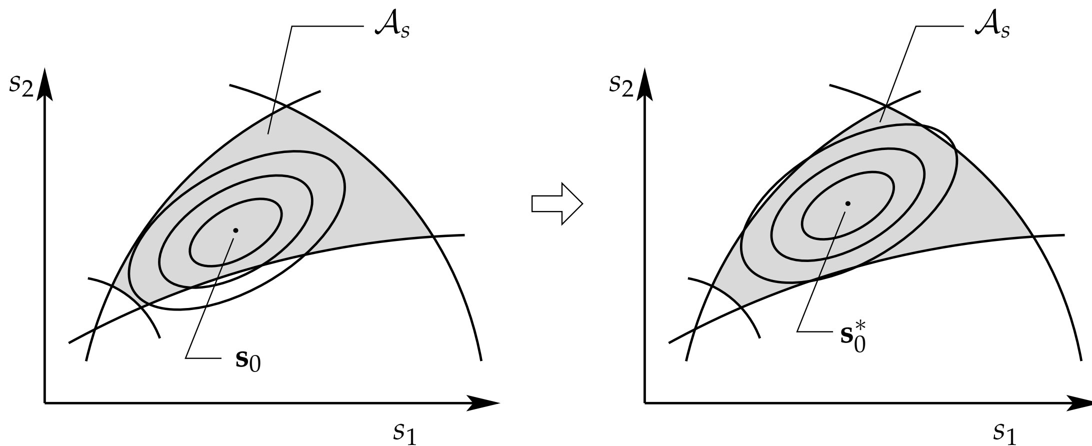


Figure 4.9: Illustration of yield optimization through the change of the nominal vector of statistical parameters.

The increase of the worst-case distance by tuning of the statistical parameters can be explained similarly. The sensitivity expressions of the worst-case distance over statistical parameter vectors in (4.71) and (4.73) indicate the directions, along which the steepest increase of the worst-case distance can be achieved by shifting the nominal statistical parameter vector $\mathbf{s}_0$. In this case, the shape of the acceptance region remains unchanged. The volume of the statistical parameter distributions inside the acceptance region is maximized by shifting the whole distribution around a new nominal vector. The effect of the nominal vector shifts on the yield optimization is illustrated in Figure 4.9.

It is worth mentioning that, the yield optimization through the tuning of the nominal statistical parameter vectors is a part of the manufacturing process design. For the process engineers, both the nominal values and standard deviations of certain transistor parameter distribution are monitored and tuned, in order to maximize the production yield. This is especially important when a new technology is applied. For circuit designers, however, this is beyond their control. They can only change the design parameters. For circuit designers working with discrete devices, they can choose from different categories with different nominal values and variations for the device parameters. This however is not within the scope of this thesis.

As indicated in Figure 4.8, when transistor aging is considered on top of the process variations of statistical parameters, the optimal design parameter vector $\mathbf{d}^*$ should shift the performance boundaries further away for those aging-sensitive performance. In other words, a more robust fresh circuit is needed to tolerate the transistor aging effects.

4.3.5 Aged Yield Analysis

Then, using AgeMOS model for the aging simulator RelXpert from Cadence [cad], the aged yield value of the obtained circuit at age t_1, i.e. the aged robustness of the circuit, can be analyzed. This step can be done by Monte-Carlo simulations on the aged circuit at t_1, or a geometric aged yield analysis based on the evaluation of aged worst-case distances at t_1, which can be formulated as

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s})|_{t=t_1}, i = 1, \ldots, n_{\mathbf{f}} \ \text{ s.t.} \max_{\boldsymbol{\theta}\in\Theta} f_i(\mathbf{s}, \boldsymbol{\theta})|_{t=t_1} \geq f_{i,U}. \tag{4.74}$$

for one of the performance f_i at t_1, whose mean value satisfies the specification $f_i \leq f_{i,U}$.

The inner optimization considers performance values inside the statistical parameter non-acceptance region of that performance f_i at t_1. Concerning the definition of $\mathcal{A}_s$ in (3.24), the non-acceptance region of statistical parameters is defined to be the set of statistical parameters for which there exists an operating parameter vector that violates the performance boundary at t_1:

$$\overline{\mathcal{A}}_{s,U,i} = \left\{ \mathbf{s} | \underset{\boldsymbol{\theta}\in\Theta}{\exists} f_i(\mathbf{s}, \boldsymbol{\theta})|_{t=t_1} \geq f_{i,U} \right\} \tag{4.75}$$

For a performance specification $f_i \leq f_{i,U}$, the definition in (4.75) is equivalent in such a case in the performance space, that the largest performance value obtained over all operating parameters is still bigger than the performance upper bound $f_{i,U}$, i.e.:

$$\mathbf{s}(t_1) \in \overline{\mathcal{A}}_{s,U,i} \Leftrightarrow \max_{\boldsymbol{\theta}\in\Theta} f_i(\mathbf{s}, \boldsymbol{\theta})|_{t=t_1} \geq f_{i,U} \tag{4.76}$$

The other cases corresponding to the specification $f_i \geq f_{i,L}$ and whether or not the mean value satisfies the specification can be formulated accordingly as:

$f_i(t_1) \leq f_{i,U}$ and $\mathbf{s}_0(t_1) \notin \mathcal{A}_{s,U}$:

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s})|_{t=t_1}, \ \text{ s.t.} \max_{\boldsymbol{\theta}\in\Theta} f_i(\mathbf{s}, \boldsymbol{\theta})|_{t=t_1} \leq f_{i,U} \tag{4.77}$$

$f_i(t_1) \geq f_{i,L}$ and $\mathbf{s}_0(t_1) \in \mathcal{A}_{s,L}$:

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s})|_{t=t_1}, \ \text{ s.t.} \min_{\boldsymbol{\theta}\in\Theta} f_i(\mathbf{s}, \boldsymbol{\theta})|_{t=t_1} \leq f_{i,L} \tag{4.78}$$

$f_i(t_1) \geq f_{i,L}$ and $\mathbf{s}_0(t_1) \notin \mathcal{A}_{s,L}$:

$$\min_{\mathbf{s}} \beta_i^2(\mathbf{s})|_{t=t_1}, \quad \text{s.t.} \min_{\boldsymbol{\theta}\in\Theta} f_i(\mathbf{s},\boldsymbol{\theta})|_{t=t_1} \geq f_{i,L} \tag{4.79}$$

The solution of the aged worst-case distance at t_1 can be derived similarly to (4.29)-(4.50) and the formulations in [Gra07] as follows.

For the cases $f \leq f_U$ and $\mathbf{s}_0(t_1) \notin \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0(t_1) \in \mathcal{A}_{s,L}$:

$$\beta_{w,L/U}(t_1) = \frac{\overline{f}_w(\mathbf{s}_0)|_{t_1} - f_{L/U}}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T|_{t_1} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})|_{t_1}}} \tag{4.80}$$

$$= \frac{\nabla f(\mathbf{s}_{w,L/U})^T|_{t_1} \cdot (\mathbf{s}_0(t_1) - \mathbf{s}_{w,L/U}(t_1))}{\sigma_{\overline{f}_w}} \tag{4.81}$$

For the cases $f \leq f_U$ and $\mathbf{s}_0(t_1) \in \mathcal{A}_{s,U}$, as well as $f \geq f_L$ and $\mathbf{s}_0(t_1) \notin \mathcal{A}_{s,L}$:

$$\beta_{w,L/U}(t_1) = \frac{f_{L/U} - \overline{f}_w(\mathbf{s}_0)|_{t_1}}{\sqrt{\nabla f(\mathbf{s}_{w,L/U})^T|_{t_1} \cdot \mathbf{C} \cdot \nabla f(\mathbf{s}_{w,L/U})|_{t_1}}} \tag{4.82}$$

$$= \frac{\nabla f(\mathbf{s}_{w,L/U})^T|_{t_1} \cdot (\mathbf{s}_{w,L/U}(t_1) - \mathbf{s}_0(t_1))}{\sigma_{\overline{f}_w}} \tag{4.83}$$

4.4 Summary

The worst-case distance of a circuit has been proposed to model the circuit robustness and the resulting yield value, considering the manufacturing process variations and various operating conditions. This chapter further extends the formulations and applications into the circuit reliability modeling and optimization considering transistor aging effects. The proposed design flow optimizes the fresh circuit with the consideration of both fresh and aged sizing rules, as well as maximum layout area constraints, in order to achieve x-sigma robustness. Then the robustness of the aged circuit is analyzed by the evaluation of aged worst-case distance values. By applying the design flow repeatedly with different maximum area constraints, the trade-off between circuit's aged robustness and the price paid in terms of circuit layout area can be obtained. Circuit designers can choose from different product reliability categories with acceptable area overhead.

Chapter 5

Aged Yield Prediction

Considering the joint effects of manufacturing process variations and time-dependent transistor aging, the aged yield value of the circuit, $Y(t)$, can be obtained in several ways. One method is to run Monte-Carlo simulations on the aged circuit, where a large number of circuit aging analysis and simulations are needed. This method lacks efficiency, although it can provide the most accurate results. The other method is based on the evaluation of the aged worst-case distance value, $\beta_w(t)$, by solving (4.74).

A different approach to evaluate the aged yield value is presented in this chapter. It is based on the sensitivity analysis of the aged worst-case distance values. The idea is inspired by the performance sensitivity over transistor aging, which is introduced in the following. The challenges of the sensitivity analysis of the aged worst-case distance over transistor aging will be discussed next. Section 5.1 presents the details of the aged worst-case distance prediction model. Section 5.2 shows the structure of the algorithm for the aged yield prediction.

The circuit performance sensitivity over transistor parameter aging can be firstly divided into two components as follows:

- the sensitivity of circuit performance over transistor parameters,
- the degradation of transistor parameters over time.

The first component describes how does a small disturbance of a certain transistor parameter influence the behavior of the circuit, while the second component, the parameter degradations over time, can be decomposed in addition into the following two parts:

- the parameter degradation as a function of stress,

- the actual stress received by each transistor.

This decomposition is based on the fact that the parameter degradation over time is a function of the stress on that transistor, which is again a function of time. Whenever the stress condition of the transistor during operation changes, the transistor parameter degradation changes as well.

The above components contributing to the circuit performance sensitivities over parameter aging can be illustrated in Figure 5.1.

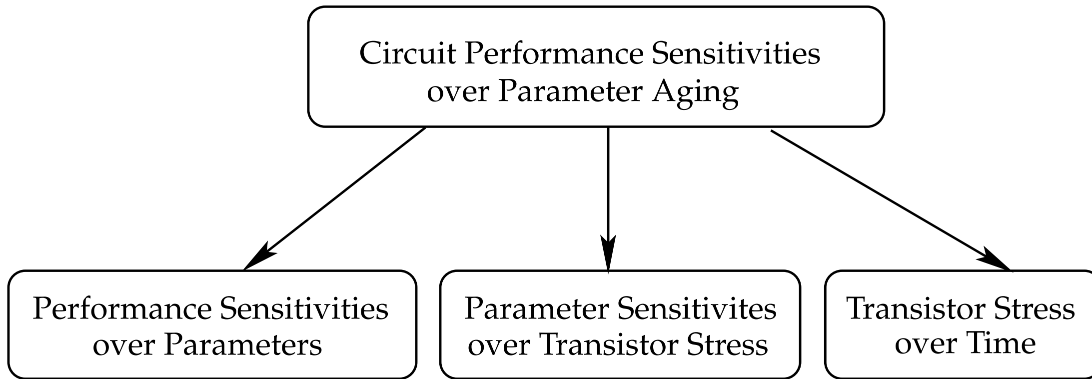


Figure 5.1: Components contributing to the circuit performance sensitivities over transistor parameter aging

Inspired by the performance sensitivity components, the aged worst-case distance and its sensitivities over transistor parameter aging can be used to predict the aged yield value of the circuit, since the aged worst-case distance has been shown in Section 4.2 as a robustness indicator of the aged circuits.

While the idea of performance sensitivities over transistor parameter aging is obvious, the modeling of the aged worst-case distance by means of its sensitivities over parameter aging faces several challenges:

- The aged worst-case parameter vector $\mathbf{s}_w(t)$ at time t is not known a priori. This means it has to be calculated or formulated to enable the formulation of aged worst-case distance.
- The performance specification in the statistical parameter space is not linear in most cases. Only considering the linear sensitivity term may result in inaccuracies.
- The dealing with the variance of statistical parameters needs special attention. A proper modeling and necessary simplifications may be needed.

The purpose of the proposed aged yield prediction model is to provide circuit designers a quick overview of the quality of their circuits after transistors degrade for some time, since the fresh worst-case distance is already available for a fresh-optimal design. After such a quick circuit reliability evaluation, degradation-sensitive performances can be identified, certain weakness in the lifetime robustness of their designs can be obtained early and quickly, proper design techniques can be applied to improve it, and the cost and time for the circuit design process towards reliability can be reduced.

5.1 Aged Worst-Case Distance Prediction Model

5.1.1 Idea

In this section, a prediction model of the aged worst-case distance in time domain is presented to speed up the analysis of corresponding the aged yield value. Only performance and statistical parameter sensitivity analysis are needed, in comparison to the Monte-Carlo simulation method and numerical optimization solutions. It is based on the linear performance model as follows. The index i of ith performance in vector $\mathbf{f}$ is left out for simplicity. Without loss of generality, only upper bound f_U is considered hereafter. The case for lower bound f_L can be derived similarly taken the different sign into consideration.

5.1.2 Linear Performance Model at t_1

At any time t_1 during the operating lifetime, the first-order Taylor expansion of performance $f(t_1)$ with respect to $\mathbf{s}(t_1)$ from the worst-case statistical parameter vector $\mathbf{s}_{w,U}(t_1)$ in the statistical parameter space is

$$\overline{f}(t_1) = f(\mathbf{s}_{w,U}(t_1)) + \nabla f(\mathbf{s}_{w,U}(t_1))^T \cdot (\mathbf{s}(t_1) - \mathbf{s}_{w,U}(t_1)) \tag{5.1}$$

By assuming a linear performance model, the sensitivity of performance over statistical parameters keeps constant, i.e.,

$$\nabla f(\mathbf{s}_{w,U}(t_1)) \equiv \mathbf{g} \tag{5.2}$$

is constant over the entire statistical parameter space at any time.

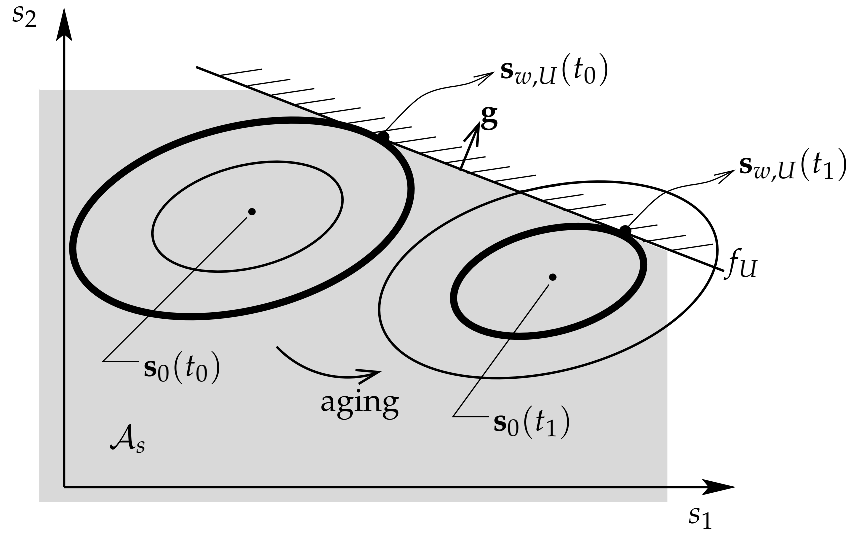


Figure 5.2: Linear performance model in the statistical parameter space.

$f(\mathbf{s}_{w,U}(t_1))$ in (5.1) is the upper bound value f_U. So from (5.1) the linear performance model at t_1 can be formulated as

$$\overline{f}(t_1) = f_U + \mathbf{g}^T \cdot (\mathbf{s}(t_1) - \mathbf{s}_{w,U}(t_1)) \tag{5.3}$$

The worst-case statistical parameter vector $\mathbf{s}_{w,U}(t_1)$ is the statistical parameter vector where the corresponding performance f reaches its boundary value f_U at t_1. It corresponds to the position in the statistical parameter space where the probability of occurrence reaches its maximum in the non-acceptance region (slashed area in Figure 5.2). A robust design indicates that such a probability of occurrence should be kept minimum, i.e., $\mathbf{s}_{w,U}(t_1)$ should be positioned furthest away from $\mathbf{s}_0(t_1)$ so that it is least sensitive to the changes of statistical parameters which may cause it fall into the non-acceptance region.

Since $\mathbf{s}(t_1) \sim \mathcal{N}(\mathbf{s}_0(t_1), \mathbf{C}(t_1))$, the mean and the variance of the linearized performance model can be formulated from (5.3) as

$$\mu(f(t_1)) = f_U + \mathbf{g}^T \cdot (\mathbf{s}_0(t_1) - \mathbf{s}_{w,U}(t_1)) \tag{5.4}$$

$$\sigma^2_{f(t_1)} = \mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g} \equiv \sigma_f^2 \tag{5.5}$$

where (5.5) is constant over time. Taking the process variation as second order effects on the sensitivity towards degradation, $\mathbf{C}(t_1)$ is assumed to be constant, i.e., $\mathbf{C}(t_1) = \mathbf{C}$ [SRRP09].

5.1.3 Mapping from t_0 to t_1

Considering parameter aging from t_0 to t_1, a first-order Taylor approximation of $\mu(f(t_1))$ with respect to t from t_0 can be expressed as

$$\overline{\mu}(f(t_1)) = \mu(f(t_0)) + \frac{\partial \mu_f}{\partial t}|_{t_0} \cdot (t - t_0) \tag{5.6}$$

From (5.4) we have

$$\mu(f(t_0)) = f_U + \mathbf{g}^T \cdot (\mathbf{s}_0(t_0) - \mathbf{s}_{w,U}(t_0)) \tag{5.7}$$

and

$$\frac{\partial \mu_f}{\partial t}|_{t_0} = \mathbf{g}^T \cdot \left(\frac{\partial \mathbf{s}_0(t)}{\partial t}|_{t_0} - \frac{\partial \mathbf{s}_{w,U}(t)}{\partial t}|_{t_0} \right) \tag{5.8}$$

It can be observed from (5.8) that the product

$$\mathbf{g}^T \cdot \frac{\partial \mathbf{s}_{w,U}(t)}{\partial t}|_{t_0} \tag{5.9}$$

remains zero, since the two vectors $\mathbf{g}$ and $\frac{\partial \mathbf{s}_{w,U}(t)}{\partial t}|_{t_0}$ are orthogonal to each other. This is easy to understand because during the aging of statistical parameters, from t_0 to t_1, the worst-case statistical parameter vector $\mathbf{s}_{w,U}(t)$ moves along the performance boundary f_U, as can be observed in Figure 5.2, while the performance gradient $\mathbf{g}$ always points to the direction that is vertical to the performance boundary in the statistical parameter space. Thus $\mathbf{g}$ is orthogonal to the vector $\frac{\partial \mathbf{s}_{w,U}(t)}{\partial t}|_{t_0}$, which is pointing in parallel to the performance boundary.

So (5.8) becomes

$$\frac{\partial \mu_f}{\partial t}|_{t_0} = \mathbf{g}^T \cdot \frac{\partial \mathbf{s}_0(t)}{\partial t}|_{t_0} \tag{5.10}$$

and (5.6) can be further expressed as

$$\overline{\mu}(f(t_1)) \approx f_U + \mathbf{g}^T \cdot (\mathbf{s}_0(t_0) - \mathbf{s}_{w,U}(t_0)) + \mathbf{g}^T \cdot \frac{\partial \mathbf{s}_0(t)}{\partial t}|_{t_0} \cdot (t_1 - t_0) \tag{5.11}$$

5.1.4 Prediction of $\beta_{w,U}(t_1)$

The prediction of the aged worst-case distance at t_1 requires the mapping of the worst-case distance from t_0 to t_1 based on sensitivity analysis, in which the results

from section 5.1.3 are needed. The key concept here is to properly derive the sensitivity parts concerning the aging of the nominal and worst-case statistical parameter vectors.

To predict $\beta_{w,U}(t_1)$, a first-order Taylor expansion of $\beta_{w,U}(t_1)$ with respect to t from t_0 is

$$\beta_{w,U}(t_1) \approx \beta_{w,U}(t_0) + \frac{d\beta_{w,U}(t)}{dt}|_{t_0} \cdot (t - t_0) \tag{5.12}$$

where the sensitivity part, $\frac{d\beta_{w,U}(t)}{dt}|_{t_0}$ can be derived using results from section 5.1.3 as follows.

Since at the worst-case statistical parameter vector at t_1, $\mathbf{s}_{w,U}(t_1)$, the corresponding level contour of $\mathbf{s}(t_1)$ is

$$\beta^2_{w,U}(t_1) = (\mathbf{s}_{w,U}(t_1) - \mathbf{s}_0(t_1))^T \cdot \mathbf{C}^{-1} \cdot (\mathbf{s}_{w,U}(t_1) - \mathbf{s}_0(t_1)) \tag{5.13}$$

It touches the performance boundary at $\mathbf{s}_{w,U}(t_1)$, which means the orthogonal on (5.13) is parallel to $\mathbf{g}$:

$$\mathbf{C}^{-1} \cdot (\mathbf{s}_{w,U}(t_1) - \mathbf{s}_0(t_1)) = \text{sign}(\lambda) \cdot \lambda \cdot \mathbf{g} \tag{5.14}$$

where

$$\text{sign}(\lambda) = \begin{cases} +1, & \mu(f(t_1)) \leq f_U\,; \\ -1, & \mu(f(t_1)) > f_U\,. \end{cases} \tag{5.15}$$

Inserting (5.14) into (5.13) and assuming the positive sign for λ hereafter, we have

$$\beta^2_{w,U}(t_1) = \lambda^2 \cdot \mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g} \tag{5.16}$$

By taking λ from (5.16) into (5.14) we obtain

$$(\mathbf{s}_{w,U}(t_1) - \mathbf{s}_0(t_1)) = \frac{\beta_{w,U}(t_1)}{\sqrt{\mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g}}} \cdot \mathbf{C} \cdot \mathbf{g} \tag{5.17}$$

Then (5.17) is taken back into (5.4):

$$\mu(f(t_1)) = f_U - \beta_{w,U}(t_1) \cdot \sqrt{\mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g}} \tag{5.18}$$

so that the worst-case distance at t_1 can be expressed as

$$\beta_{w,U}(t_1) = \frac{f_U - \mu(f(t_1))}{\sqrt{\mathbf{g}^T \cdot \mathbf{C} \cdot \mathbf{g}}} \tag{5.19}$$

Then from (5.19) and (5.11), and applying the results from (4.12), the worst-case distance degradation rate in (5.12) can be formulated as

$$\frac{d\beta_{w,U}(t)}{dt}|_{t_0} = -\frac{1}{\sigma_f} \cdot \mathbf{g}^T \cdot \frac{\partial \mathbf{s}_0(t)}{\partial t}|_{t_0} \tag{5.20}$$

which differs from (5) in [SRRP09].

From (5.20) it is clear that the evaluation of the worst-case distance degradation rate for a performance upper bound involves only multiple sensitivity evaluations which can be carried out efficiently. Especially in the context of this thesis, both σ_f and $\mathbf{g}$ remain constant, requiring an one-time evaluation only. The sensitivity of $\mathbf{s}_0(t)$ over t is calculated by the finite-difference approximation.

In the finite-difference approximation, for a function f of x, $f(x)$, the sensitivity of f over x, $f'(x)$, is approximated by

$$f'(x) \approx \frac{f(x + \triangle x) - f(x)}{\triangle x} \tag{5.21}$$

The values of $\mathbf{s}_0(t)$ at various ages are obtained from the aging simulator described in Section 4.3.1, then the corresponding sensitivity and the worst-case distance degradation rate can be evaluated.

Thus, by taking (5.20) back into (5.12), the values of $\beta_{w,U}(t_1)$ at time t_1 can be predicted efficiently without searching for the worst-case statistical parameter vector $\mathbf{s}_{w,U}(t_1)$ through iterative optimization method.

5.1.5 Second Order Sensitivity Term

The above prediction model for the aged worst-case distance can be further extended by building up a quadratic prediction model. It is formulated by second-order Taylor expansion in time domain as:

$$\beta_{w,U}(t_1) \approx \beta_{w,U}(t_0) + \frac{d\beta_{w,U}(t)}{dt}|_{t_0} \cdot (t_1 - t_0) + \frac{1}{2} \cdot \frac{d^2\beta_{w,U}(t)}{dt^2}|_{t_0} \cdot (t_1 - t_0)^2 \tag{5.22}$$

The additional second-order sensitivity term of the worst-case distance over time in (5.22), $\frac{d^2\beta_{w,U}(t)}{dt^2}|_{t_0}$, can be obtained by the formulation of $\mu(f(t_1))$ in (5.4) and $\beta_{w,U}(t_1)$ in (5.19) as

$$\frac{d^2\beta_{w,U}(t)}{dt^2}|_{t_0} = -\frac{1}{\sigma_f} \cdot g^T \cdot \frac{\partial^2 \mathbf{s}_0(t)}{\partial t^2}|_{t_0} \tag{5.23}$$

The second-order sensitivity of statistical parameters over time, $\frac{\partial^2 \mathbf{s}_0(t)}{\partial t^2}|_{t_0}$, captures the fact that the degradation models of most statistical parameters are non-linear in time domain [McP07]. Multiplicated by the performance sensitivity over statistical parameters, (5.23) can better capture the non-linear effects during the prediction of the aged worst-case distance.

5.2 Algorithm for the Aged Yield Prediction

This section presents the structure of the algorithm to predict the aged worst-case distance values and the corresponding aged yield values based on the above proposed prediction models.

Step		
choose one performance f with its specification f_U, let $j = 0$		(I)
calculate $\beta_{w,U}(t_j)$ of f using geometric yield approximation		(II)
calculate $\mathbf{g}$ and σ_f of f using (5.5)		(III)
	increase j by 1	(IV)
	sensitivity analysis for $\partial \mathbf{s}_0(t_j)/\partial t_j$ from finite-difference approximation	(V)
	sensitivity analysis for $\partial^2 \mathbf{s}_0(t_j)/\partial t_j^2$ from finite-difference approximation	(VI)
	calculate $d\beta_{w,U}(t_j)/dt_j$ using (5.20) and $d^2\beta_{w,U}(t_j)/dt_j^2$ using (5.23)	(VII)
	predict $\beta_{w,U}(t_j)$ using (5.22)	(VII)
	predict $Y_U(t_j)$ using (4.61)	(VIII)
until $j = j_{\max}$		(IX)
finish		(X)

Figure 5.3: Overview of the algorithm to predict aged worst-case distances and aged yield for one of the performance at different ages

Firstly, the worst-case distances $\beta_{w,U}(t_0)$ for one performance feather f at t_0 is calculated, using geometric yield approximation method in Chapter 4. In practice, since the fresh yield optimization is a necessary step during the circuit design process, these $\beta_{w,U}(t_0)$ can be obtained from the results of the fresh yield optimization.

Then, the standard deviation of the linear performance model, σ_f, as well as the sensitivity of f over statistical parameters $\mathbf{s}$ are prepared. Again, these values are already available during the fresh yield analysis based on worst-case distance process.

For each time point t_j of interests, the sensitivity analysis of $\mathbf{s}_0(t_j)$ over time is performed based on the finite difference approximation. The results of these first-order sensitivities are then applied into the calculation of second-order sensitivity of $\mathbf{s}_0(t)$ over time. Based on these results, the first- and second-order sensitivity of $\beta_{w,U}(t)$ over time can be obtained, using (5.20) and (5.23) respectively. Finally, the aged worst-case distance $\beta_{w,U}(t_j)$ can be predicted. The corresponding aged yield value $Y_U(t_j)$ can be obtained accordingly.

The degradation of the aged worst-case distance as well as aged yield can then be obtained if the above process is repeated for all of the time points of interests. The whole structure of the prediction algorithm is shown in Figure 5.3.

5.3 Summary

Considering both manufacturing process variations and transistor aging during lifetime operation, this chapter proposes a modeling and prediction framework to predict the aged worst-case distance value and the corresponding lifetime robustness of analog circuits. The proposed method is based on the sensitivity analysis of transistor parameters over aging, as well as the sensitivity analysis of the circuit robustness over transistor parameters. It does not involve either analytical formulation of circuit performance or Monte-Carlo simulations. In comparison with the aged yield analysis based on the geometrical yield modeling presented in the last chapter, the proposed method is more efficient in obtaining the aged worst-case distance values. Using the proposed method, circuit designers can obtain quickly an overview of the lifetime robustness of their designs, since the fresh worst-case distance is already available for a fresh-optimal design. Certain weakness in the lifetime robustness of their designs can be obtained early and quickly, thus reducing the redesign cost.

Chapter 6

Experimental Results

This chapter presents the experimental results on the proposed sizing flow presented in Chapter 4, as well as the results on the new prediction framework presented in Chapter 5. The experiments are performed on different types of operational amplifiers, which serve as basic building blocks in analog circuitry. They are based on an 180nm industrial technology.

The experiments are run on a 64-bit Linux server with eight 2.33GHZ Xeon CPUs and 2G memory. The AgeMOS model parameters are obtained from an industry partner. By using the RelXpert simulator, which has both NBTI and HCI degradation models, the aged BSIM modelcard and the aged netlist for a specific time and temperature can be generated. Other types of aging simulators with respective aging models can be applied as well.

6.1 Miller Operational Amplifier

6.1.1 Circuit Topology

Figure 6.1 shows the schematic of the two-stage Miller operational amplifier used in our experiments [LS94]. The first stage of the circuit is made up of a pair of PMOS transistors with the current mirror MN1 and MN2 as an active load. The second stage is a CMOS inverter with MN3 as driver and MP5 as an active load. The compensating capacitor C_{miller} connects the output of the circuit to the input.

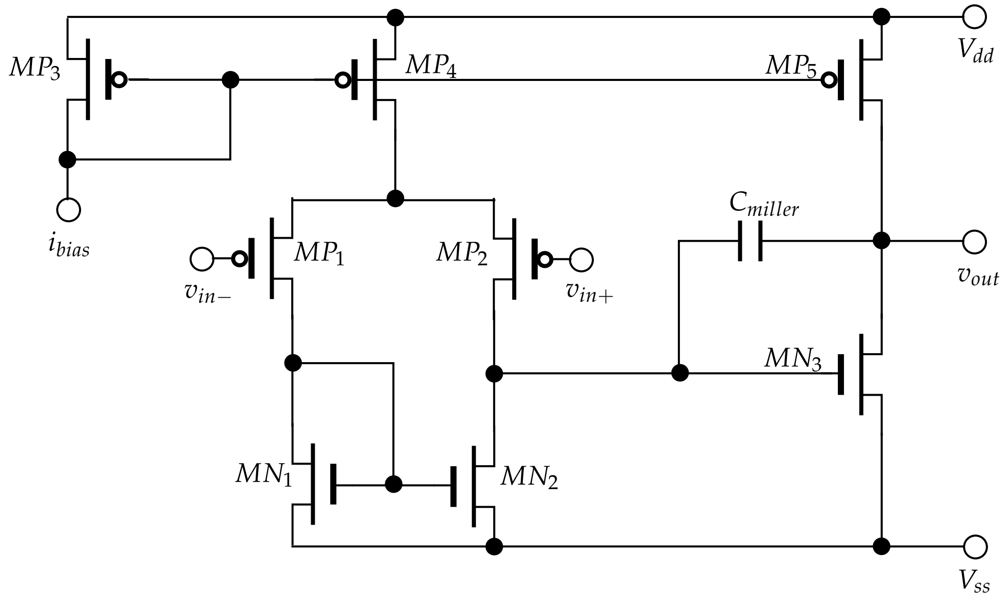


Figure 6.1: Circuit schematic of the Miller operational amplifier

6.1.2 Circuit Performances and their Specifications

In this study, there are seven performance specifications totally for the Miller operational amplifier, as listed in Table 6.1.

Table 6.1: List of performance specifications for the Miller operational amplifier.

Performance	DC Gain (dB)	GBW (MHz)	PM (degree)	SR (v/μs)	CMRR (dB)	Swing (%)
Specification	≥ 80	≥ 2	≥ 60	≥ 3	≥ 80	≥ 90

For the discussion of the experimental results from circuit simulation later, the analytical form of each performance is prepared here first. A complete process of derivations can be referred to [LS94].

- DC Gain
  DC Gain specifies the low frequency gain of the amplifier. It is a major performance of an amplifier which implies the ability of the amplification. For the first stage, its gain A_1 is given by

$$A_1 = \frac{g_{m1}}{g_{o1}} \tag{6.1}$$

where g_{m1} is the transconductance of transistor MP1, $g_{o1} = g_{oMP2} + g_{oMN2}$ is the load conductance of the first stage. For the second stage which is an inverter, its gain A_2 is given by

$$A_2 = \frac{g_{m3}}{g_{o2}} \tag{6.2}$$

where g_{m3} is the transconductance of transistor MN3, $g_{o2} = g_{oMN3} + g_{oMP5}$ is the load conductance of the second stage. Then the DC Gain at low frequencies is given by

$$DC\ \ Gain = A_1 A_2 = \left(\frac{g_{m1}}{g_{o1}}\right)\left(\frac{g_{m3}}{g_{o2}}\right) \tag{6.3}$$

- GBW
  Gain-Bandwidth product (GBW) is the frequency value at which the gain drops to 0dB. As the name implies, it can also be defined as the product of the DC Gain and the bandwidth. Here, the "bandwidth" of the amplifier refers to the dominant pole frequency, at which the gain starts to drop (more exactly, drops by -3dB). As a constant value with various gain and its frequency, GBW describes the amplifier's gain behavior with frequency. For example, if DC Gain is 10000, or $20\log(10000) = 80$dB, and GBW is 2MHz, the bandwidth then is 200Hz (at this frequency the gain is actually (80-3=77)dB). When the gain drops to 1, or 0dB, the corresponding frequency is 2MHz. GBW helps the designers to know the bandwidth value for a certain gain. A larger GBW implies a wider frequency at which the maximal gain can be obtained.
  Analytically, since the dominant pole comes from the result of the Miller effect of C_{miller}, the bandwidth can be approximated by

$$bandwidth \approx \frac{g_{o1}}{2\pi A_2 C_{miller}} \tag{6.4}$$

  GBW can be obtained by the product of (6.4) and DC Gain, given by (6.3), approximately as

$$GBW \approx \frac{g_{o1}}{2\pi A_2 C_{miller}} \cdot \left(\frac{g_{m1}}{g_{o1}}\right) \cdot A_2 \tag{6.5}$$

$$= \frac{g_{o1}}{2\pi C_{miller}} \tag{6.6}$$

- PM
  Phase margin (PM) is an important indicator of the stability of the amplifier. Since the frequency where the gain drops to 0dB is represented as GBW, PM measures the difference between the phase at the frequency GBW and $-180°$, shown as

$$PM = \varphi(f = GBW) - (-180°) \tag{6.7}$$

- SR
  Slew rate (SR) is the slope of the ramp at the output node of the amplifier when a large input step signal is applied. It has both rising and falling parts. In this study, only the rising slew rate is considered. SR measures how fast the output signal follows the input signal. It is measured as

$$SR = \max\left(\frac{\Delta V_{out}}{\Delta T}\right) \tag{6.8}$$

  i.e., the maximum ratio between the change of the output voltage and the time required to achieve such a change.
  In the two-stage Miller amplifier, it is expressed as

$$SR = \frac{I_B}{C_{miller}} \tag{6.9}$$

  where I_B is the current flowing through the Miller capacitor to charge it. It arises from the fact that the slewing phenomena occurs when a constant output current of the first stage charges the compensating capacitor of the second stage. The low pass character of the second stage behaves similar to an integrator, where an increasing output is produced when a constant input is applied.
- CMRR
  Common mode rejection ratio (CMRR) is the ratio between amplifier's differential gain and common mode gain. Ideally, the common mode gain should be zero, since the goal of a differential amplifier with a differential pair as the input stage is to amplify only the differential component of the input signal, while reject totally the common-mode component. The resulting CMRR should be infinite in the ideal case. But in reality, due to the finite output impedance of the current source of the differential input stage, as well as the asymmetries in the input pair, it is a finite value. The computation of CMRR, in dB, is given by

$$CMRR = 20 \cdot \log|\frac{A_{diff}}{A_{comm}}| \tag{6.10}$$

$$= 20 \cdot \log[\frac{g_{m1}}{g_{o1}} \cdot 2g_{mn1}R_o] \tag{6.11}$$

  where g_{m1} is the transconductance of MP1, $g_{o1} = g_{oMP2} + g_{oMN2}$ is the load conductance of the first stage, g_{mn1} is the transconductance of MN1, R_o is the output resistance of the current source MN1 and MN2. CMRR can be improved by increasing the input transconductance g_{m1}, and by taking a current source with high output resistance R_o.

- Swing
  Swing at the output node is the maximum voltage range it can achieve. It is measured as a relative percent between maximum output voltage and the supply voltage. In reality, the maximum output voltage is limited by two factors: V_{ds} of MP5 which keeps it in saturation, and the output current driving the load. As mentioned in [Raz01], the maximum voltage swing trades with device size and bias currents and hence speed. It is the principle challenge to obtain a large swing in modern analog design.

6.1.3 Results on Aged Yield Optimization

Table 6.2 shows the result using the proposed fresh circuit optimization and lifetime yield analysis flow. They are obtained based on circuit simulations. For the five performances considered, the corresponding worst-case distance values as well as the total yield values for the fresh circuit and the 10-year-old circuit are listed. It is clear from the table that the lifetime robustness of the circuit is ensured, with an aged yield of 96.9% even after 10 years.

Table 6.2: Simulation results of worst-case distance after applying the proposed optimization flow for the fresh and aged Miller operational amplifier.

	t = 0	t = 10 years
DC Gain	4.790	4.470
GBW	5.901	5.720
PM	6.382	6.155
SR	5.441	2.030
CMRR	4.423	2.896
Yield	100%	96.9%

It can be observed from Table 6.2 that the SR and CMRR are critical performances considering aging after 10 years. They are justified as follows.

To measure SR, a positive voltage step is applied to v_{in+}, which turns off the PMOS transistor MP2. The current from the current source MP4 then flows through MP1 and MN1, as well as through MN2 as a copy. Since MP2 is off, this current will be drawn through the capacitor C_{miller}. The SR of the amplifier is then determined by the speed of charging/discharging of this C_{miller}. When transistor aging occurs,

the current through MP4 degrades over time due to the change of its v_{th}, as already demonstrated in [JRSR05]. Thus the speed of charging/discharging of the capacitor C_{miller} decreases over time.

CMRR on the other hand measures the ratio of differential gain to common mode gain. Ideally it should be infinite with common mode voltages being totally rejected, but in practice it is limited due to the finite output impedance of the current source of MN1 and MN2 as well as the asymmetries in the input transistors MP1 and MP2. Analytically it depends on the transconductance gm of MN1 and MP1. During transistor aging, both of these quantities degrade over time, resulting in a significant loss of CMRR over time.

The overall additional area of the above optimized circuit is 50%. A detailed list of the additional area of each device for the above optimized circuit is shown in Figure 6.2, where MN1 and MN2, as well as MP1 and MP2 have the same size. As can be seen, after fresh circuit optimization, the gate area of MP3 has a maximum increase up to 75.9%, increasing the robustness of the current driving capability over aging. This is achieved together with the additional gate area of MP4 and MP5. The current source MN1 and MN2 also see a gate area enhancement for the similar reason. The gate area increase of input pair MP1 and MP2 improves the matching property.

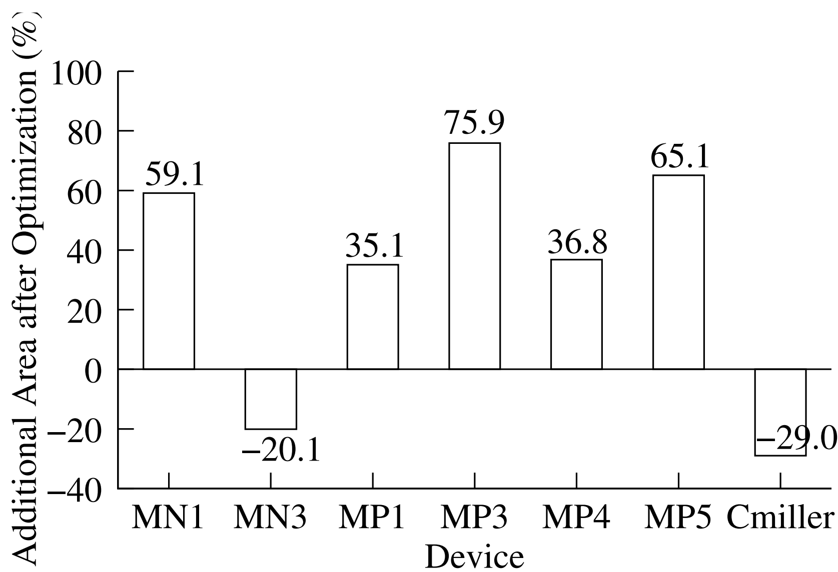


Figure 6.2: Additional area of each device after fresh circuit optimization on the Miller operational amplifier (MN1 and MN2, MP1 and MP2 have the same size).

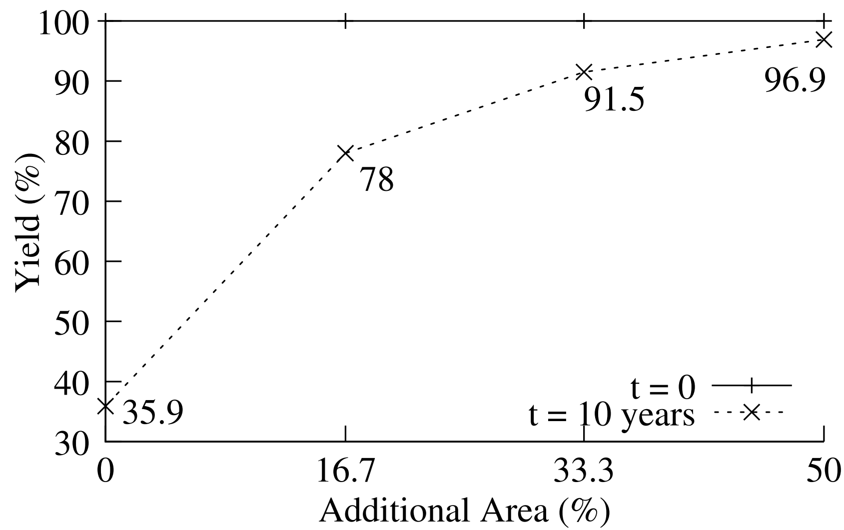


Figure 6.3: Yield at lifetimes t=0 and t=10 years for four designs with increasing additional area requirements on the Miller operational amplifier.

Furthermore, we have performed four different optimizations according to Figure 4.6 with increasing maximum area value A_{max} as a constraint, and compare the 10-year robustness with corresponding areas.

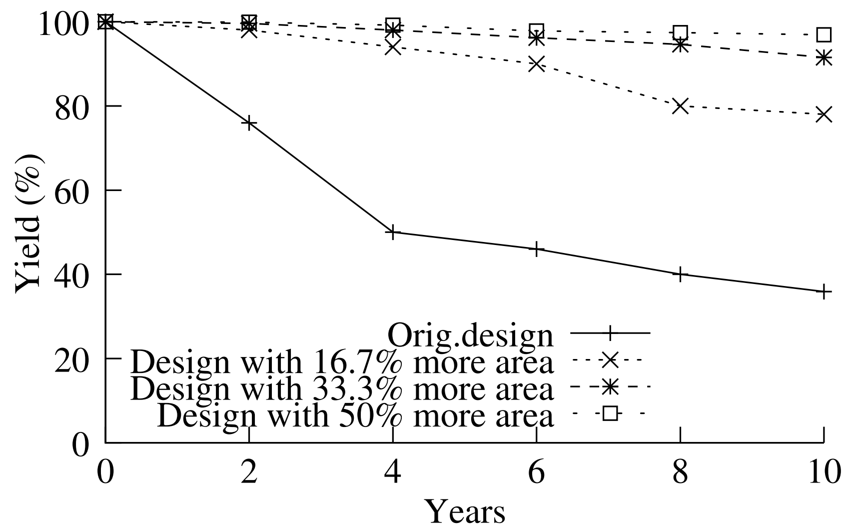


Figure 6.4: Aging of yield value over time for four different Miller operational amplifier designs with different layout areas. The respective lifetime ends if the aged yield drops to a certain boundary.

Figure 6.3 shows the results for the fresh and 10-year-old circuit. As can be seen, a bigger aged yield value always requires more circuit area. In comparison to an initial design which achieves an optimized 10 years yield value of 35.9%, it needs 16.7% more area for an optimized aged yield value of 78%, or 33.3% more area for an optimized aged yield value of 91.5%, or 50% more area for 96.9% aged yield.

Figure 6.4 shows the aging of yield values from their fresh value to various time points for the above four different designs with different layout area. It is clear from the figure that, concerning the aged yield values, the design with 50% more area is the most robust one which ages relatively slowly in comparison to other realizations.

The aged yield value indicates the aged robustness of the circuit. When it drops to a certain pre-defined value, the lifetime of the circuit products according to that definition ends. The above experimental results verify that a longer circuit lifetime requires more total area to be spent in layout. By using the proposed sizing flow with maximum area constraints, designers can ensure the circuit robustness in operational lifetime with a certain layout area consumption.

6.1.4 Results on Aged Yield Prediction

To verify the prediction model presented in Chapter 5 on the Miller operational amplifier, the aged worst-case distance values obtained through a geometric aged yield analysis based on the evaluation of aged worst-case distances by solving (4.74), and through the prediction model (5.22) are compared for the five performances listed in Table 6.2. The comparison results are presented in Table 6.3.

Table 6.3: Aged worst-case distance prediction results in comparison with accurate values for different performance features of the Miller operational amplifier at t=10 years.

	Accurate	Predict	Error	Speedup
DC Gain	4.470	4.602	3.0%	4.6X
GBW	5.720	5.362	6.3%	6.5X
PM	6.155	6.348	3.1%	6.2X
SR	2.030	2.124	4.6%	6.5X
CMRR	2.896	2.630	9.2%	7.4X
Average			5.24%	6.22X

As can be seen from the table, for the five performances considered, the predicted aged worst-case distances match very close to the accurate aged worst-case distance values, with an average error of 5.24%. A clear speedup by using the proposed prediction framework can be observed. On average they are 6.22 times faster in comparison to the solutions obtained through the geometric aged yield analysis.

A detailed comparison on prediction results and relative errors at various time points of interests for the performance SR and CMRR are plotted in Figure 6.5. It is clear from the figure that the results using the proposed prediction framework can well track the aging of the worst-case distances at multiple time points of interests during the lifetime. As a result, circuit designers can have a quick image of their design robustness with relatively small error by applying the proposed prediction framework.

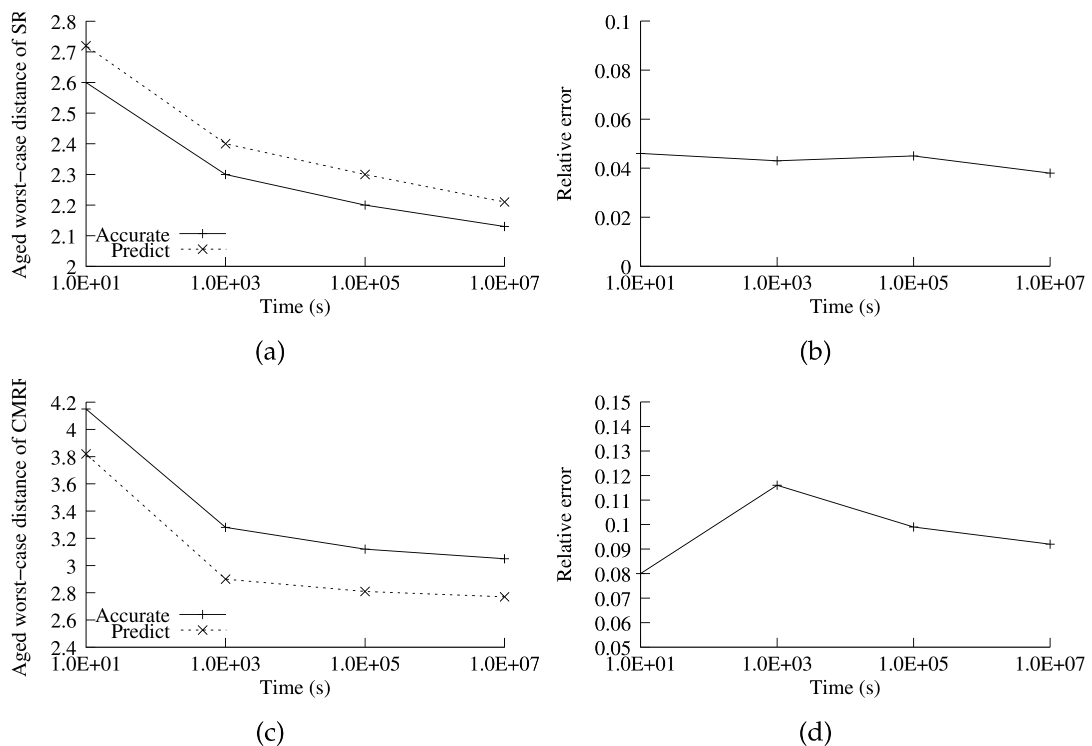


Figure 6.5: Comparison results of $\beta_w(t)$ at respective time points for SR (a) and CMRR (c), and the corresponding relative errors for SR (b) and CMRR (d) of the Miller operational amplifier.

6.2 Folded Cascode Operational Amplifier

Another experimental investigation is performed on the folded cascode operational amplifier, as shown in Figure 6.6. Folded cascode operational amplifier uses wide swing current mirror to achieve higher speed. The performances and their specifications for the folded cascode operational amplifier are the same as those for the Miller operational amplifier, as listed in Table 6.1.

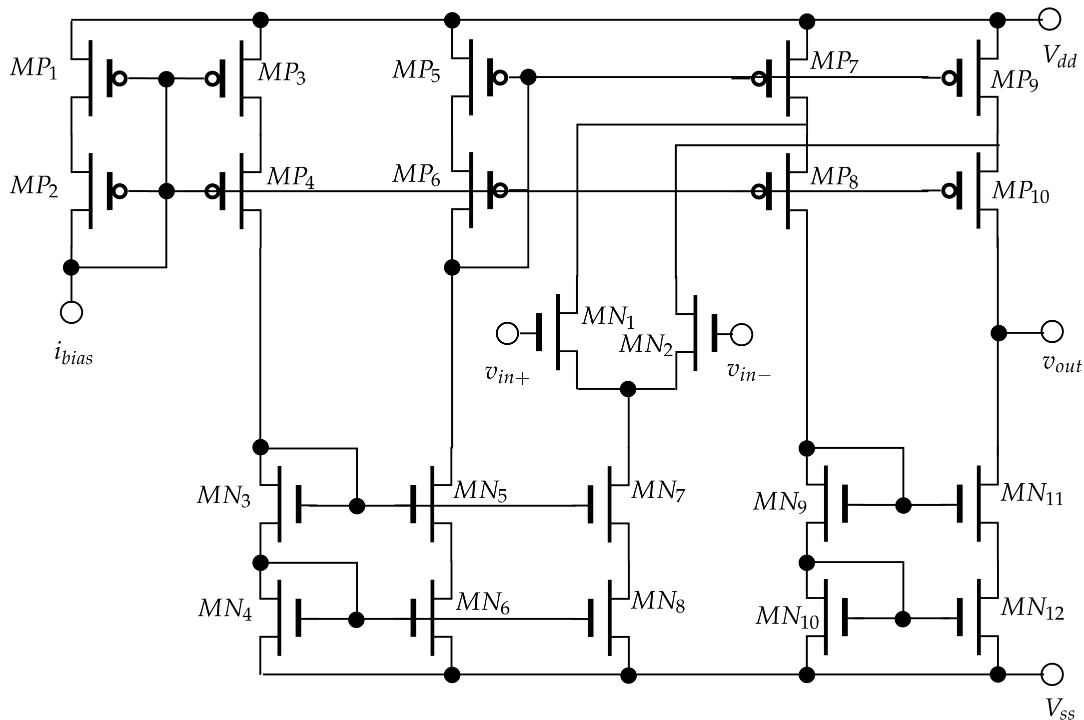


Figure 6.6: Circuit schematic of the folded cascode operational amplifier

6.2.1 Results on Aged Yield Optimization

Using the proposed fresh circuit optimization and lifetime yield analysis flow, the worst-case distance values as well as the total yield values for the fresh circuit and the 10-year-old circuit are listed in Table 6.4. It is clear from the results in Table 6.4 that the lifetime robustness of the folded cascode operational amplifier is achieved, with a 10-year aged yield of 94.6%.

Four different optimizations on the folded cascode operational amplifier are performed according to Figure 4.6 with increasing maximum area value A_{max} as a con-

Table 6.4: Simulation results of worst-case distance after applying the proposed optimization flow for the fresh and aged folded cascode operational amplifier.

	t = 0	t = 10 years
DC Gain	4.761	4.289
GBW	6.372	5.837
PM	7.760	6.219
SR	5.280	2.873
CMRR	5.095	1.914
Yield	100%	94.6%

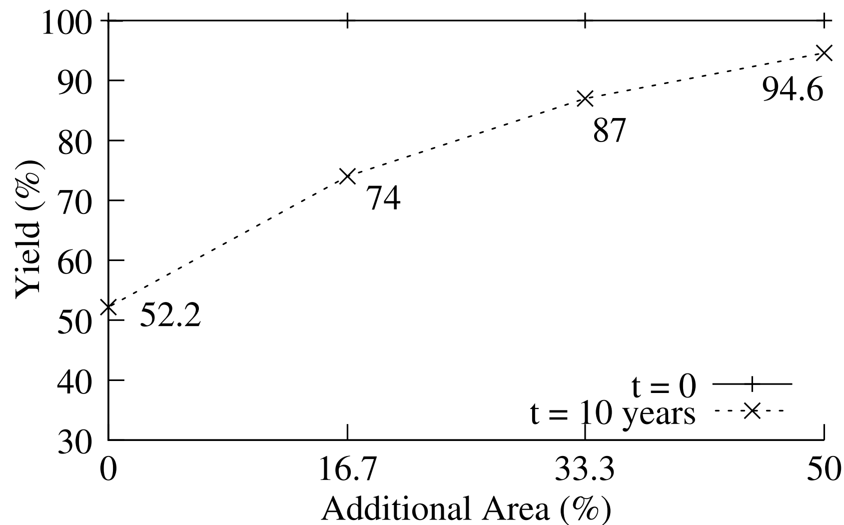


Figure 6.7: Yield at lifetimes t=0 and t=10 years for four designs with increasing additional area requirements on the folded cascode operational amplifier.

straint. The trade-off between the 10-year robustness and the corresponding areas of the four designs are presented in the following.

Figure 6.7 shows the results for the fresh and 10-year-old circuit. As can be seen, a bigger aged yield value always requires more circuit area. In comparison to an initial design which achieves an optimized 10 years yield value of 52.2%, it needs 16.7% more area for an optimized aged yield value of 74%, or 33.3% more area for an optimized aged yield value of 87%, or 50% more area for 94.6% aged yield.

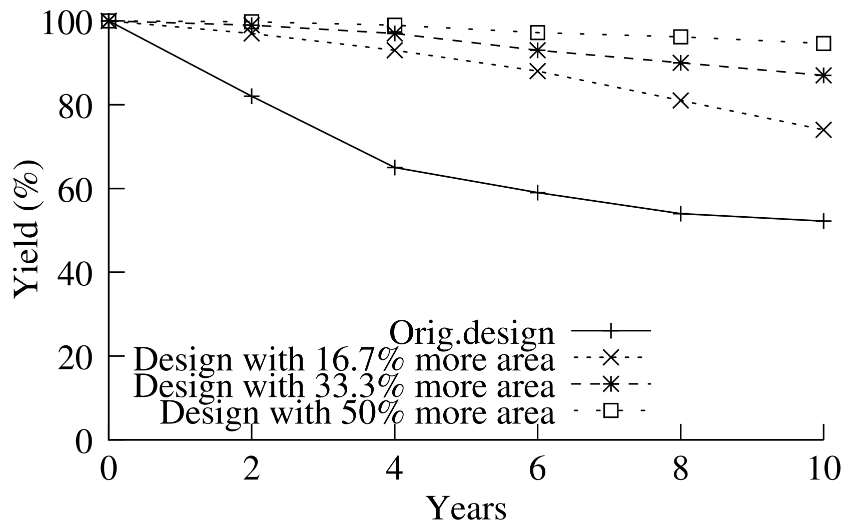


Figure 6.8: Aging of yield value over time for four different folded cascode operational amplifier designs with different layout areas. The respective lifetime ends if the aged yield drops to a certain boundary.

Figure 6.8 shows the aging of yield values from their fresh value to various time points for the above four different designs. The experimental results on the folded cascode operational amplifier also verify the fact that, spending more layout area can ensure a better lifetime robustness of the circuits. By using the proposed sizing flow with maximum area constraints, designers can ensure the circuit robustness in operational lifetime with a certain layout area consumption.

6.2.2 Results on Aged Yield Prediction

To verify the prediction model presented in Chapter 5 on the folded cascode operational amplifier, the aged worst-case distance values obtained through the geometric aged yield analysis based on the evaluation of aged worst-case distances by solving (4.74), and through the prediction model (5.22) are compared for the five performances listed in Table 6.4. The comparison results are presented in Table 6.5.

As can be seen from the table, for the five performances considered on the folded cascode operational amplifier, the predicted aged worst-case distances also match very close to the accurate aged worst-case distance values, with an average error of 6.10%. A clear speedup by using the proposed prediction framework can also be observed in this case, with an average value of 5.98 times in comparison to the solutions obtained through the geometric aged yield analysis.

Table 6.5: Aged worst-case distance prediction results in comparison with accurate values for different performance features of the folded cascode operational amplifier at t=10 years.

	Accurate	Predict	Error	Speedup
DC Gain	4.289	4.590	4.9%	4.2X
GBW	5.837	5.984	2.5%	6.8X
PM	6.219	6.618	6.4%	5.7X
SR	2.873	2.684	6.6%	6.3X
CMRR	1.914	1.720	10.1%	6.9X
Average			6.10%	5.98X

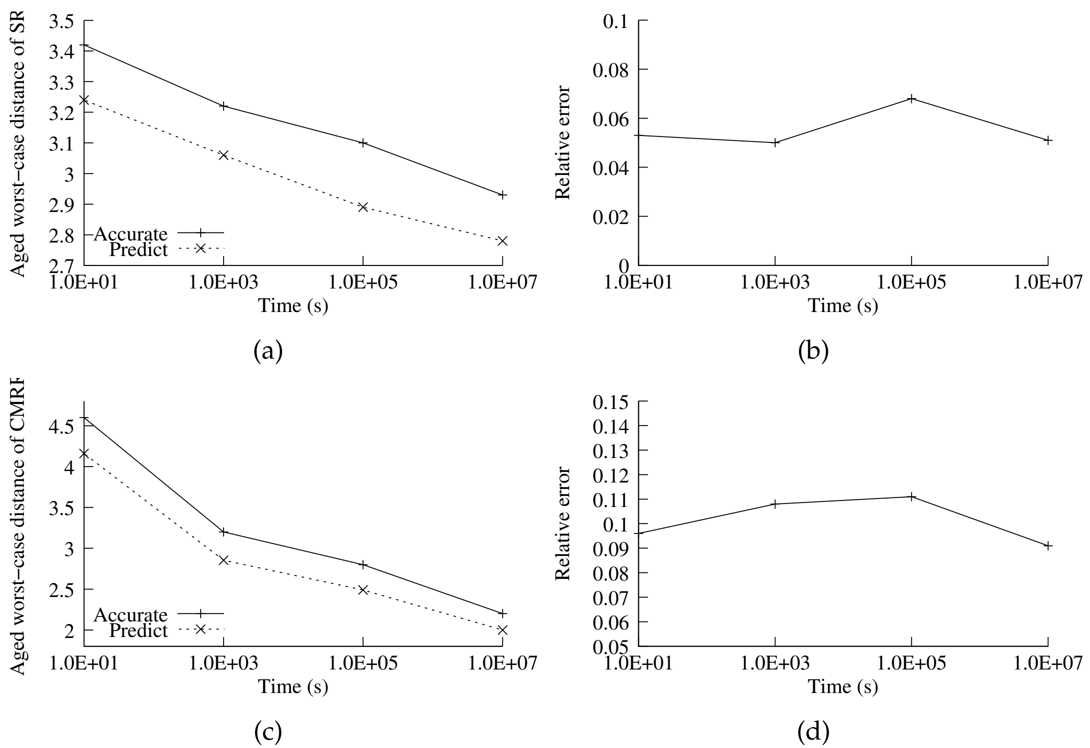


Figure 6.9: Comparison results of $\beta_w(t)$ at respective time points for SR (a) and CMRR (c), and the corresponding relative errors for SR (b) and CMRR (d) of the folded cascode operational amplifier.

A detailed comparison on prediction results and relative errors at various time points of interests for the performance SR and CMRR are plotted in Figure 6.9. It is clear from the figure that for the folded cascode operational amplifier, the results using the proposed prediction framework can also well track the aging of the worst-case distances at multiple time points of interests during the lifetime. Thus the proposed framework to predict the aged worst-case distance serves as a fast overview into the lifetime robustness of the circuits.

6.3 Summary

This chapter presents the experimental results of the proposed fresh circuit optimization and lifetime yield analysis flow, as well as the modeling and prediction framework to predict the aged worst-case distance value and the corresponding lifetime robustness of analog circuits. According to the result of this trade-off analysis, a longer circuit lifetime requires more total area to be spent in layout, and designers can ensure the circuit robustness with certain layout area consumption. The results on the aged worst-case distance prediction models, on the other hand, verify that the models are accurate enough to provide a fast overview of the lifetime robustness of the analog circuits.

Chapter 7

Conclusion

Semiconductor manufacturing process variations and transistor aging during lifetime operations are the two main challenges raised by the continuous scaling of semiconductor technologies. Although the higher chip density with a lower cost per transistor as well as the improved circuit performance are contributed by the advanced technologies, designers must ensure the robustness of their circuit designs early during the design phase to tolerate those above mentioned uncertainties. This task is getting more and more complicated with increased number of transistor parameters and design complexities in the new generations of technologies.

In comparison to the digital counterpart, the design of analog integrated circuits is still mainly done manually. However, there has been a trend in recent years towards an automatic sizing of analog integrated circuits, by the close interaction between analog designers and automatic sizing software tools. These tools can help analog designers in improving their design qualities with a better overview of the design space and a more intelligent recognition of important analog circuit structures, as well as a better solution in terms of sizing. In addition, these automatic sizing tools need to consider the emerging effects caused by the new manufacturing technologies and physical effects in order to improve the quality of the solutions they provide to the designers. Thus there is a strong need from the semiconductor industry for both analog circuit designers and automatic sizing tool developers, to have a better understanding into the emerging reliability challenges and deeper improvement in the whole design flow to cope with the new physical effects.

This thesis proposes new solutions to the joint effects of process variations and transistor aging, with new sizing flow during design phase to ensure a robust circuit design in operational lifetime, as well as a new prediction framework to help designers evaluate their design robustness in operational lifetime.

The new sizing flow presented in Chapter 4 is based on the evaluation and optimization of the worst-case distance for fresh circuits with checking of sizing rules for both fresh and aged circuits. The worst-case distance has been proved to be an effective measurement to the circuit robustness in terms of a number of sigma. Originated for the modeling of statistical manufacturing process variations, worst-case distance also applies in the case where the statistical parameters drift over time due to transistor aging. Thus the evaluation of the fresh and aged worst-case distance provides insights into the circuit robustness for both fresh and aged circuits. Checking of sizing rules for both fresh and aged circuits, on the other hand, ensures the robustness further into lifetime operation, since certain constraints on transistor node voltages are ensured in the sizing flow. Overall, the proposed flow captures the robustness measures based on fresh and aged worst-case distances, optimize the fresh worst-case distances, with consideration of both fresh and aged sizing rules to ensure the lifetime robustness.

The new prediction framework presented in Chapter 5 further speeds up the evaluation of the aged circuit robustness. The prediction of the aged worst-case distance, in this case, is based on an analytical evaluation model, instead of the solution from an iterative numerical optimization algorithm. The sensitivity analysis of the statistical parameters over time is obtained using aging simulators with specified time points of interests. It does not involve either analytical formulation of circuit performances or Monte-Carlo simulations. Circuit designers can capture quickly an overview of the lifetime robustness of their designs, and certain weakness in the lifetime robustness of their designs can be obtained early and quickly, thus reducing the redesign cost.

In conclusion, the presented methods provide new solutions to the emerging joint effects of process variations and transistor aging in the scaling semiconductor manufacturing technologies, with new sizing flow during design phase to ensure a robust circuit design in operational lifetime, as well as a new prediction framework to help designers to predict their design robustness.

Bibliography

[AGW94] K. Antreich, H. Graeb, and C. Wieser. Circuit analysis and optimization driven by worst-case distances. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 13(1):57–71, 1994.

[AHY87] S. Aur, D.E. Hocevar, and P. Yang. HOTRON-A circuit hot electron effect simulator. In *Proceedings of the IEEE International Conference on Computer-Aided Design (ICCAD)*, pages 256–259, 1987.

[AKPR07] M. Alam, K. Kang, BC Paul, and K. Roy. Reliability-and process-variation aware design of VLSI circuits. In *Proceedings of the 14th IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, pages 17–25, Bangalore, India, 2007.

[AKVM07] M. A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra. A comprehensive model for PMOS NBTI degradation: Recent progress. *Microelectronics Reliability*, 47(6):853–862, June 2007.

[AMHH99] Hany L. Abdel-Malek, Abdel-Karim S. O. Hassan, and Mohamed H. Heaba. A boundary gradient search technique and its applications in design centering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 18(11):1654–1660, November 1999.

[AS94] S. A. Aftab and M. A. Styblinski. IC variability minimization using a new C_p and C_{pk} based variability/performance measure. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 149 – 152, London , UK, May 1994.

[AVK08] M.A. Alam, D. Varghese, and B. Kaczer. Theory of breakdown position determination by voltage-and current-ratio methods. *IEEE Transactions on Electron Devices*, 55(11):3150–3158, 2008.

[AWS02] M.A. Alam, B.E. Weir, and P.J. Silverman. A study of soft and hard breakdown-Part II: Principles of area, thickness, and voltage scaling. *IEEE Transactions on Electron Devices*, 49(2):239–246, 2002.

[Bak08] R. Jacob Baker. *CMOS circuit design, layout and simulation*. IEEE Press, John Wiley & Sons, Inc., 2nd edition, 2008.

[BC10] Robert Brayton and Jason Cong. NSF Workshop on EDA: Past, Present, and Future. *IEEE Design & Test of Computers*, 27(3):62–74, May-June 2010.

[BGT81] Robert G. Bland, Donald Goldfarb, and Michael J. Todd. The ellipsoid method: a survey. *Operational Research*, 29(6):1039–1091, November-December 1981.

[BJ77] Peter W. Becker and Finn Jensen. *Design of Systems and Circuits for Maximum Reliability or Maximum Production Yield*. McGraw-Hill Book Company, 1977.

[Bla69] J.R. Black. Electromigration – A brief survey and some recent results. *IEEE Transactions on Electron Devices*, 16(4):338–347, 1969.

[cad] http://www.cadence.com.

[CCL+03] G. Chen, K. Y. Chuah, M. F. Li, D. S. H. Chan, C. H. Ang, J. Z. Zhcng, Y. Jin, and D. L. Kwong. Dynamic NBTI of pMOS transistors and its impact on device lifetime. In *Proceedings of International Reliability Physics Symposium*, pages 196–202, 2003.

[CFSL10] F.R. Chouard, M. Fulde, and D. Schmitt-Landsiedel. Reliability assessment of voltage controlled oscillators in 32nm high-κ metal gate technology. In *Proceedings of the 2010 European Solid-State Circuits Conference (ESSCIRC)*, pages 410–413, Seville, Spain, 2010.

[CGRP09] Ashutosh Chakraborty, Gokul Ganesan, Anand Rajaram, and David Z. Pan. Analysis and optimization of NBTI induced clock skew in gated clock trees. In *Proceedings of the Design, Automation & Test in Europe (DATE)*, pages 296 – 299, Nice, France, April 2009.

[CMFSL11a] Florian Raoul Chouard, Shailesh More, Michael Fulde, and Doris Schmitt-Landsiedel. An aging suppression and calibration approach for differential amplifiers in advanced CMOS technologies. In *Proceedings of the 2011 European Solid-State Circuits Conference (ESSCIRC)*, pages 251–254, Helsinki, Finland, 2011.

[CMFSL11b] Florian Raoul Chouard, Shailesh More, Michael Fulde, and Doris Schmitt-Landsiedel. An analog perspective on device reliability in 32nm high-κ metal gate technology. In *Proceedings of 2011 IEEE International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS)*, Cottbus, Germany, 2011.

[DG98] Geert Debyser and Georges Gielen. Efficient analog circuit synthesis with simultaneous yield and robustness optimization. In *Proceedings of IEEE/ACM International Conference on Computer-Aided Design*, pages 308 – 311, November 1998.

[DGSL09] S. Drapatz, G. Georgakos, and D. Schmitt-Landsiedel. Impact of negative and positive bias temperature stress on 6T-SRAM cells. *Advances in Radio Science*, 7:191–196, 2009.

[DH77] S. Director and G. Hachtel. The simplicial approximation approach to design centering. *IEEE Transactions on Circuits and Systems*, 24(7):363 – 372, July 1977.

[DHGSL10] Stefan Drapatz, Karl Hofmann, Georg Georgakos, and Doris Schmitt-Landsiedel. Impact of fast-recovering NBTI degradation on stability of large-scale SRAM arrays. In *Proceedings of the 2010 European Solid-State Device Research Conference (ESSDERC)*, pages 146–149, Seville, Spain, 2010.

[Die07] B. Dierickx. Scaling below 90nm: Designing with unreliable components, 2007.

[DK95] Abhijit Dharchoudhury and S. M. Kang. Worst-case analysis and optimization of VLSI circuit performances. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 14(4):481 – 492, April 1995.

[DS98] Norman Richard Draper and Harry Smith. *Applied Regression Analysis*, volume 326 of *Wiley Series in Probability and Statistics*. Wiley, New York, 3rd edition, 1998.

[ea06] I. Ahsan et al. RTA-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology. In *Symposium on VLSI Technology Digest of Technical Papers*, 2006.

[eet] http://eetimes.eu/en/learning-center/automotive-featured.html.

[Esh92] K. Eshbaugh. Generation of correlated parameters for statistical circuit simulation. *IEEE Transactions on Computer-Aided Design*, 11(10):1198–1206, 1992.

[FAH+08] Thomas Fischer, Ettore Amirante, Karl Hofmann, Martin Ostermayr, Peter Huber, and Doris Schmitt-Landsiedel. A 65nm test structure for the analysis of NBTI induced statistical variation in SRAM transistors. In *Proceedings of the 2008 European Solid-State Device Research Conference (ESSDERC)*, pages 51–54, Edinburgh, Scotland, UK, September 2008.

[FAH+09] Thomas Fischer, Ettore Amirante, Peter Huber, Karl Hofmann, Martin Ostermayr, and Doris Schmitt-Landsiedel. A 65 nm test structure for SRAM device variability and NBTI statistics. *Solid-State Electronics*, 53(7):773–778, July 2009.

[Fle87] Roger Fletcher. *Practical Methods of Optimization*. John Wiley & Sons Ltd, second edition, 1987.

[GDWM+08] G. Gielen, P. De Wit, E. Maricau, J. Loeckx, J. Martín-Martínez, B. Kaczer, G. Groeseneken, R. Rodríguez, and M. Nafría. Emerging yield and reliability challenges in nanometer CMOS technologies. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 1322–1327, Munich, Germany, 2008.

[GP09] Helmut Graeb and Xin Pan. Optimierung integrierter Schaltungen im Hinblick auf Alterungseinfluesse. In *newsletter edacentrum*, pages 11–14, October 2009.

[Gra07] Helmut Graeb. *Analog Design Centering and Sizing*. Springer, The Netherlands, 2007.

[GS97] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, second edition, July 1997.

[GWA93] Helmut E. Graeb, Claudia U. Wieser, and Kurt J. Antreich. Improved methods for worst-case analysis and optimization incorporating operating tolerances. In *Proceedings of the 30th ACM/IEEE Design Automation Conference*, pages 142 – 147, Dallas, Texas, June 1993.

[GZEA01] H. Graeb, S. Zizala, J. Eckmueller, and K. Antreich. The sizing rules method for analog integrated circuit design. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pages 343–349, San Jose, CA, USA, 2001.

[HH64] J. M. Hammersley and D. C. Handscomb. *Monte Carlo methods*. London Methuen & Co. Ltd, 1964.

[Hjo80] Urban Hjorth. A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rates. *Technometrics*, 22(1):99–107, February 1980.

[HLT83] Dale E. Hocevar, Michael R. Lightner, and Timothy N. Trick. A study of variance reduction techniques for estimating circuit yields. *IEEE Transactions on Computer-Aided Design*, 2(3):180–192, July 1983.

[HTH+85] C. Hu, S.C. Tam, F.C. Hsu, P.K. Ko, T.Y. Chan, and KW Terrill. Hot-electron-induced MOSFET degradation– model, monitor, and improvement. *IEEE Journal of Solid-State Circuits*, 20(1):295–305, 1985.

[Hu92] C. Hu. IC reliability simulation. *IEEE Journal of Solid-State Circuits*, 27(3):241–246, 1992.

[I.T] I.T.R.S. *International Technology Roadmap for Semiconductors (ITRS)*. http://www.itrs.net, 2009 edition.

[JRSR05] NK Jha, PS Reddy, DK Sharma, and VR Rao. NBTI degradation and its impact for analog circuit reliability. *IEEE Transactions on Electron Devices*, 52(12):2609–2615, 2005.

[JWL82] K. Jain, C. G. Willson, and B. J. Lin. Ultrafast deep UV lithography with excimer lasers. *IEEE Electron Device Letters*, 3(3):53–55, March 1982.

[KD95] Kannan Krishna and Stephen W. Director. The linearized performance penalty (LPP) method for optimization of parametric yield and its reliability. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 14(12):1557–1568, December 1995.

[KFHR01] M. Karam, W. Fikry, H. Haddara, and H. Ragai. Implementation of hot-carrier reliability simulation in Eldo. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 5, pages 515–518, Sydney, NSW , Australia, May 2001.

[Kil] Jack Kilby. Miniaturized electronic circuits. U.S. Patent 3,138,743. Issued June 23, 1964 (filed Feb. 6, 1959).

[KKAR06] Kunhyuk Kang, H. Kufluoglu, M. A. Alain, and K. Roy. Efficient transistor-level sizing technique under temporal performance degradation due to NBTI. In *Proceedings of the IEEE International Conference on Computer Design (ICCD)*, pages 216–221, San Jose, CA, USA, October 2006.

[KKW03] Georgia-Ann Klutke, Peter C. Kiessler, and M. A. Wortman. A critical look at the bathtub curve. *IEEE Transactions on Reliability*, 52(1):125–129, March 2003.

[LFG10] Bo Liu, Francisco V. Fernández, and Georges Gielen. An accurate and efficient yield optimization method for analog circuits based on computing budget allocation and memetic search technique. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 1106 – 1111, Dresden, Germany, 2010.

[LFJG09] Bo Liu, Francisco V. Fernández, Dimitri De Jonghe, and Georges Gielen. Less expensive and high quality stopping criteria for MC-based analog IC yield optimization. In *Proceedings of the 16th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 267–270, Yasmine Hammamet, Tunis, December 2009.

[Li10] Yan Li. *Robust Design of DRAM Core Circuits - Yield Estimation and Analysis by A Statistical Design Approach*. Phd thesis, Technische Universitaet Muenchen, Munich, Germany, 2010.

[Lie06] J. Lienig. Introduction to electromigration-aware physical design. In *Proceedings of the IEEE International Symposium on Physical Design*, 2006.

[LK89] Y. Leblebici and SM Kang. Simulation of MOS circuit performance degradation with emphasis on VLSI design-for-reliability. In *Proceedings of the IEEE International Conference on Computer Design (ICCD): VLSI in Computers and Processors*, pages 492–495, Cambridge, MA , USA, Octoboer 1989.

[LMM06] Z. Liu, BW McGaughy, and JZ Ma. Design tools for reliability analysis. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, pages 182–187, San Francisco, CA, USA, 2006.

[LQH+06] X. Li, J. Qin, B. Huang, X. Zhang, and J.B. Bernstein. A new SPICE reliability simulation method for deep submicrometer CMOS VLSI circuits. *IEEE Transactions on Device and Materials Reliability*, 6(2):247–257, 2006.

[LS94] Kenneth R. Laker and Willy M. C. Sansen. *Design of Analog Integrated Circuits and Systems*. McGraw-Hill, Inc, 1994.

[LSZ+09] Y. Lu, L. Shang, H. Zhou, H. Zhu, F. Yang, and X. Zeng. Statistical reliability analysis under process variation and aging effects. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, pages 514–519, San Francisco, CA, USA, 2009.

[LWH+07] Hong Luo, Yu Wang, Ku He, Rong Luo, Huazhong Yang, and Yuan Xie. Modeling of PMOS NBTI effect considering temperature variation. In *Proceedings of the 8th IEEE International Symposium on the Quality Electronic Design (ISQED)*, pages 139–144, San Jose, CA, USA, March 2007.

[LWM+93] MM Lunenborg, PBM Wolbert, PBL Meijer, T. Phat-Nguyen, and JF VerWeij. Press-a circuit simulator with built-in reliability model for

hot-carrier degradation. In *Proceedings of the European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF)*, pages 157–161, 1993.

[Mas09] Tobias Massier. *On the Structural Analysis of CMOS and Bipolar Analog Integrated Circuits*. Phd thesis, Technische Universitaet Muenchen, Munich, Germany, 2009.

[Mas10] Tobias Massier. *On the Structural Analysis of CMOS and Bipolar Analog Integrated Circuits*. Phd thesis, Technische Universitaet Muenchen, Munich, Germany, 2010.

[McP07] J.W. McPherson. Reliability trends with advanced CMOS scaling and the implications for design. In *Proceedings of the 2007 IEEE Custom Integrated Circuits Design (CICC)*, pages 405–412, San Jose, CA, USA, September 2007.

[MDJG12] E. Maricau, D. De Jonghe, and G. Gielen. Hierarchical analog circuit reliability analysis using multivariate nonlinear regression and active learning sample selection. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 745–750, Dresden, Germany, 2012.

[MFCSL11] Shailesh More, Michael Fulde, Florian Chouard, and Doris Schmitt-Landsiedel. Reducing impact of degradation on analog circuits by chopper stabilization and autozeroing. In *Proceedings of the 12th IEEE International Symposium on the Quality Electronic Design (ISQED)*, pages 1–6, Santa Clara, CA, USA, 2011.

[MG09] E. Maricau and G. Gielen. Efficient reliability simulation of analog ICs including variability and time-varying stress. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 1238–1241, Nice, France, 2009.

[MG10] E. Maricau and G. Gielen. Variability-aware reliability simulation of mixed-signal ICs with quasi-linear complexity. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 1094–1099, Dresden, Germany, 2010.

[MG11] E. Maricau and G. Gielen. Stochastic circuit reliability analysis. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 1–6, Grenoble, France, 2011.

[MGS08] T. Massier, H. Graeb, and U. Schlichtmann. The sizing rules method for CMOS and bipolar analog integrated circuit synthesis. *IEEE Trans-*

actions on Computer-Aided Design of Integrated Circuits and Systems, 27(12):2209–2222, December 2008.

[Ml95] G.E. Mueller-l. PASTA – the characterization of the inherent fluctuations in the fabrication process for circuit simulation. *International journal of circuit theory and applications*, 23(4):413–432, 1995.

[Moo65] Gordon E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8), April 1965.

[Moo75] Gordon E. Moore. Progress in digital integrated electronics. In *International Electron Device Meeting*, pages 11–13, 1975.

[Nas08] Sani R. Nassif. Process variability at the 65nm node and beyond. In *Proceedings of the IEEE Custom Integrated Circuits Conference (CICC)*, pages 1–8, San Jose, CA, USA, 2008.

[Noy] Robert Norton Noyce. Semiconductor device and lead structure. U.S. Patent 2,981,877. Issued Apr. 25, 1961 (filed July 30, 1959).

[NW00] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2000.

[PDW89] Marcel J. M. Pelgrom, Aad C. J. Duinmaijer, and Anton P. G. Welbers. Matching properties of MOS transistors. *IEEE Journal of Solid-State Circuits*, 24(5):1433–1440, October 1989.

[PG09] Xin Pan and Helmut Graeb. Degradation-aware analog design flow for lifetime yield analysis and optimization. In *Proceedings of the 16th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pages 667–670, Yasmine Hammamet, Tunis, December 2009.

[PG10a] Xin Pan and Helmut Graeb. Lifetime yield optimization: Towards a robust analog design for reliability. In *Handout of the 2010 Design, Automation and Test in Europe (DATE) University Booth*, Dresden, Germany, March 2010.

[PG10b] Xin Pan and Helmut Graeb. Reliability analysis of analog circuits by lifetime yield prediction using worst-case distance degradation rate. In *Proceedings of the 11th IEEE International Symposium on the Quality Electronic Design (ISQED)*, pages 861–865, San Jose, CA, USA, March 2010.

[PG10c] Xin Pan and Helmut Graeb. Reliability analysis of analog circuits using quadratic lifetime worst-case distance prediction. In *Proceedings of the*

2010 IEEE Custom Integrated Circuits Design (CICC), San Jose, CA, USA, September 2010.

[PG11a] Xin Pan and Helmut Graeb. Lifetime yield optimization of analog circuits considering process variations and parameter degradations. *Intech Publishing: Advances in Analog Circuits*, pages 131–146, February 2011.

[PG11b] Xin Pan and Helmut Graeb. Reliability optimization of analog circuits with aged sizing rules and area trade-off. In *Proceedings of the edaWorkshop11*, Dresden, Germany, May 2011.

[PG12] Xin Pan and Helmut Graeb. Reliability optimization of analog integrated circuits considering the trade-off between lifetime and area. *Journal of Microelectronics Reliability, Elsevier*, 52(8):1559–1564, August 2012.

[PH93] Shaowei Pan and Yu Hen Hu. Pyfs - a statistical optimization method for integrated circuit yield enhancement. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 12(2):296–309, February 1993.

[Pha06] Hoang Pham, editor. *Springer Handbook of Engineering Statistics*. Springer-Verlag London Limited, 2006.

[PSV01] K. Ponnambalam, Abbas Seifi, and Jiri Vlach. Probabilistic design of systems with general distributions of parameters. *International Journal of Circuit Theory and Applications*, 29(6):527–536, November/December 2001.

[QS08] Zhenyu Qi and Mircea R. Stan. NBTI resilient circuits using adaptive body biasing. In *Proceedings of the 18th ACM Great Lakes symposium on VLSI*, pages 285–290, Orlando, Florida, USA, May 2008.

[Rau02] Stewart E. Rauch. The statistics of NBTI-induced V_T and β mismatch shifts in pMOSFETs. *IEEE Transactions on Device and Materials Reliability*, 2(4):89–93, December 2002.

[Raz01] Behzad Razavi. *Design of analog CMOS integrated circuits*. The McGraw-Hill Book Company, 2001.

[RI97] Jane Radatz and Institute of Electrical and Electronics Engineers. *The IEEE Standard Dictionary of Electrical and Electronics Terms*. Institute of Electrical & Electronics Engineers, 1997.

[RMY03] S. Rangan, N. Mielke, and E.C.C. Yeh. Universal recovery behavior of negative bias temperature instability. In *IEEE International Electron Device Meeting 2003 Technical Digest*, pages 14.3.1–14.3.4, 2003.

[Sah10] Samar K. Saha. Modeling process variability in scaled CMOS technology. *IEEE Design & Test of Computers*, 27(2):8–16, March-April 2010.

[SB03] Dieter K. Schroder and Jeff A. Babcock. Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing. *Jounal of Applied Physics*, 94(1):1–18, 2003.

[Sch66] Y. A. Schreider. *The Monte Carlo method*, volume 87 of *International Series of Monographs in Pure and Applied Mathematics*. Pergamon Press, New York, NY, 1966.

[Sch97] Robert R. Schaller. Moore's law: past, present and future . *IEEE Spectrum*, 34(6):52–59, June 1997.

[Sch03] Frank C. Schenkel. *Tolerance Analysis and Design Centering of Analog Circuits, with Consideration of Mismatch*. Phd thesis, Technische Universitaet Muenchen, Munich, Germany, 2003.

[SEGA99] R. Schwencker, J. Eckmueller, H. Graeb, and K. Antreich. Automating the sizing of analog CMOS circuits by consideration of structural constraints. In *Proceedings of the Design, Automation and Test in Europe (DATE)*, pages 323 – 327, Munich, Germany, March 1999.

[SHL89] BJ Sheu, WJ Hsu, and BW Lee. An integrated-circuit simulator–RELY. *IEEE Journal of Solid-State Circuits*, 24:473–477, 1989.

[SP81] K. Singhal and J. Pinel. Statistical design centering and tolerancing using parametric sampling. *IEEE Transactions on Circuits and Systems*, 28(7):692–702, July 1981.

[SPV99] Abbas Seifi, K. Ponnambalam, and Jiri Vlach. A unified approach to statistical design centering of integrated circuits with correlated parameters. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 46(1):190–196, January 1999.

[SR85] R. S. Soin and P. J. Rankin. Efficient tolerance analysis using control variates. *IEE Proceedings G Electronic Circuits & Systems*, 132(4):131–142, August 1985.

[SRRP09] U. Sobe, K.H. Rooch, A. Ripp, and M. Pronath. Robust analog design for automotive applications by design centering with safe operating

areas. *IEEE Transactions on Semiconductor Manufacturing*, 22(2):217–224, 2009.

[SSG97] Peter J. Smith, Mansoor Shafi, and Hongsheng Gao. Quick simulation: A review of importance sampling techniques in communications systems. *IEEE Journal on Selected Areas in Communications*, 15(4):597–613, May 1997.

[STPW76] Terrence R. Scott and JR. T. P. Walker. Regionalization: A method for generating joint density estimates . *IEEE Transactions on Circuits and Systems*, 23(4):229–234, April 1976.

[TR07] C.M. Tan and A. Roy. Electromigration in ULSI interconnects. *Materials Science and Engineering: R: Reports*, 58(1-2):1–75, 2007.

[VOX09] B Vaidyanathan, A Oates, and Y Xie. Intrinsic NBTI-variability aware statistical pipeline performance assessment and tuning. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 164–171, San Jose, CA, USA, 2009.

[VOXW09] B Vaidyanathan, A Oates, Y Xie, and Y Wang. NBTI-aware statistical circuit delay assessment. In *Proceedings of the 10th International Symposium on Quality Electronic Design (ISQED)*, pages 13–18, San Jose, CA, USA, 2009.

[VWC06] Rakesh Vattikonda, Wenping Wang, and Yu Cao. Modeling and minimization of PMOS NBTI effect for robust nanometer design. In *Proceedings of the ACM/IEEE Design Automation Conference*, pages 1047–1052, San Francisco, CA, USA, July 2006.

[Wan] Frank Wanlass. Low stand-by power complementary field effect circuitry. U.S. Patent 3,356,858. Issued Dec. 5, 1967 (filed June 18, 1963).

[wika] http://en.wikipedia.org/wiki/High-k_dielectric.

[wikb] http://en.wikipedia.org/wiki/Smartphone.

[wikc] http://en.wikipedia.org/wiki/tablet_personal_computer.

[WRK+07] W. Wang, V. Reddy, A.T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao. Compact modeling and simulation of circuit reliability for 65-nm CMOS technology. *IEEE Transactions on Device and Materials Reliability*, 7(4):509–517, 2007.

[WRY+08] Wenping Wang, Vijay Reddy, Bo Yang, Varsha Balakrishnan, Srikanth Krishnan, and Yu Cao. Statistical prediction of circuit aging under pro-

cess variations. In *Proceedings of the IEEE Custom Integrated Circuits Conference (CICC)*, pages 13–16, San Jose, CA, USA, 2008.

[WSH00] E.Y. Wu, J.H. Stathis, and L.K. Han. Ultra-thin oxide reliability for ULSI applications. *Semiconductor Science and Technology*, 15:425–435, 2000.

[WVO97] Jacek Wojciechowski, Jiri Vlach, and Leszek Opalski. Design for non-symmetrical statistical distributions. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(1):29 – 37, January 1997.

[XCS+03] X. Xuan, A. Chatterjee, AD Singh, NP Kim, and MT Chisa. IC reliability simulator ARET and its application in design-for-reliability. In *Proceedings of the 12th Asian Test Symposium (ATS)*, pages 18–21, Xi' an, China, November 2003.

[YKHT87] Tat-Kwan Yu, Sung Mo Kang, I. N. Haji, and T. N. Trick. Statistical performance modeling and parametric yield estimation of MOS VLSI. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 6(6):1013 – 1022, November 1987.

[YQD+09] B. Yan, J. Qin, J. Dai, Q. Fan, and J. Bernstein. Reliability simulation and circuit-failure analysis in analog and mixed-signal applications. *IEEE Transactions on Device and Materials Reliability*, 9(3):339–347, September 2009.

List of Figures

1.1 Shift of the performance distributions from 300 Monte-Carlo simulation samples on a fresh (triangles) and 5-year-old (squares) Miller operational amplifier. . . . 3
1.2 Typical design flow of analog integrated circuits [Bak08] 4

2.1 The Bathtub Curve with effects of the device increasing wearout degradations. . . . 14
2.2 Effects of NBTI 16
2.3 Effects of HCI 18
2.4 General workflow of RelXpert 21
2.5 Repetitive simulation workflow of ELDO 22

3.1 Mapping from the statistical parameter space onto the performance space $\mathcal{F}$ considering parameter aging 36
3.2 Acceptance region (marked in grey) on the statistical parameter space (left) and the performance space (right) 38
3.3 Acceptance region (marked in grey) on the statistical parameter space (left) and the performance space (right) with transistor aging 40
3.4 Example: a simple current mirror 43

4.1 The relationship between yield analysis and worst-case analysis. . . . 48
4.2 Partial acceptance region $\mathcal{A}_{s,i}$ (in grey) in the two-dimensional statistical parameter space with one performance specification $f_{i,U}$ (slashed curve). . . . 49
4.3 The idea of the worst-case distance β_w. $\overline{\mathcal{A}}_s$ is the linear bounded approximated acceptance region. . . . 51
4.4 Yield values over different worst-case distances. . . . 53
4.5 The degradation of worst-case distance from t_0 to t_1 with respect to one performance specification f_U (slashed curve). . . . 59
4.6 Core of the proposed reliability optimization flow with one maximum area constraint. . . . 62

4.7 The generation and simulation flow of the aged circuit. The two grey boxes indicate the steps where fresh and aged sizing rules are checked respectively. 63
4.8 Illustration of yield optimization through the change of design parameters. 68
4.9 Illustration of yield optimization through the change of the nominal vector of statistical parameters. 69

5.1 Components contributing to the circuit performance sensitivities over transistor parameter aging 74
5.2 Linear performance model in the statistical parameter space. 76
5.3 Overview of the algorithm to predict aged worst-case distances and aged yield for one of the performance at different ages 80

6.1 Circuit schematic of the Miller operational amplifier 84
6.2 Additional area of each device after fresh circuit optimization on the Miller operational amplifier (MN1 and MN2, MP1 and MP2 have the same size). 88
6.3 Yield at lifetimes t=0 and t=10 years for four designs with increasing additional area requirements on the Miller operational amplifier. 89
6.4 Aging of yield value over time for four different Miller operational amplifier designs with different layout areas. The respective lifetime ends if the aged yield drops to a certain boundary. 89
6.5 Comparison results of $\beta_w(t)$ at respective time points for SR (a) and CMRR (c), and the corresponding relative errors for SR (b) and CMRR (d) of the Miller operational amplifier. 91
6.6 Circuit schematic of the folded cascode operational amplifier 92
6.7 Yield at lifetimes t=0 and t=10 years for four designs with increasing additional area requirements on the folded cascode operational amplifier. 93
6.8 Aging of yield value over time for four different folded cascode operational amplifier designs with different layout areas. The respective lifetime ends if the aged yield drops to a certain boundary. 94
6.9 Comparison results of $\beta_w(t)$ at respective time points for SR (a) and CMRR (c), and the corresponding relative errors for SR (b) and CMRR (d) of the folded cascode operational amplifier. 95

List of Tables

3.1 Example: sizing rules for a simple current mirror. 43

4.1 Worst-case distances and the corresponding yield value. 53

6.1 List of performance specifications for the Miller operational amplifier. 84
6.2 Simulation results of worst-case distance after applying the proposed optimization flow for the fresh and aged Miller operational amplifier. . 87
6.3 Aged worst-case distance prediction results in comparison with accurate values for different performance features of the Miller operational amplifier at t=10 years. 90
6.4 Simulation results of worst-case distance after applying the proposed optimization flow for the fresh and aged folded cascode operational amplifier. 93
6.5 Aged worst-case distance prediction results in comparison with accurate values for different performance features of the folded cascode operational amplifier at t=10 years. 95

Abstract in German

Im Zuge der fortschreitenden Skalierung integrierter Prozesstechnologien wird die Zuverlässigkeit analoger Schaltungen ein wichtiges Anliegen der Halbleiterindustrie. Diese Arbeit schlägt eine effiziente Methode zur Dimensionierung von analogen integrierten Schaltungen im Hinblick auf die Robustheit über die Lebensdauer hinweg vor. Die Methode beruht auf der Analyse und Optimierung der frischen Worst-Case-Abstände aller Schaltungseigenschaften als Robustheitsmaß bezüglich Fertigungsprozessschwankungen und Alterungseffekten der Transistoren. Während der Optimierung werden Dimensionierungsregeln für die frische und die gealterte Schaltung und Nebenbedingungen für die Fläche überprüft. Der Trade-off zwischen Schaltungslebensdauer und dem Preis, der im Hinblick auf Layoutfläche gezahlt wird, wird im Detail untersucht. Zur Beschleunigung der Abschätzung der Lebensdauerrobustheit stellt die Arbeit einen neuen Ansatz vor, bei dem die Worst-Case-Abstände der gealterten Schaltung mittels Empfindlichkeitsanalysen der frischen Schaltung abgeschätzt werden.