A METHODOLOGY FOR MODELING THE MANUFACTURABILITY OF INTEGRATED CIRCUITS

by
Eric David Boskin

Memorandum No. UCB/ERL M95/26

20 April 1995

A METHODOLOGY FOR MODELING THE MANUFACTURABILITY OF INTEGRATED CIRCUITS

by

Eric David Boskin

Memorandum No. UCB/ERL M95/26

20 April 1995

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

A Methodology for Modeling the Manufacturability of Integrated Circuits

by

Eric David Boskin

Abstract

A Methodology for Modeling the Manufacturability of Integrated Circuits

by

Eric David Boskin

Doctor of Philosophy in Engineering-Electrical Engineering and Computer Sciences

University of California at Berkeley

Professor Costas J. Spanos, Chair

This thesis presents a unified approach to Design for Manufacturability (DFM). A methodology is introduced wherein the three technology organizations within a company, process development, circuit design, and product engineering participate in the characterization, modeling and improvement of the manufacturability of IC products.

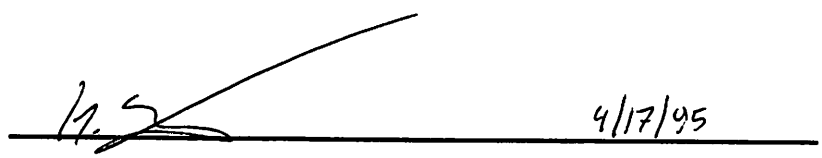
The methodology is based on an IC fabrication line description consisting of a small set of measurable process parameters which describe the variation in performance seen on a manufacturing line. These parameters are utilized in a physically based MOSFET parameter extraction technique which accurately predicts transistor characteristics over a wide range of process variation. The device models drive Monte Carlo circuit simulations which, in conjunction with manufacturing data, are utilized to build applications which improve product manufacturability.

Three specific applications are discussed. First, a performance prediction model was developed which uses the process parameters as measured on the manufacturing floor to predict the performance of fabricated integrated circuits before packaging and final test. The second application uses the performance prediction model for statistical process control (SPC) on product performance. Finally, design engineers use circuit simulation to systematically improve the manufacturability of the EPROM product for low power

operation by decreasing the sensitivity of the circuit to process variation. The fabrication line description, device models, and the three applications were developed and tested on a 1 Mbit EPROM produced with an industrial 1.2 μm CMOS process.

The advantages of the DFM approach developed in this thesis are the focus on manufacturing applications and the high level integration of simulation models and manufacturing data. The results of this thesis include production-ready applications, such as an IC performance prediction model and an SPC procedure, which can be implemented on an IC manufacturing line to improve efficiency and detect problems, thereby lowering costs. Also, the use of common, measurable process parameters integrate the design and manufacturing applications, providing additional, important cross-checks which verify the models and applications before their use.

The dissertation of Eric David Boskin is approved:



4/17/95

Professor Costas J. Spanos — Date

Committee Chairman

To Trudi

for her unending love and support and

to Natalie

for bringing so much happiness into my life.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my most heartfelt appreciation and gratitude to my research advisor, Professor Costas J. Spanos, for his support, guidance and encouragement throughout my doctoral program. He set a standard for hard work, dedication and excellence that I hope to keep with me in all my future pursuits. I would also like to thank the rest of my committee members, Professors Bernhard Boser, Ping Ko and John Rice, for their comments, suggestions and guidance of my thesis work. Finally, I am grateful to Dean David Hodges for his support of my project and of Computer Integrated Manufacturing research at Berkeley.

I would like to give special thanks to Dr. Sherry Lee, who not only graciously answered an unending series of questions during our five years sitting beside each other, but gave me support in meeting life's everyday challenges. I would also like to acknowledge my fellow members of the BCAM group, including Roawen Chen, Mark Hatzilambrou, Herb Huang, Anna Ison, Sovarong Leang, Tony Miranda, David Mudie, Xinhui Niu, Manolis Terravitis, Pam Tsai and Crid Yu, as well as BCAM graduates Bart Bombay, Dr. Norman Chang, Dr. Raymond Chen, Zeina Daoud, Haifang Guo, Mehdi Hosseini, Hao-Cheng Liu, Shang-Yi Ma, Dr. Zhi-min Ling, Tom Luan, Professor Gary May, Dr. Fariborz Nadi, Dave Rodriquez, John Thomson and Eddie Wen, for making this time much more productive and enjoyable.

I would like to acknowledge other Berkeley students and staff for their contributions. Thanks to Dr. Ken Nishimura for his help with computers and transistors, Andy Burstein, Dr. Paul Haskell and Dave Lidsky for reminding me to exercise, Dave Cline for accompanying me to Physics Seminars, Sean Cunningham for much needed study breaks, Dr. Cormac Conroy, Dr. Cindy Keys, Dr. Paul Landman, Dr. Klaus Schuegraf and Professor Greg Uehara for their friendship and support. I would especially like to Dr. Sheila Humphreys and Pam Atkinson for providing me with excellent opportunities for personal growth and diversity.

As this research program was very integrated with development at ATMEL Corporation, I would like to thank several ATMEL employees. I especially thank George Korsh, whose expertise, guidance and vision played a vital role in this research. I acknowledge May Lai and Tim Pearce for the special testing, Mehdi Jazaeri for using the DFM software and doing the design changes on the EPROM circuits, Dr. Michael Chern, Ed Hui and Dr. Gust Perlegos for many valuable discussions, Nelson Ingersoll for tremendous support on the CIM system, Dean Allum and Todd Randazzo for their invaluable contributions to the parameter extraction methodology, Richard Lee and Steve Ozoa for assistance on the UNIX system, and Dan Terry and Bill Brisko for many interesting discussions.

I would also like to thank Dr. Len Gruber of Digital, Professor Linda Milor at the University of Maryland, and Bill Davis of Motorola for their important contributions.

I would like to thank my friends for helping me through this effort. First, let me thank Denny and Tamra Renfrow for responding to my endless email, helping me with awk and my car engine, feeding and housing me on my trips to ATMEL and for being such special friends. Special thanks to Steve and Julie Lassig and Frank and Leslie Van Veenendaal for their long friendships and encouragement. I thank Bruce Sindahl for sharing his unique perspective on life and Dr. John Kibarian for his vision and for reminding me to finish this thesis.

I would like to acknowledge the tremendous support of my family, especially my mother Ellie and father Marvin, who taught me the value of education, knowledge, hard work and love. Thanks to my brothers Michael and Stephen, my cousins Melissa and Jeff, for all their love and support, and to Warren for the trips to high country. Lastly and most importantly, I want to express my love for my wife Trudi and my daughter Natalie for putting up with my long hours and giving me all the love I could ever ask for, and then more.

I most gratefully acknowledge the joint sponsors of this work, ATMEL Corporation, the SRC, the California Competitive Technology Program, and SEMATECH.

Chapter 1

Introduction

In today's competitive semiconductor manufacturing environment, it is important to anticipate the effects of processing variation during the design of an integrated circuit (IC) product, and then characterize and control this variation while producing it [1][2][3]. In this work we present a methodology for characterizing process variation, enabling both circuit design improvements and manufacturing control applications. Manufacturability modeling includes not only techniques to improve the manufacturability of circuit designs, such as improving circuit yield, but also tying these techniques directly to the manufacturing line.

The organization of a corporation involved in the manufacture of semiconductor products typically contains three business units. Commonly referred to as design, process development and manufacturing, these three organizations have grown to be separate sectors within the company. In fact, the success of the IC industry has in part been due to the use of a design style, first formalized by Mead and Conway [4], which separates IC designers from most considerations of the technology and the manufacturing process. However, the increasing cost and complexity of a modern IC manufacturing line is necessitating increased communication between designers, manufacturing and technology to

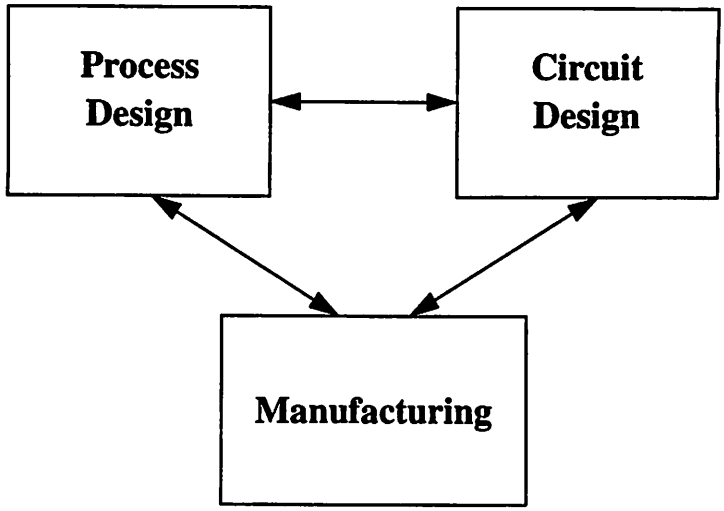


Figure 1-1 The Infrastructure of Design for Manufacturability

deliver profitable products in a timely fashion. Design for Manufacturability (DFM) techniques strive to impact the design and manufacture of an IC product in light of specific technology and manufacturing processes in order to improve the manufacturability of the product. As shown in Figure 1-1, manufacturability implies the integration of process design, circuit design and manufacturing activities. This integration will enhance the communication between these sectors while improving manufacturing capability.

1.1 Motivation

The semiconductor manufacturing process is best described as a series of steps, up to 300 in a modern CMOS process, which turn a bare silicon wafer into packaged ICs. These steps include the introduction and redistribution of impurities into the silicon, the growth or deposition of layers on the wafer and the patterning of the these layers. Finally, the wafer is tested and the die on the wafer are separated and put into packages. As with any manufacturing process, there are uncontrolled variations in the process which cause product quality to vary. DFM includes the techniques for characterization of this variation and methods to predict and minimize its effect.

Historically, the focus of DFM work has been *yield* prediction and optimization. Simply stated, the yield of an IC is the fraction of manufactured parts which can be sold to customers. Yield can be further decomposed into the fractions of parts sold at different performance levels, reflecting the distribution of product performance caused by uncontrolled process variation. Previous work in DFM analyzed the distribution of product performance, and optimized the circuit design or fabrication process for maximum yield. DFM for circuit or process optimization is motivated by the monetary reward of increased yield, including both reduced scrap and having fewer products manufactured with low performance which must be sold at a reduced price.

Recently, there has been increased focus on improving the *learning curve*, defined as the amount of time from the introduction of a new process or circuit until an acceptable level of yield is established. DFM activity, and in particular, improved communication between process, design and manufacturing, can decrease the time required to reach high yield levels. This brings both a monetary saving and a competitive advantage to an organization, justifying the DFM process.

In addition to these underlying motivations, the goal of this thesis is to develop a DFM approach which allows the development of applications for use directly in IC manufacturing. Such applications will improve the efficiency and reduce the cost of IC manufacturing. Given the high cost of IC manufacturing lines and the product wafers themselves, the potential monetary savings from manufacturing applications is quite large. For example, the statistical process control (SPC) application developed in this thesis will generate an alarm if a lot is incorrectly tested. If a lot contains several thousand parts worth an additional $10 as fast parts, $20,000 can be gained by the successful retesting of a single lot.

1.2 Thesis Overview

The methodology presented in this work is based on an IC fabrication line description consisting of a small set of measurable process parameters which describe the variation in performance seen on a manufacturing line. The methodology models IC product performance in terms of this simple set of process characterization data. These parameters are utilized in a physically based MOSFET parameter extraction technique which combines physical measurements, global optimization and regression modeling of key fitting parameters to accurately predict transistor characteristics over a wide range of process variation. This straight-forward parameter extraction methodology is applied to a 1.2 μm CMOS EPROM process using a SPICE Level 3 MOSFET model with a resulting accuracy of better than 10% across a large range of process and temperature variation.

The MOSFET model in conjunction with manufacturing data is used to develop DFM applications. The first application is a model for use in manufacturing for predicting the performance of products before final test. The same set of process parameters as is used in the MOSFET model can be measured during production, both in-line and at electrical test, and used to predict the performance, such as the operating frequency of a microprocessor, of fabricated parts. A compact, product-specific performance prediction model is built from a combination of simulation results and manufacturing data. The performance prediction model developed for the access time of the EPROM achieved a standard deviation of regression of less than 2% of the access time. This high level of accuracy is needed because of the large expense of incorrect predictions.

The second application is a statistical control procedure for IC performance, which is applied to EPROM production. Control charts based on the prediction model are developed for fabrication line statistical process control (SPC). These control charts can be used to identify process shifts or the incorrect performance testing of EPROM product lots.

The third, more traditional application utilizes the MOSFET model for statistical circuit simulation to test and improve the manufacturability of an IC product. As an example, the manufacturability of a 1 Mbit CMOS EPROM intended for low voltage application is investigated. Monte Carlo experiments using circuit simulation are used to find the sensitivity of the EPROM to process variation.

1.3 Thesis Organization

This thesis is organized as follows. Chapter 2 defines the manufacturability problem and discusses previous work in this area. Chapters 3 and 4 present the theory underlying manufacturability modeling. Chapter 3 discusses the fabrication line description, the device modeling method, and Monte Carlo experiments using circuit simulation to model circuit performance, and Chapter 4 presents the applications for improving IC manufacturability. The next two chapters implement the methodology presented in Chapters 3 and 4. In Chapter 5, the fabrication line description and device models are built for a production 1.2 μm CMOS EPROM process. Then, Chapter 6 uses the models from Chapter 5 to build the three manufacturability applications for a commercial 1 Mbit EPROM. Finally, Chapter 7 discusses conclusions and future work.

Chapter 2

Problem Formulation and Previous Work

This chapter presents the formal problem definition and previous work in manufacturability. An understanding of the problem and previous work toward solutions will set the context for the direction and contribution of this research. The chapter begins by introducing notation and continues by formulating the yield problem. Previous work is presented and discussed, including the limitations which motivate the methodology pursued in this work.

2.1 Notation and Variable Definitions

The notation used in this report is as follows. For a variable x:

- x lower-case, indicates a scalar,
- $\mathbf{x}$ lower-case and bold, indicates a column vector with elements x_i,
- $\mathbf{X}$ upper-case and bold, indicates a matrix with columns $\mathbf{x}_i$ and elements x_{ij}.

This report will develop relationships between process parameters and IC product performance. An IC process is the specific series of steps used to manufacture a semiconductor product, with each step requiring specific materials and settings on the manufacturing equipment. *Process parameters* refer to measurable quantities which characterize the

result of a small number of steps, for example, the sheet resistance or thickness of a layer or the length and width of a feature. Examples of process parameters in an MOS process include oxide thickness and polysilicon linewidth. The result of the IC fabrication process are circuits, which must meet certain specifications to be sold by the manufacturer. These specifications will be referred to as the circuit *performances*, and include the speed of a microprocessor or memory device, and the gain and bandwidth of an amplifier.

Manufacturability models will explain the variation in IC product performance in terms of a set of process parameters. The process parameters will be denoted with the variable x. The vector containing each parameter will be denoted $\mathbf{x}$, while $\mathbf{X}$ is the $(n \times p)$ matrix containing n samples of p process parameters. Similarly, the performance of the circuit will be denoted by the variable y. The range of variation seen on the manufacturing line across the group of process parameters or performances forms a multidimensional region which will be referred to as a *process* or *performance space*, respectively.

The variation of both the process parameters and circuit performances depend on the group which is sampled. There are several components to the variation, including die, wafer, lot and total variation. The calculated variance of process parameters or performances will increase as the sample is taken from within a wafer, within a lot or across several lots. It is important to use the appropriate grouping for each application. Variance will be denoted by σ^2, subscripted by either the variable name or the sampling group, as appropriate. When considering multiple variables, it is important to consider not only their variances, but also how strongly their variation is related to each other. This information is contained in the variance-covariance matrix, denoted as $\mathbf{V}(\mathbf{x})$.

2.2 The Manufacturing Yield Problem Formulation

Manufacturability is commonly associated with increasing the yield of a circuit [1], which is formalized as a problem of yield estimation [5] with a subsequent optimization. Yield is defined as the fraction of manufactured parts which meet all product specifica-

tions and can therefore be delivered to customers. The range of variation across the product performances defines a performance space, R_y. In general, upper and lower bounds are associated with each performance specification. Therefore, the region of acceptable IC performance, A_y, is defined as:

$$A_y = \{y \in R_y | \text{each y is within specification}\} \quad (1)$$

The performances of a given IC design are determined by the values of the process parameters for each manufactured lot. This leads to the concept of the *acceptability region*, A_x, which is the region within all process space, R_x, which results in acceptable IC performance, that is:

$$A_x = \{x \in R_x | y \in A_y\} \quad (2)$$

Given the definition of the acceptability region, the yield, Y, of a circuit can be calculated as:

$$Y = \int_{A_x} f_x(x)\, dx \quad (3)$$

where f_x is the joint probability density function for the process variables.

Up to this point we have considered a fixed circuit design, however, the circuit topology and transistor sizes also effect yield. In fact, choosing the sizes of transistors for maximum yield under the given performance constraints is often the goal of manufacturability optimization. Including the circuit design parameters, **q**, the overall manufacturability optimization problem can be written as:

$$\max_{x,q} \left\{ Y(x,q) = \int_{A_x(q)} f_x(x)\, dx \right\} \quad (4)$$

Modifying the circuit design parameters impacts yield by changing the acceptability region.

It is commonly assumed that the process parameters can be decomposed into a nominal process, $\mathbf{x}_0$, and the normally distributed random variation of each parameter, **d**, often called the *process disturbances*, that is:

$$x = x_0 + d \tag{5}$$

$$d \sim N\left(0, \sigma_x^2\right) \tag{6}$$

Therefore, the targets of optimization are the circuit design parameters, **q**, the nominal process, $\mathbf{x}_0$, and the process variability, **d**.

Circuit design optimization improves yield by altering the circuit given a fixed process. The design parameters, **q**, include the length and width of transistors and the topology of the circuit. This work will include the manufacturability application of circuit design optimization.

Previous work in manufacturability optimization has adjusted the nominal process to optimize yield through *design centering*. Design centering starts by finding the region of acceptability for a process, and moves the process targets so that nominal conditions produces circuits centered in the region of acceptability, hence optimizing the yield [6]. This work will use a fixed nominal process.

The process disturbances are mainly the result of equipment variation. Although methods to reduce equipment variation are starting to be developed, manufacturability work has generally assumed the disturbances are fixed. This work will develop methods to understand and characterize process variability, but not reduce it.

2.2.1 Parametric versus Functional Yield

The yield problem is divided into two areas, functional and parametric yield. Functional yield loss, defined as the fraction of manufactured parts which fail to operate at any level of performance, is commonly associated with defects introduced during manufacturing that cause circuit malfunction. Parametric yield loss is the fraction of functional parts

which do not operate to the desired performance specifications, for example, circuits which operate only at very low speeds contribute to parametric yield loss. Parametric yield loss is usually caused by the process disturbances, that is, the small perturbations in the manufacturing process such as a change in the critical dimension of the transistor geometry [7].

Good examples of functional Design for Manufacturability (DFM) techniques are redundancy and error correction. In these techniques, circuitry is added with the specific goal of fixing potential functional defects. In fact, redundancy is being used in certain applications, especially memories, to "repair" slow circuits. One design trade-off is that the area of the IC is increased to add the circuitry which allows proper circuit operation in the case of spot defects. Also, there is performance overhead associated with these techniques, so that the performance of the circuit decreases. However, redundancy and error correction greatly increase the yield of memory ICs and they are widely used techniques.

However, there is no technique with such success in improving the parametric yield of integrated circuits such as microprocessors, PALs, etc. Despite widespread concern and papers in the literature, there are few widely accepted DFM techniques. The methodology presented in this report focuses on parametric yield improvement.

2.3 Previous Work

The field of statistical circuit design represents the first systematic efforts to model and improve parametric yield [8]. The underlying concept is that perturbations in the manufacturing process change the performance of the silicon devices and therefore cause the performance yield fluctuations seen in final test. There are several important components in statistical design, as shown in Figure 2-1. First, the manufacturing process must be described in a way which characterizes the process perturbations responsible for yield loss. Second, this manufacturing line variation must be mapped into the performance variation of the fabricated devices. Often, a combination of process and device simulation is

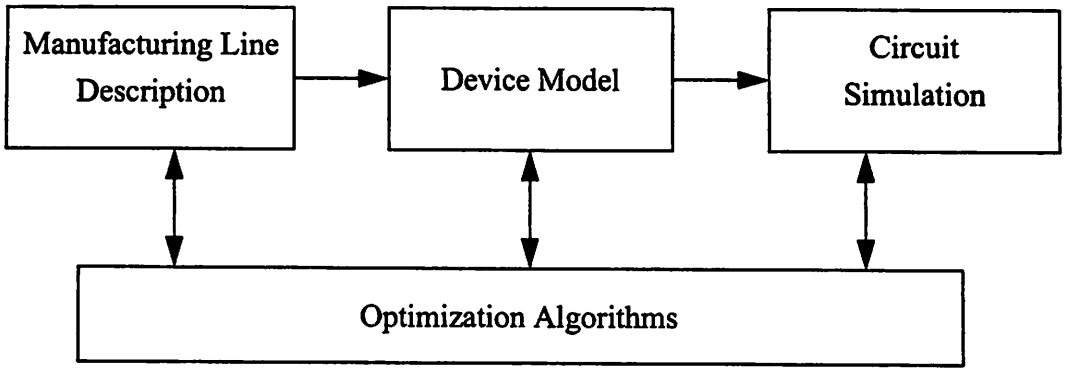


Figure 2-1 Elements of Statistical Circuit Design

used to map process variation into the changes in transistor characteristics. Third, circuit simulation provides the link between the transistor characteristics and the final circuit performance. Last, an algorithm is used to optimize the yield.

Despite these efforts, DFM still has found limited use in the semiconductor industry. However, Motorola [9], Philips [10], Texas Instruments [1] and Harris Corporation [11] have developed statistical design tools which influenced this project. This research will focus on the difficulties faced by the application of statistical design techniques. In the following subsections, each of the components in Figure 2-1 will be briefly examined. This section will finish with a discussion of the important issue of manufacturing integration. These issues will be revisited in greater detail as the methodology is presented in Chapter 3.

2.3.1 Manufacturing Line Description

Early work in statistical modeling relied heavily on the use of process simulation to describe the fabrication process. This use of simulation brings the accompanying problem of tuning the simulator to match the actual manufacturing line [12]. For example, the mean and standard deviations of process simulation variables such as the time in a furnace

step or the diffusivity of boron would be adjusted so that simulation results would match the statistics of measurements such as the MOS transistor channel length and consequently the device current. In other words, early statistical device models were verified by matching actual device or circuit performance distributions [10][11]. Historical electrical test data was used to determine the mean and standard deviation of device current, and then the tuning process would match the distribution produced in simulation to the historical device current distribution.

There are several problems with this approach. First, tuning is often a time consuming procedure. Second, there is no guarantee that the mapping of the process to the device characteristics is correct simply because the distributions match. Third, this approach assumes a stable production line which has attained a state of statistical control. Modern fabrication lines, however, are constantly undergoing evolutionary improvements, and a more accurate and timely verification strategy is needed. In other words, there is no mechanism to alarm when the statistical model is no longer valid, and if there were, it is not convenient to re-tune the models. In short, the problem with previous DFM work centers on the fact that these methods are not sufficiently integrated with the manufacturing line.

These problems are addressed in this work by avoiding the tuning problem entirely. This work does not rely on process simulation, instead, it uses transistor characterization data collected during production in conjunction with straightforward device physics. The use of performance prediction models which are used on the manufacturing line in conjunction with physically based device models improves the mapping from the process to the device characteristics. Further, the performance prediction model will generate an alarm when the simulation model no longer matches the manufacturing line output. In the case of an alarm, the physical basis of the device model will allow for a simple mechanism to align the models to the current manufacturing process. Another important feature of our work is that the use of manufacturing data ensures that the parameters are measurable

using production line capable techniques, and that they explain most of the variation seen on an actual fab line. The manufacturing applications detailed in this paper show that this method can be driven effectively by fabrication line data.

2.3.2 Device Modeling

A transistor, or any semiconductor device model, is used to calculate the response of the device to the voltage and current conditions in a circuit. This response models the I-V curve, which is the current versus voltage characteristic that defines the electrical properties of the device. In this work, the term device model refers to a compact model such as a SPICE transistor model, as opposed to a numerical model which simulates the physics of the device. The device model is comprised of a set of equations with coefficients that are specified to fit the transistor I-V curves for a specific process. This set of coefficients is referred to as the *device model parameters,* and the procedure for specifying them is called *parameter extraction.*

Early work in statistical circuit design allowed the use of a different set of model parameters for each value of the process parameters to be modeled. For each point in process space, process and device simulation would produce an I-V curve. Parameter extraction was accomplished by simple optimization algorithms. One problem with this approach is the large number of parameter sets required. The second, more important issue is the loss of physical relationships between the process and the device. With a large number of optimized models, it can be difficult to extract the physical cause from changes in device performance.

The work presented in this report is based on physically based parameter extraction techniques. The physical basis allows a smaller number of parameters to cover the process space and retains the relationship between the process and devices. The techniques used in this work have been strongly influenced by industry research in this area. Motorola has developed a physically based bipolar model [13] which has been utilized in a statistical

design tool which includes designed experiments to study performance variation of IC designs [9]. Similarly, Philips has developed an analytical parameter extraction technique [14] which was used for sensitivity and process capability analysis [10].

2.3.3 Circuit Simulation and Optimization

The use of the SPICE circuit simulator to estimate circuit performance is common practice and will be used in this work. Gradient and computer-based experimental design methods have been used to increase yield by optimizing the transistor sizes or topology of a circuit for a specific process. The gradient method is a standard non-linear optimization technique which can be implemented within a circuit simulator for use in yield optimization [15]. Experimental design software packages have focused on the use of Taguchi Robust Design methods to select optimum transistor sizes [16][17][18].

In contrast, our work does not use formal optimization techniques to improve yield, as the focus of this work is integration with manufacturing. Instead, graphical descriptions of the process sensitivities combined with the insight of the circuit designer are used to modify a design.

2.3.4 Integration with the Manufacturing Line

In general, there has been insufficient focus in statistical design and DFM research on integration with manufacturing. Two good examples of manufacturing applications of statistical modeling have focused on reducing the amount of testing done at final test. Brockman used statistical modeling to eliminate high and low temperature testing of ICs based on room temperature results [19]. Milor optimized the ordering of final tests to minimize the tester time required to identify failing parts [20]. However, the development of these prior manufacturing applications were based solely on simulation results.

Although previous efforts in physically based statistical device modeling have developed MOS transistor models which are integrated with manufacturing data, they have pri-

marily been applied to yield prediction and circuit optimization during the IC design phase. However, there are important advantages to be obtained in implementing manufacturability solutions which further integrate with a specific manufacturing line.

To address this problem, this work shifts the focus of DFM activity from IC design to the integration of design, process development and manufacturing. To this end, the basis for this work is a single set of parameters, measurable on the manufacturing line which support both design and manufacturing applications. The advantages of this approach include the use of these parameters in the device model, which avoids the problems of optimized parameters, and their routine measurement on the production line, supporting the tracking and control of product performance. Importantly, these parameters have physical meaning to the process, design and manufacturing groups, fostering inter-group discussion of manufacturability issues.

This work developed a manufacturing application to include in the DFM methodology, that is, a performance prediction model for IC binning. This application served to validate the manufacturing line, transistor and circuit models, as well as contributing to the manufacturing line itself. A design oriented application is also presented, which is used to improve the manufacturability of a 1 Mbit EPROM.

2.4 Summary

This chapter discussed the formalism of the manufacturability problem and previous work toward solutions. The limitations of statistical circuit design were presented, focusing on the limited integration of manufacturing into the DFM process. Integration will be the emphasis of the DFM methodology presented in Chapter 3, which discusses the theory underlying manufacturability modeling. and specifically each block shown in Figure 2-1.

Chapter 3

Manufacturability Modeling Methodology

This chapter gives an overview of the techniques and theory used for manufacturability modeling. The chapter begins with a brief presentation of a specific application developed during this research - performance prediction modeling - to give the context for the discussion. Then, given the application, the technique for mapping process variation into circuit variation is presented. The mapping is comprised of three sections, the fabrication line description, the device model and the use of circuit simulation.

3.1 Overview of Performance Prediction Model Building

The application developed in this work is for the performance-binning[1] of manufactured parts [21]. In-line measurements and electrical test results for a manufactured wafer can be used to predict circuit performance before final wafer test or assembly. With "just in time" manufacturing planning, only parts which meet performance requirements need be assembled.

1. Commodity ICs, such as memories, microprocessors, EPLDs, etc., are performance tested and divided into speed bins, and then priced accordingly.

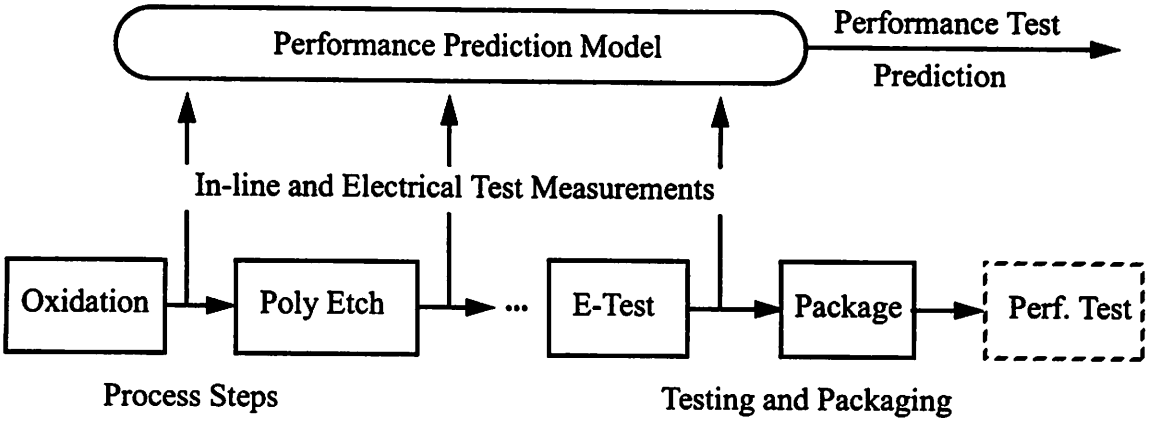


Figure 3-1 The Performance Prediction Method.

An overview of the manufacturing application of the performance prediction model is shown in Figure 3-1. The manufacturing process is represented by a series of steps, such as gate oxide growth, polysilicon gate definition, etc. The outcome of each step can be measured in-line during processing, and can be used to continuously refine an estimate of the performance of the produced ICs. After processing, a series of electrical tests are done on each wafer. These measurements are commonly used to monitor the fabrication process, and as a first test to assure the compliance of the manufactured goods.

There are important manufacturing uses for the performance prediction model. The performance of the circuits on wafers can be estimated early in the manufacturing process; first, from the in-line test data, and then from electrical test results, all before assembly of the die into packages. In-line binning will improve manufacturing planning by providing actual wafer and lot specific yield estimates instead of *a priori* estimates. This improvement in manufacturing scheduling and planning is particularly important for fabrication lines running at capacity with specific performance targets to produce, making it possible to expedite higher performance wafers through the fabrication line, or scrap the lowest performing ones [22]. After the early prediction based on in-line measurements, electrical

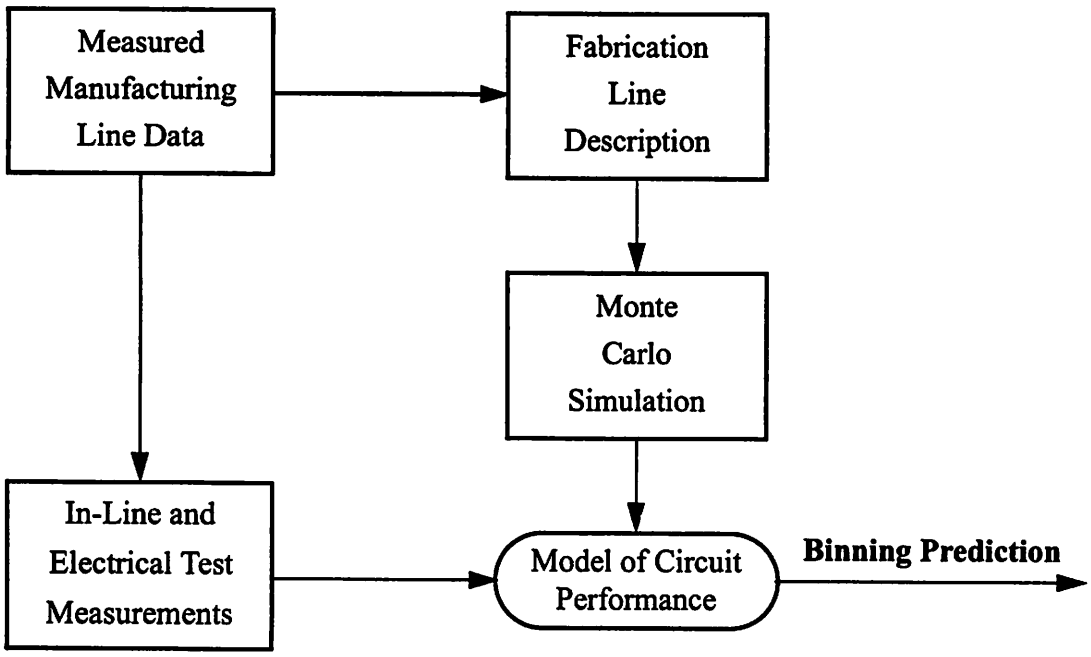


Figure 3-2 Overview of Performance Model Development

test data can be used to produce a more accurate estimation of performance. This can further aid in planning, particularly in the case where final performance testing can not occur until after packaging. For example, fast parts can be directed to be packaged in higher performance packages.

Figure 3-2 gives an overview of performance prediction model development. First, a description of the fabrication line is needed to simulate the process. This description, including the identification of the key process variables and their nominal values, range and correlation, is created from measurements taken on the manufacturing line. Using this description, instances of the process are generated for use in a Monte Carlo simulation. The device model, discussed in Section 3.4, is used to map the process into a SPICE transistor model for use in circuit simulation. The results of the simulation give the circuit performance at each point in the process. Using the process instances and the simulation

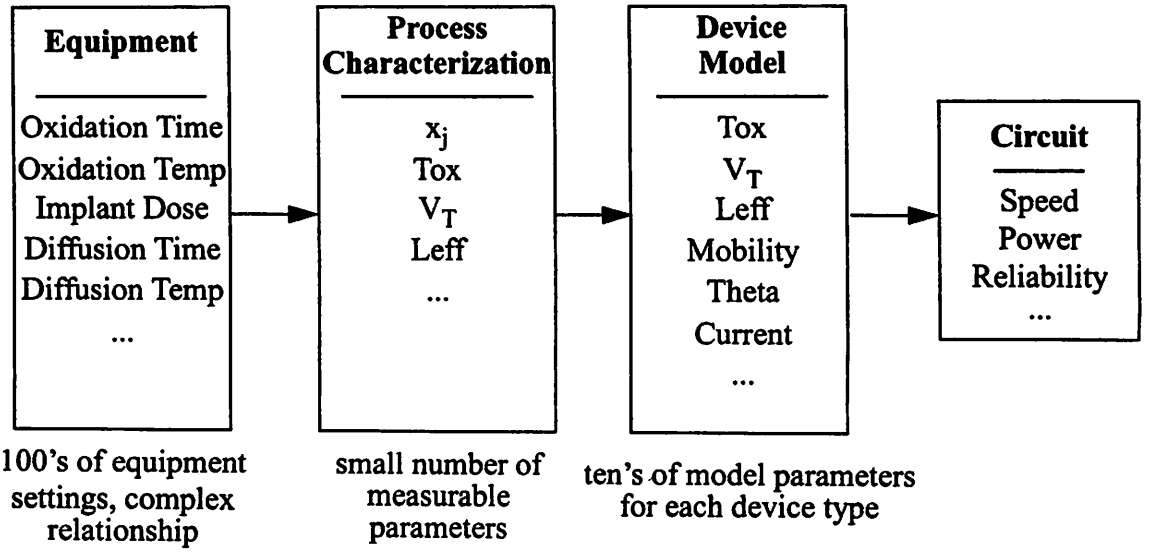


Figure 3-3 Manufacturability Modeling Regimes (MOS process example)

results, a performance model is built which predicts the performance of manufactured parts from in-line process measurements.

3.2 Modeling Hierarchy

Figure 3-3 illustrates the regimes in which manufacturability modeling can be accomplished [23]. In general, the variables listed at one level are used to model the variables at the next level. Process simulation models usually work at the level of equipment settings, such as the time and temperature settings for an oxidation step. These settings, also known as the *recipe,* determine the junction depth, doping level, or layer thickness, *etc.*, which results from the process step. Process simulators model the complex physical mechanisms governing diffusion, oxidation, *etc.*, to calculate the physical device characteristics from variables defining the chamber environment. After process simulation, a device simulator uses the cross-section, profile and Poisson's equation to calculate the device I-V characteristics. The I-V characteristics are then used for device model parameter extraction. In

terms of a MOS transistor model, the parameters in the model include oxide thickness and threshold voltage, as well as mobility reduction and short channel effects.

The advantages of combined process and device simulation include the ability to investigate device characteristics before the manufacturing process is complete, and study details of the device structure. However, a modern IC process uses dozens of pieces of equipment and hundreds of steps, making it difficult to accurately determine these characteristics using simulation.

The manufacturability modeling methodology presented in this work finds the process characterization parameters by measurement rather than through simulation. This reduction in the level of complexity is possible because these few, measurable process characteristics can accurately determine the device properties. Further, the physical basis of the compact device simulation models is exploited, maintaining a strong relationship between the device model parameters and the manufacturing process. These critically important relationships [13][23] which enable the methodology can be summarized as:

1. *There are a few measurable process characterization parameters which account for the variability seen in product performance.*
2. *There is a straightforward, physical mapping between the measurable process characterization parameters and the device model parameters.*

Finally, the device model is used to determine the performance of the final circuit product. A designed experiment is run to determine the circuit performance across the space of process variability. In this work, the experiment was designed using a Monte Carlo method, in order to get even coverage of the process space regardless of dimensionality. The Monte Carlo technique selects random points across the process space. At each point, the process parameters are used to create a SPICE transistor model, which are used in a circuit simulation to determine circuit performance at that point in the process. Thus, circuit simulation is used to map the process space into the circuit performance space.

The next three Sections of this chapter will look in detail at the theory behind each level in this hierarchy: the fabrication line description, mapping the process to the device model, and mapping the device model to circuit performance.

3.3 Fabrication Line Description

The objective of the fabrication line description is to help create a computer-based experiment in order to fit response surface models that relate process measurements to product performance. Given this, the description should capture the maximum possible range of variation seen on the fabrication line. For experimental economy, the concept of parameter correlation is used to eliminate simulations for unlikely combinations of parameter values.

The modern semiconductor manufacturing process, containing hundreds of steps, has inherent variation which effects the performance of fabricated ICs. Despite the complexity of semiconductor processing, a few measurable parameters can explain the bulk of the variation in the performance of an IC. In other words, the process is not modeled on the equipment settings, but rather on a small set of parameters which describe the results of the process. These parameters can be divided into two sets, the parameters which are important in explaining the variation for all ICs, and the set of parameters specific to each design.

Parameter sets suitable for general characterization of a modern MOS [1] or bipolar process [13] have been previously described. This work extends the description for an MOS process to include analog circuit design considerations such as the transistor body effect. The key to formulating the parameter set for a given process is to find the parameters which are both measurable and explain the performance variation seen in the process.

Some product specific parameters must also be considered. For the EPROM example presented in this work, the gate oxide used in the EPROM cell and the resistivity of the wordline layer are included as additional parameters. In certain analog circuit designs,

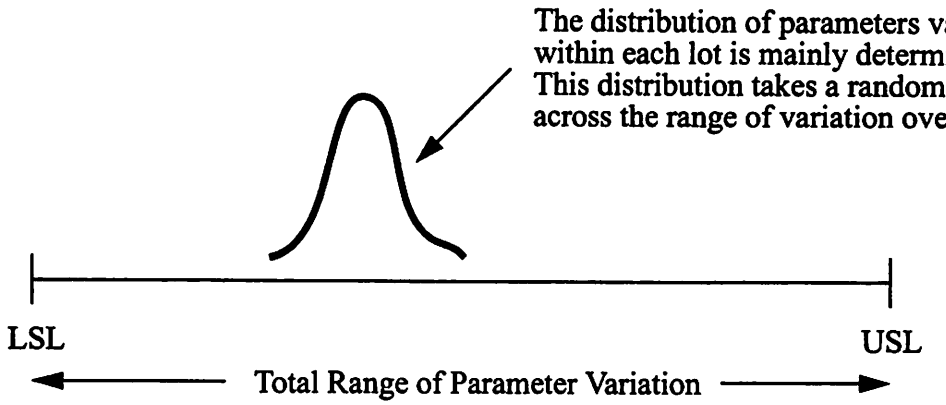


Figure 3-4 Process Parameter Distribution

there are components which need exact matching characteristics. Characterization of the mismatch between adjacent structures on a die is another example of a product specific parameter which may be considered. This work has not included the effect of mismatch, because constraints on the test pattern used in this project did not allow the addition of a structure to characterize mismatch in production. Good references on the subject include [24] and [25].

3.3.1 Parameter Distribution

Statistical simulation of fabrication lines is based on using random number generators to create instances of the process parameters drawn from their distributions. Previous work in statistical circuit design has commonly modeled process variables using a normal distribution. The model for the distribution of process parameters used in this work is shown in Figure 3-4. Within each lot, the distribution of the parameters can be approximated by a normal distribution. However, the cause of the distribution is not random error, rather, there are deterministic patterns caused by equipment variation [26]. For example, fixed patterns across the field of a stepper [27] may appear as normally distributed variation. This deterministic pattern takes a random walk across the allowed range of the pro-

cess, limited by the process specifications. Over time, the overall distribution of a parameter within reasonable control resembles a normally distributed variable, and this approximation allows an accurate calculation of the yield over long production cycles.

In this work, normal distributions are used for the process control application because an approximation to the actual overall distribution is required. However, because there will be periods of time where a manufacturing line is producing lots near the specification limits, uniform distributions are used when the goal is to capture and model performance dependencies over the entire process range. The range of the uniform distributions are equivalent to the statistical process control range.

3.3.2 Parameter Correlation

The set of process parameters used to model the fabrication line forms a multidimensional space bounded by the range seen in each of the parameters. The region can be further limited by taking into account the fact that the process parameters, as is shown in the left hand side of Figure 3-5, are usually dependent on each other. Limiting the range allows for better coverage of process space with fewer experiments. The various parameters measured in-line and at electrical test often show strong correlation to each other, reflecting the physical limitations of the process, such as the sharing of specific process steps. For example, the reductions in the channel length (critical dimension) of n- and p-channel transistors are highly correlated because they share the common processing steps which define polysilicon linewidth.

To generate instances for simulation of two correlated variables x_1 and x_2 with given means (μ) and standard deviations (σ), the following empirical formulas were derived:

$$x_1 = \sigma_{x1}\sqrt{\frac{1}{a^2+b^2}}(av+bw) + \mu_{x1} \tag{7}$$

$$x_2 = \sigma_{x2}\sqrt{\frac{1}{c^2+d^2}}(cv+dz) + \mu_{x2} \tag{8}$$

where v, w, and z are normalized random numbers, and a, b, c and d are determined by the correlation (ρ), such that:

$$\rho_{x1,x2} = \frac{ac}{\sqrt{(a^2+b^2)\,(c^2+d^2)}} \tag{9}$$

The values for $\rho_{x,y}$ are chosen based on our experience with the process and to be consistent with the objective to eliminate simulations for unlikely combinations of process parameter values.

3.4 Device Modeling Method

To simulate the effect of process variation on a circuit, the connection between the process parameters and the device model parameters must be established. Therefore, given the fabrication line description, the next step is to map the process into the device model parameters, creating transistor models for use in circuit simulation. One of the most important activities in manufacturability modeling is to efficiently extract device model parameters to cover the range of the process, without losing information in the mapping. This section will review approaches for creating the device models, and further develop the relationships stated in Section 3.2 which underlie this methodology.

3.4.1 Previous Approaches

Traditionally, circuit designers have used worst-case transistor models for simulation. Such models are optimized to fit specific process "corners," usually called "fast" and "slow" cases. One problem with such "worst-case" models is that they overestimate the range of the process, thereby increasing the design time, as designers create circuits which operate under pessimistic process conditions. In addition, a modern IC process often does not contain "corners," as there is no single combination of process parameters which results in the "slowest" transistor for every circuit topology.

In statistical circuit design, the process to device mapping is normally done through coupled process and device simulators, such as SUPREM [28], MINIMOS [29] or FABRICS [3], however, these methods are often cumbersome. As previously described, process simulators model variation using the physics of individual process steps, and the resulting complexity and tuning problems make it difficult to apply this method in commercial fabrication lines.

In both worst-case and statistical modeling, mapping from the process to the device is closely tied to device parameter extraction. Parameter extraction refers to the method used to find the values for the device model parameters. Given an I-V curve, which can be produced from either a process and device simulation or a worst-case process, mapping essentially implies extracting the parameters for the given model. Another problem with previous approaches is the use of optimization for parameter extraction. It is a common approach to allow all the device parameters to vary during the extraction. Although this achieves the best fit, all physical meaning can be lost.

Other prior work has used Principal Component Analysis[2] to perform this mapping [30]. Briefly, this strategy develops statistical models for the device model parameters in terms of the Principal Components of the process parameters. Principal Component techniques have been shown to increase the goodness of fit of the device model over the range of process variation. In fact, the methodology used in this work also uses statistical models for this purpose, although without the complexity of Principal Components. However, the use of Principal Components in the device model adds a level of difficulty in relating circuit simulation results to specific process variables [31].

2. Principal Component Analysis will be described in more detail in Chapter 4, when it is utilized in the development of the performance prediction model.

3.4.2 Physically Based Parameter Extraction

The approach taken in this work utilizes a physically based parameter extraction technique. This method, introduced in the work of Davis [13] and Tuinhout [14], bases the device model parameters on the measured process characteristics. In this method, the process characterization parameters and device model parameters have a straight-forward mapping to measurements, allowing for easy and efficient mapping of the process to the device model. Using this approach, a device model based on the Level 3 SPICE model was developed for use in the simulation phase of this work.

As stated in Section 3.2, the first key point in this methodology is that the overwhelming percentage of IC performance variation is explained by the variation in a small set of measurable parameters. Instead of using coupled process and device simulation, the methodology begins with the small set of measurable process parameters. These parameters, which characterize the range of variation seen in the process, are the basis of a statistical transistor model. Varying these parameters across the process space allows circuit simulation anywhere in the range of process variation. The physically based statistical transistor models alleviate these difficulties or process and device simulation, yet still allow circuit simulations to be run anywhere in the range of the process.

The second key to efficiently connecting the fabrication line and device models is to have a simple mapping among the important parameters. In this case, the measurable process parameters in the fabrication line model correspond directly to underlying device model parameters. For example, in MOS processes, the gate oxide thickness is both a measurable process characteristic and a specific device model parameter. This provides a direct link between the process and device models.

The process parameters map not only to the directly corresponding device model parameters, but also to other parameters through second order effects. For example, in an MOS device the effective mobility is a function of the gate oxide thickness and must be

modeled as such for accuracy over the entire range of oxide thickness variation. Prior approaches to the second order mapping have included both physical and statistical methods. The most powerful, but complex method is to use physical models for fitting parameters. Due to the complexity, simpler statistical techniques are used here to model key device parameters, such as mobility reduction, in terms of the measured process parameters [33]. It is advantageous to limit the complexity of the relationship between the physical and the device fitting parameters in obtaining a good fit, so that the communication between manufacturing and design organizations is not lost. Importantly, the physically correct process parameters are chosen as the variables for the statistical models.

3.5 Designed Monte Carlo Experiments using Circuit Simulation

Given a device model which describes the transistor characteristics at any point in the process, a designed experiment can be run to investigate circuit performance over the process space. Two dimensional examples of this experimental space are shown in Figure 3-5. The plot on the left hand side of Figure 3-5 describes the variation in the effective channel length of n and p channel transistors. The rectangle represents "worst-case" corners for these variables, ignoring their correlation. Each variable is described by a minimum and maximum value, and a simulation would be run at each corner of the rectangle.

The striped area inside the rectangle is the region which accounts for the correlation in the parameters[3]. The regions inside the striped area represent regions of increasing process control. Modern IC fabrication lines occasionally run out of statistical control. Therefore, the experiment must be run on region large enough so that the results are valid whenever the process is within specifications, even when the process is out of statistical control.

The plot on the right side of Figure 3-5 shows the performance of the manufactured circuit, where the two dimensions used are delay and power. Depending on the perfor-

3. In the specific case of effective channel length, the correlation is caused by the sharing of the polysilicon etch step.

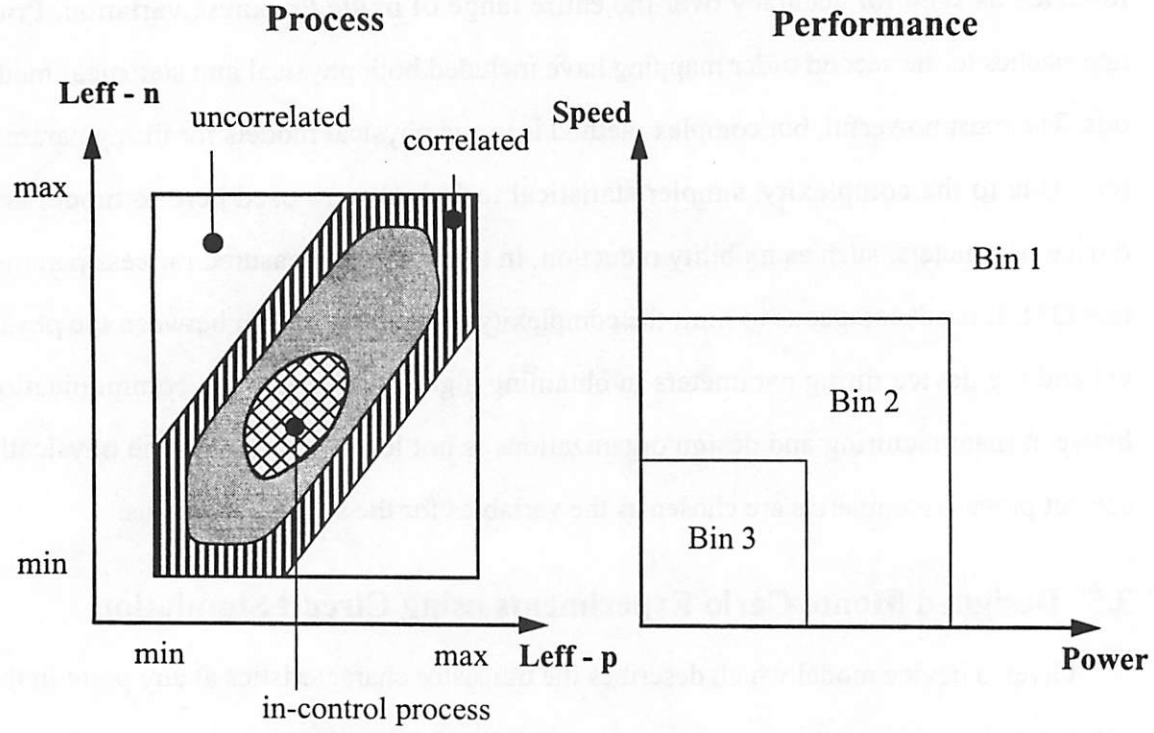


Figure 3-5 Mapping the Process Space into the Performance Space

mance of the circuit determined at final test, the manufacturer assigns it to the appropriate bin and prices the part accordingly. Although identical in design and manufacturing cost, the highest performance parts are sold at a premium.

The goal of the designed experiment is to map the process space into the performance space, that is, to determine what the corresponding performance will be for each point in the process. The experiment involves selecting a set of points in process space. For each point selected, a device model is created and a circuit simulation is run to find the circuit performance. However, unlike the simple two dimensional example of Figure 3-5, more parameters are required to generate a good description of a modern MOS process. Therefore, Monte Carlo techniques are used to cover the process space with a reasonable number of experiments.

Experimental designs such as orthogonal arrays and latin hypercubes have previously been used to reduce the number of simulations required to cover this space

[9][16][17][18]. Often, a screening experiment is utilized to limit the number of parameters considered in the experiment. These experimental design techniques are not utilized in this work because, given the parameter correlation in the fabrication line description, the process space is no longer orthogonal.

The Monte Carlo technique requires a large number of circuit simulations to converge on accurate results for the experiment. Previous work has used macromodeling techniques which develop regression models for performance from a limited number of circuit simulations and then use the regression models for the larger number of simulations needed to complete the mapping [2][34]. Although such techniques which optimize the Design for Manufacturability (DFM) process are important, they were not considered in this work. Creating accurate macromodels saves simulation time, but improves neither the robustness of the DFM process nor the integration with the manufacturing line, which remain the focus of this methodology.

3.6 Summary

This chapter developed the theory underlying manufacturability modeling. Specifically, the techniques for mapping the manufacturing process to the final circuit characteristics were explored. Together, a fabrication line model, device models, and circuit simulation enable IC DFM. Most importantly, a physically based parameter extraction methodology for device modeling eliminates the difficulties of worst-case and simulation based methods, and enables the use of more powerful DFM applications [9][32].

So far, model verification has mainly been between adjacent levels of the modeling hierarchy. For example, the device model has been verified in that the simulated I-V curve matches the curve which produced the model. Chapter 4 will focus on the applications of the manufacturability models. Additional cross-checks at the process, device and circuit levels verify and integrate the models, closing the loop between design and manufacturing.

Chapter 4

Design For Manufacturability Applications

Chapter 3 discussed the development of a fabrication line description and device models which give the circuit designer the ability to simulate a design across the range of the process. Their purpose is to build applications for use in manufacturing and circuit design. The particular manufacturing applications discussed here are driven by response surface models which predict product performance before assembly. For the circuit designer, the models presented in Chapter 3 enable advanced Design for Manufacturability (DFM) techniques which improve the robustness of circuit performance in light of IC manufacturing line variation. Importantly, these applications cross-check each other. Manufacturing line measurements verify the design models, improving confidence in the model predictions, and hence the design improvements, before the improved design is fabricated. This chapter discusses the development of applications and the verification strategy, as well as the CIM infrastructure necessary to implement the method.

Previous DFM work has mainly been focused on optimizing the circuit design of an IC product to improve yield. However, it is important to include manufacturing data and applications in DFM work to validate the models and to effectively bring the manufactur-

ing organization into the DFM process. The first manufacturing application developed in this work is a performance prediction model which predicts the performance of the fabricated parts from in-line and electrical test measurements. The use of in-line measurements will allow estimation of product performance early in the fabrication process, improving manufacturing planning. Electrical test data will provide a more accurate performance prediction, allowing fast parts to be directed into the highest performance packages.

The second application developed in this chapter utilizes the performance prediction model as the basis of a statistical process control (SPC) procedure for the manufacturing line. In today's IC factory, SPC charts are maintained on low level parameters for specific pieces of equipment, such as etch rate, deposition rate, uniformity and particle counts. The performance prediction model raises the opportunity for a high level monitor on the entire fabrication line. The setup and use of a control chart for product performance will be discussed.

The third and final application developed in Chapter 4 utilizes Monte Carlo techniques which take advantage of the physically based device models to investigate the performance of the circuit over the range of variation seen on the manufacturing line. This stands in contrast to traditional worst-case design, where the designer only has models to simulate at "fast" and "slow" corners of p- and n-channel device saturation currents. Sensitivities of critical subcircuits to process variation will be identified, and design changes will be tested to ensure the yield limiting sensitivities are decreased.

4.1 Statistical Methods

Before the first application is discussed, two statistical methods used in developing the applications, Principal Component Analysis (PCA) and linear regression, will be briefly reviewed, and the advantages of combining PCA and linear regression are presented.

4.1.1 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used to describe the relationships within a set of variables. PCA accomplishes this by transforming the data set into new, orthogonal variables. The new variables, called *principal components*, are linear combinations of the original variables, each explaining the largest possible variation in the original data [35]. In addition to producing uncorrelated variables, PCA can be used to reduce the dimensionality of a problem without losing much information. Since typically a small set of principal components explains most of the variance in a data set, the complexity of the problem can be reduced by choosing only the first few principal components for analysis. The "rule of thumb" for including principal components in further analysis is to retain just enough principal components to explain 80 or 90% of the variance in the original dataset.

The principal component *scores*, also called simply the principal components, are given by:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{G}^{T} \tag{10}$$

where $\boldsymbol{X}$ $(n \times p)$ is the matrix containing n measurements of the p process parameters, such as oxide thickness, $\boldsymbol{G}$ is the $(p \times p)$ matrix of principal component *loads* and $\boldsymbol{Z}$ is the $(n \times p)$ matrix of the principal components. The principal component loads are the eigenvectors of the correlation matrix of the original data, $\boldsymbol{X}$.

4.1.2 Linear Regression

In contrast to PCA, which describes the relationship within a set of variables, linear regression is a technique to build a relationship between independent variables and one or more dependent variables. In this case, the relationship will be constructed between the process parameters, $\boldsymbol{X}$ $(n \times p)$, and the product performance, $\mathbf{y}$ $(n \times 1)$. Linear regression postulates a model of the form:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{11}$$

where $\underset{\sim}{\beta}$ is the $(p \times 1)$ vector[1] of model coefficients and $\underset{\sim}{\varepsilon}$ is the value by which each observation deviates from the model. A column of 1's is often included in the **X** matrix to produce a constant term in the model.

In statistical terminology, a linear regression implies that the model is linear in the coefficients being estimated, that is, linear in $\underset{\sim}{\beta}$. The models built in this work are linear in this sense. Another common use of the term "linear" is to describe whether the model is linear in the variables. Better referred to as the order of the model, this work does not include non-linear terms such as quadratic or interaction terms in the matrix of independent variables. This is not a limitation of the method, as significant results were achieved using linear models. The use of nonlinear terms in the performance prediction model will be discussed further in Chapter 6.

Using the method of least squares [37], which estimates the model coefficients by minimizing the sum of squares of the errors in the model, the standard normal equations calculate the estimates for the coefficients in the model, $\hat{\underset{\sim}{\beta}}$ as:

$$\hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T y \tag{12}$$

and the variance-covariance matrix for the estimates of the model coefficients is given by:

$$V\left(\hat{\underset{\sim}{\beta}}\right) = \sigma^2 (X^T X)^{-1} \tag{13}$$

where σ^2 is the variance of the residuals.

4.1.3 Principal Component Regression

Instead of performing linear regression on the original process parameter data, **X**, the regression can use the principal components of the data, **Z**, that is, the transformed data can be used as the independent variable in the model. In this case, the principal compo-

1. Vectors assigned a Greek letters will have be denoted by a '~' underneath the character rather than a bold character.

nents are used as inputs to build a model for circuit performance. The form of the performance prediction model built from the principal components is simply:

$$y = \underset{\sim}{\beta} Z + \underset{\sim}{\varepsilon} \tag{14}$$

and the standard normal equations for linear regression are used, utilizing the principal components, **Z**, of the original data set. The use of the principal components of a data set for linear regression is commonly termed *principal component regression* (PCR).

There are several important advantages of PCR over linear regression on the original data. First, the presence of correlation creates difficulties in using least squares regression [37]. As was previously explained, the process parameter data is highly correlated. Correlation between the independent variables increases the variance in the estimates of the regression model coefficients, increasing the difficulty in finding a statistically significant model. It can be shown that the variance of $\underset{\sim}{\hat{\beta}}$ decreases as the eigenvalues of $(X^TX)^{-1}$ increase, that is, as the columns of **X** become mutually orthogonal [38].

It is insightful to look at the case where there are two independent variables, in which case:

$$V\left(\hat{\beta}_i\right) = \left(\frac{1}{1-\rho^2}\right)\left(\frac{\sigma^2}{\sum_{j=1}^{n} x_{ij}}\right) \quad i=1,2 \tag{15}$$

where *i* will be 1 or 2 to signify one of the two model coefficients, *n* is the number of observations and ρ is the correlation coefficient between the two variables [39]. This example shows how the variance of the model coefficients decreases as the correlation decreases. Using the uncorrelated principal components instead of the original, highly correlated data, increases the likelihood of finding a significant model.

The second advantage arises from the fact that a few principal components explain most of the variance in the data. Therefore, a linear regression using the principal compo-

nents as the independent variables will likely require fewer variables in the model than if the original data is used. The prediction error of the fitted model increases with p, the number of independent variables in the model, as given by:

$$E\left\|y - X\hat{\beta}\right\|^2 = \left\|E(y) - E\left(X\hat{\beta}\right)\right\|^2 + p\sigma^2 \tag{16}$$

where E denotes the expectation operator and $\|v\|$ denotes the norm of vector v [38]. This shows that when the model is used for prediction, the expected value of the residuals, $E\left\|y - X\hat{\beta}\right\|^2$, tends to increase as the number of variables in the model increases, unless adding the variable brings a significant reduction in the "bias" of the model.

4.2 Manufacturing Application - IC Performance-Binning

A detailed view of performance model building is shown in Figure 4-1, which expands the overview shown in Figure 3-2. This section details the development of the performance prediction model, discussing the Monte Carlo simulation, Principal Component Analysis, linear regression and model verification.

4.2.1 Monte Carlo Simulation

Monte Carlo simulation is used to drive a computer-based experiment for building performance prediction models. The process description begins with the identification of the critical model parameters. The mean value, spread and correlation of the parameters are found directly from measurements made on the process. The advantage of the Monte Carlo technique is that unlike experimental design techniques such as factorial or Latin Hypercube, the accuracy of the Monte Carlo convergence on an estimate of response is independent of the dimensionality of the problem, depending only on the number of trials run [36]. This is of great value due to the high dimensionality of the fabrication line model.

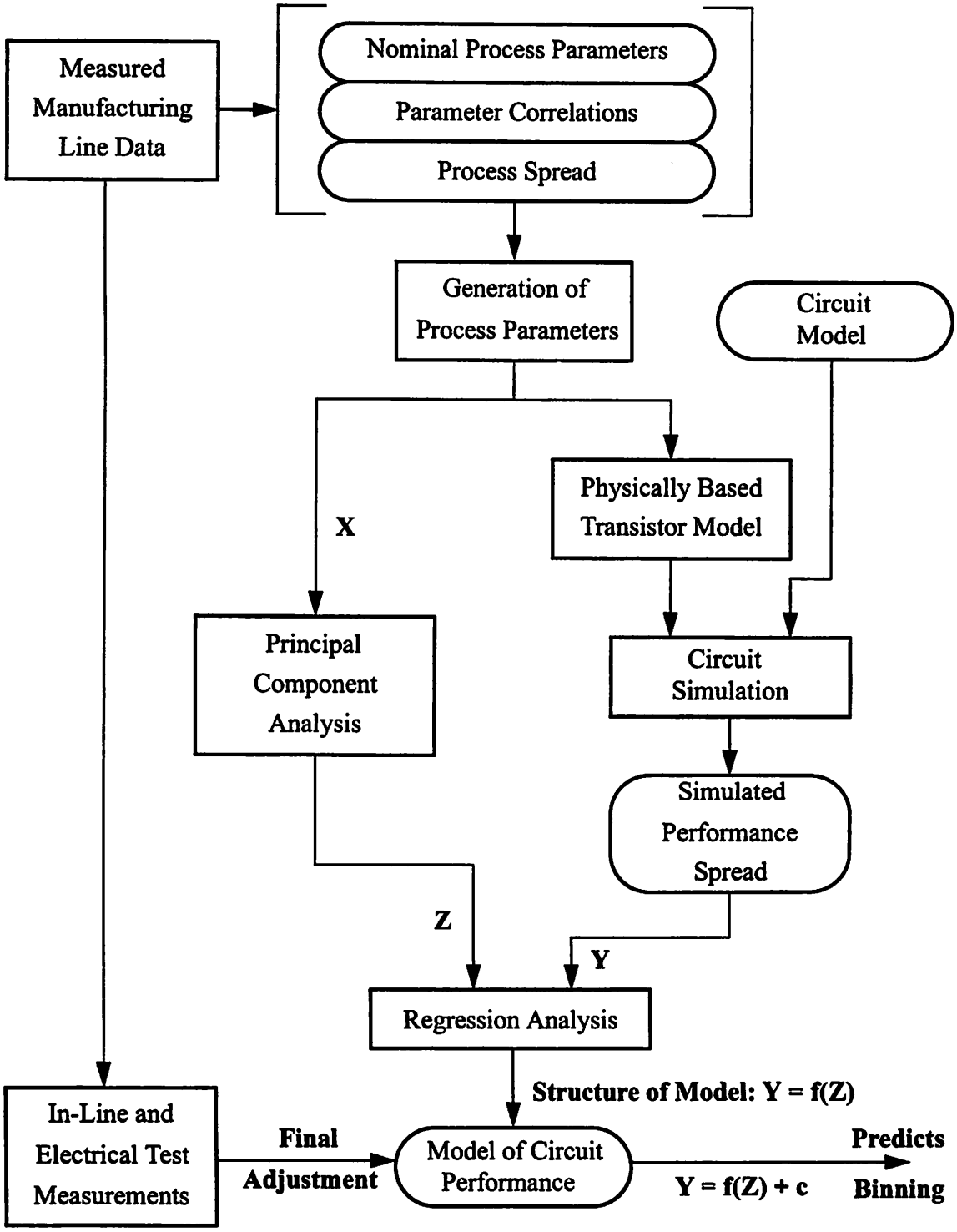


Figure 4-1 Detailed View of Performance Model Development

A Monte Carlo simulation is run which covers the maximum operational range of process variation. The process parameters are mapped into transistor models, as was explained in Section 3.4, and then circuit simulations are run to predict circuit performance over this range of variation.

4.2.2 Principal Component Transformation

Using the generated process parameters as the independent variables and the access time predicted by simulation as the dependent performance variable, a linear regression model can be built to predict the circuit performance from the measured parameters. PCA was used to transform the data set in this application to produce important advantages. First, PCA produces orthogonal variables for regression from the original, highly correlated dataset. Second, PCA reduces the dimensionality of the problem, as a few principal components will likely be sufficient to capture most of the variation in the original dataset, and reduces the prediction error of the fitted model. Together, these advantages will decrease the complexity of the linear regression and increase the likelihood of finding a significant regression model.

4.2.3 Linear Regression on Simulation Results

The principal components are used as inputs to the linear regression to build a model for circuit performance. The issue of which principal components are used in the regression equation will be discussed in more detail in Chapter 6, when a regression model is built from production data.

4.2.4 Manufacturing Verification and Final Adjustment

When a statistical model is built from a set of data, it is important to verify the model using data that was not used to generate the model. In this case, the regression model built from simulation data will be tested and verified using data from the manufacturing line.

This not only verifies the model itself, but forces agreement between the simulation models and manufacturing data.

The inputs to the performance prediction model are the measurable process parameters, such as oxide thickness and effective channel length for an MOS process. These parameters and the performance of fabricated circuits can be measured on manufactured products. The measured process parameters are used in the model to generate a performance prediction, which can be compared to the measured performance for that circuit.

Due to inaccuracies in the device model, circuit model and measurements, there may be a difference between the predicted and measured performance for the verification data. Therefore, the manufacturing data is used to estimate and correct the nominal performance value of the model by adjusting the constant term in the model. When the performance model prediction is in statistical agreement with the final test results, the process, device and performance binning models are all verified to be correct, closing the loop between design and manufacturing.

4.3 Statistical Process Control for Product Performance

The next application is a control procedure based on the performance prediction model. A control chart for product performance would add high level monitoring capability by signaling an alarm if the process was drifting away from the predictions of the model. The control chart would be a first step towards reducing the variability associated with the manufacturing process, as the alarms generated could be used to invoke high level process control algorithms to correct process drifts. Since this chart will be driven by prediction data, necessary corrections will be signalled early. It would also allow the calculation of the capability of the entire manufacturing line.

Current work in the industry focuses on equipment level control of lithography, thin film deposition and removal, *etc.* [40][41]. New research in factory level control will examine propagating specifications up and down the factory, as shown in Figure 4-2. This

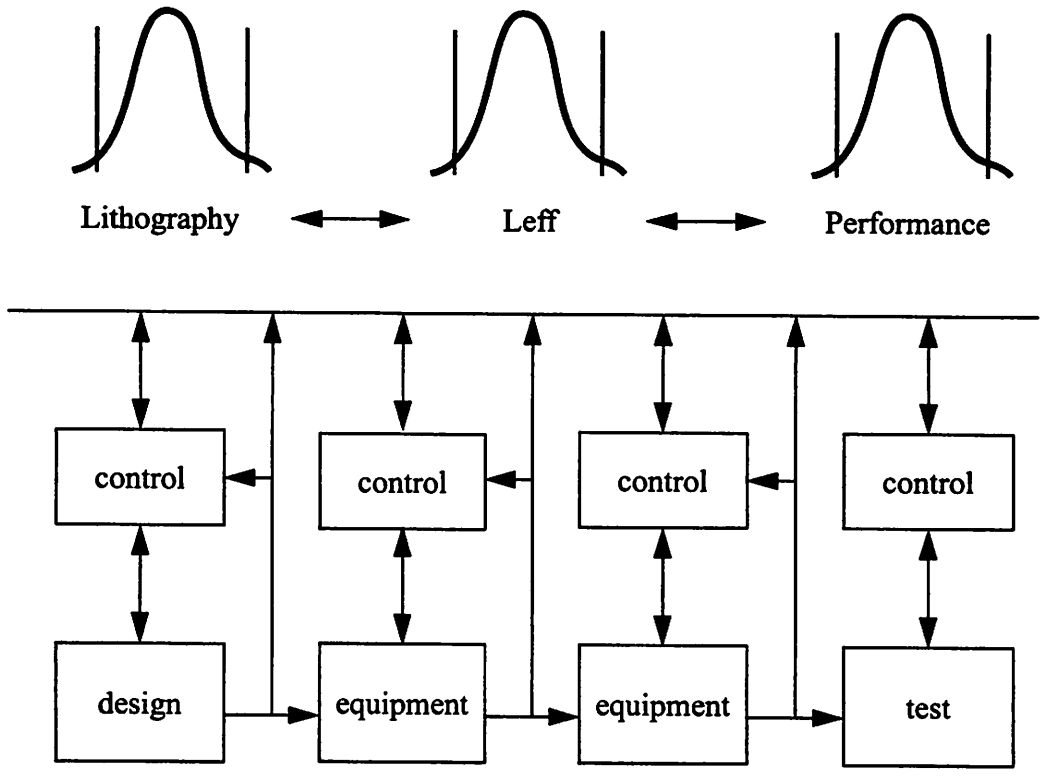


Figure 4-2 Process Control and Manufacturability

propagation must be based on the relationships between equipment, process parameters and product performance. The performance prediction model is a vehicle for propagating specifications up into the factory based on the most important metric - final product performance. The control chart is a step towards factory level control, as it generates an alarm which initiates corrective action.

There are two types of control procedures which could be used to monitor product performance. The first, utilizing the performance prediction model, compares the actual performance distribution of each lot with the model prediction. This control chart would signal an alarm when there is a shift in the characteristics of the manufacturing line or when a lot has been incorrectly tested. The control procedure would not raise an alarm

with the production of fast or slow parts, rather, when the in-line and electrical test measurements no longer accurately predict the performance of the fabricated product.

The second control procedure would assume an overall distribution of output performance for a given product and signal an alarm when actual performance falls outside of this desired distribution. The control limits could be set from the distribution, as in an $\bar{x}$ chart, or if the capability of the process is high, the performance specifications could be used, as in an acceptance chart to control fraction non-conforming [42]. Corrective action taken in response to the alarms generated by this control chart would move product performance back towards the target.

This section will first present some background in the distribution and testing of product performance, before discussing the details of these two control procedures.

4.3.1 Performance Distribution

The distribution of product performance must be specified to establish statistical process control strategies. Assuming a stationary, normal performance distribution for a manufacturing line is somewhat incorrect, as was the assumption of a normal distribution of the process parameters. The main reason, as discussed in Chapter 3, is that performance variance is not caused by random error, rather, our experience in this project showed that the variance of performance within each lot was mostly caused by deterministic spatial patterns in the manufacturing equipment [26].

However, the control procedures detailed in this work will generate control limits assuming a normal distribution of performance at both the lot and product levels. The assumption of normality greatly simplified the setup of the control chart and still allowed good results to be achieved. Future work on control charts for performance will want to revisit and improve upon this assumption.

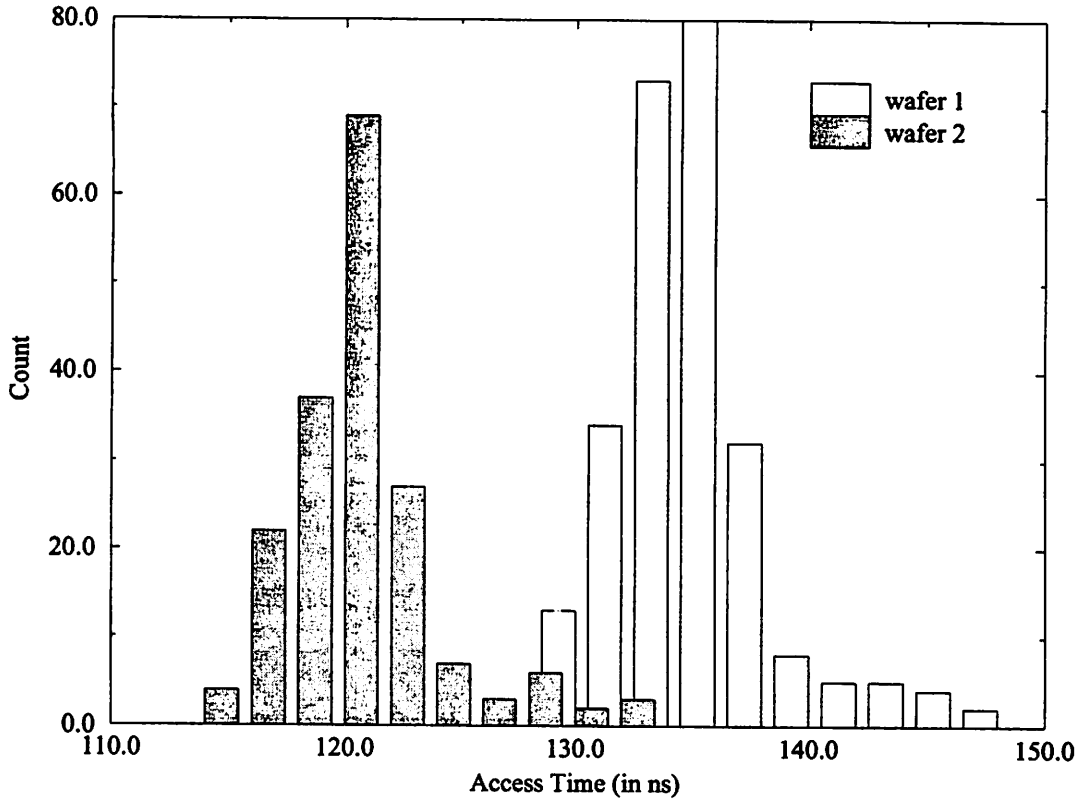


Figure 4-3 Detailed Histogram of Product Performance for Two Wafers

Histograms for two wafers with different average performance are shown in Figure 4-3. The performance specification shown is the access time, or speed of a memory IC. The shape and range of the histograms are similar, despite a 15% difference in average access time, because of deterministic equipment variation. A normal distribution is a good approximation for the shape of each histogram. Importantly, the determinism in the spread implies that the accurate prediction of the mean value of performance of the lot allows the number of parts in each bin to be calculated.

The setup of the control limits will utilize the normal distribution, which defines the probability of a variable x being less than or equal to a as:

$$P\{x \le a\} = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 dx \equiv \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad (17)$$

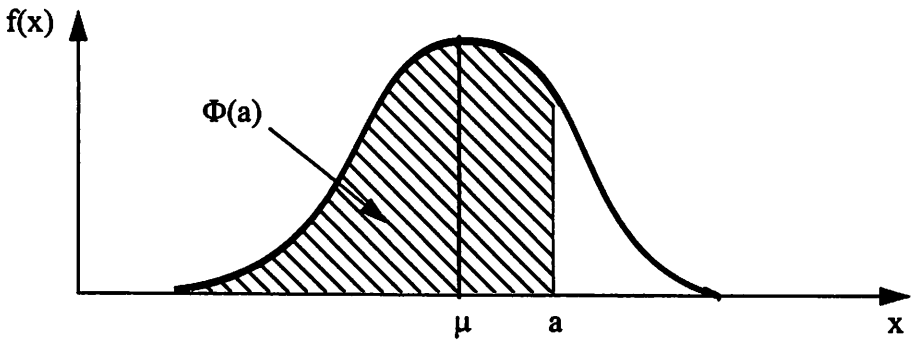


Figure 4-4 Calculating Probabilities using the Normal Distribution

where x has a mean μ and standard deviation σ. This is illustrated graphically in Figure 4-4.

4.3.2 Impact of Product Binning

Control procedures for product performance must account for the fact that actual performance is not recorded exactly at final test, rather, parts are usually sorted into bins which contain a specific range of performance. For example, consumers can purchase only 33 or 50 MHz microprocessors, and this greatly simplifies the testing and marketing of the part. Therefore, the fraction of parts in a lot sorted into each bin will be utilized in the control procedure to determine whether the process is out of control. Unfortunately, recording only the fraction in each bin rather than the absolute performance of each part, makes it impossible to calculate the variance of the lot and therefore a control chart on the range of performance in each lot cannot be maintained.

For example, suppose that the product shown in Figure 4-3 is sold with 120 ns[2] and 150 ns specifications on access time performance. In this case, all the parts from wafer 1 and approximately half of the parts from wafer 2 should be sold as 150 ns parts, with the other half of wafer 2 sold as 120 ns parts. The control procedures developed in this section

2. The symbol ns stands for nanoseconds, or billionths of a second, the common time unit in digital electronics.

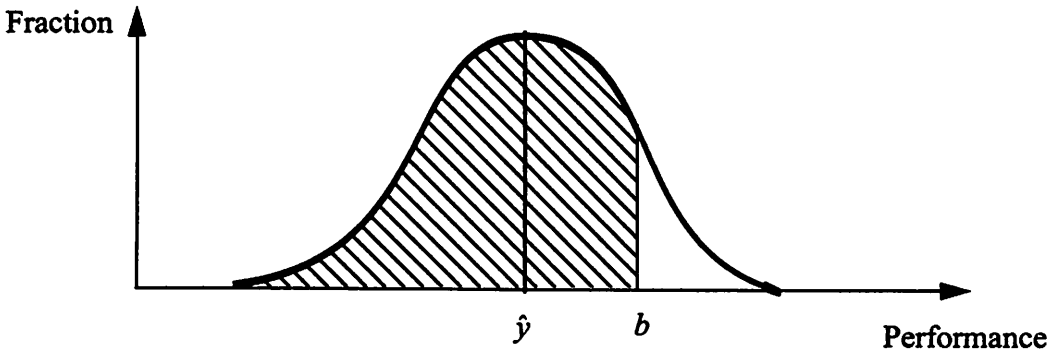
will assure that the fractions from actual testing agree with either the fractions predicted by the performance prediction model or the overall distribution of product performance.

4.3.3 Setup of a Model Based SPC Procedure for Product Performance

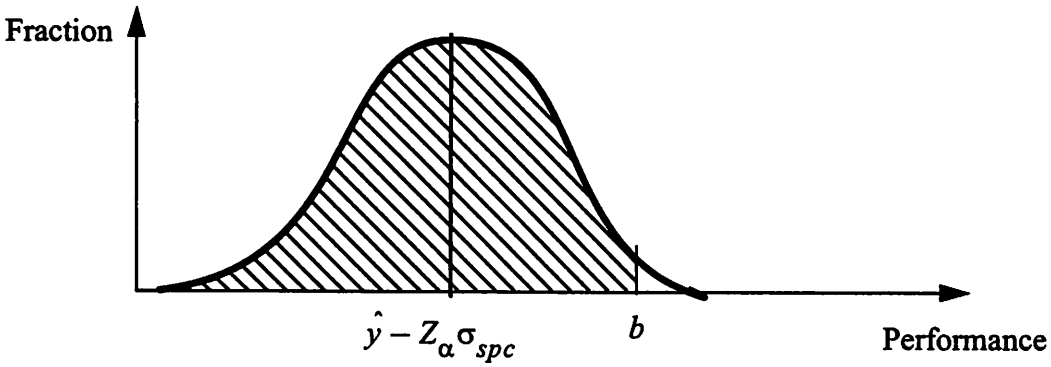
A model based control procedure for product performance would compare the fraction of parts in a lot which fell into each performance bin against the fraction predicted by the performance prediction model. It would detect process shifts which cause actual performance to deviate from the prediction of the model, including incorrectly tested lots. To generate a prediction of the bin fractions from the performance prediction model, the electrical test results for each manufactured lot are used to predict the average performance of the lot. Given this mean value, the fraction of parts which should occur within each bin can be calculated by assuming a known variance for the performance of the lot.

The control limits will be calculated for each lot, because the expected fraction in each bin will depend on the distance between the mean performance for the lot and the edge of a bin. The limits are set by shifting the mean value for the lot and calculating the number of parts in each bin, as shown in Figure 4-4. The control procedure will issue an alarm if the actual fraction of parts in any bin is outside of the control limits. To calculate the limits, both the error of the model prediction and the Type I error must be accounted for. The Type I error, or false alarm rate, is the probability of generating an alarm even though the process is in control. A Type I error rate of α is obtained from a control limit shifted by $Z_\alpha\sigma$ from the mean, where Z_α is the upper α percentage point of the standard normal.

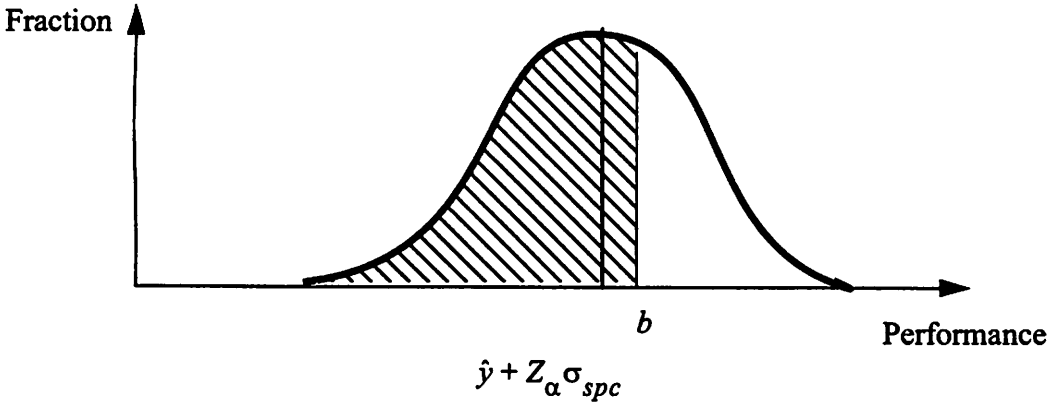
To account for model prediction error, the mean is shifted relative to the standard deviation of regression of the performance prediction model, σ_{model}, and to ensure a low Type I error, the mean is shifted relative to the standard deviation of the performance of the lot, σ_{lot}. Assuming these sources of error are independent, a standard error for the control procedure, σ_{spc}, can be calculated as:



Predicted performance implies nominal fraction of parts in each bin.



Calculation of the upper control limit for lower bin.



Calculation of the lower control limit for lower bin.

Figure 4-5 Calculating Control Limits for the Fraction of Parts in a Bin for a Model Based Control Chart - The shaded region in each of the lower two plots represents one limit to the expected fraction.

$$\sigma_{spc} = \sqrt{\sigma^2_{model} + \sigma^2_{lot}} \tag{18}$$

In the case of two bins separated by a performance specification b, the control limits are given by:

$$LCL = \Phi\left(\frac{b + (\hat{y} + Z_\alpha \sigma_{spc})}{\sigma_{lot}}\right) \tag{19}$$

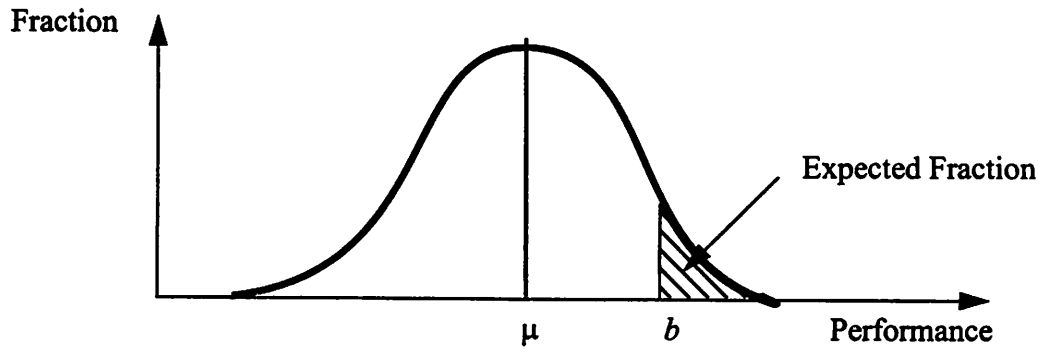
$$UCL = \Phi\left(\frac{b + (\hat{y} - Z_\alpha \sigma_{spc})}{\sigma_{lot}}\right) \tag{20}$$

where $\hat{y}$ is the performance predicted by the model and Z_α is set to give an α-level sensitivity and false alarm rate.

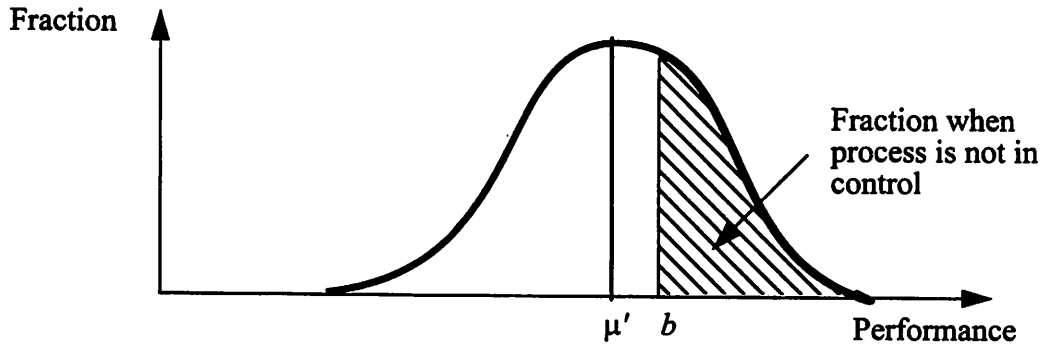
In the case where the performance distribution covers two bins, the control limits on the second bin are simply the complementary fractions from the first bin. It is a straight forward extension to calculate the limits for the case where the performance distribution spans more than two bins. However, some care must be taken in calculating the limits for the bin containing the peak of the distribution. The limits for that bin are best obtained by subtracting the minimum (or maximum) amounts from the other bins from the total number or fraction.

4.3.4 Setup of a Control Procedure for Overall Product Performance

The control procedure for overall product performance will signal an alarm if the process is drifting to produce mostly low or high performance products. The model for this chart is a fab producing a product whose performance is normally distributed with mean μ, and total variance $\sigma^2_{product}$. This procedure will count the number of parts in the fast and slow bins, generating an alarm when the fraction is over the control limit. This control chart has a fixed control limit for each bin. For example, with a bin specifications of b, as



Distribution when the process is in control



Shifted distribution when the process is not in control

Figure 4-6 Control Procedure for Overall Product Performance

shown in Figure 4-6, the fraction of parts which should be found in the highest performance bin should be less than:

$$CL = 1 - \Phi\left(\frac{b - \mu + Z_\alpha \sigma_{model}}{\sigma_{product}}\right) \tag{21}$$

where, again, Z_α is set to give an α-level sensitivity and false alarm rate.

Several lots should be averaged for each point on the control chart to maintain a low false alarm rate. A single lot may contain a large fraction in the fast bin even with the process in control. Western Electric rules should be applied to this control chart to catch trends or drifts in product performance.

This control chart has a Type II error of:

$$\beta = \Phi\left(\frac{b-\mu'}{\sigma_{product}}\right) \tag{22}$$

where β is the probability of not detecting the shift on the first sample and μ' is the shifted mean. For example, the probability of not detecting a shift in the mean from μ to b is 0.5. The probability of detecting the shift decreases as the size of the shift in the mean to detect decreases.

4.3.5 Process Capability

Process capability is an increasingly important metric for establishing how suitable a manufacturing process is for production. Process capability, C_p, is defined as:

$$C_p = \frac{USL-LSL}{6\sigma} \tag{23}$$

where USL and LSL are the upper and lower specification limits and σ is the standard deviation of the process. Current industry practice is to calculate capability at the level of a piece of equipment or process step. However, there would be a great advantage to calculating the capability of the overall process. For example, after making a change in the process flow or a piece of equipment, it could be verified that the capability of the process had actually improved.

The control limits and $\sigma_{product}$ used in the control chart for overall product performance could be used for the USL, LSL and σ in a calculation of C_p for a process. If binning is used in normal production performance testing, occasional detailed testing to estimate $\sigma_{product}$ would be required. Motorola [9] and Philips [10] have implemented a C_p metric for performance using this concept in their DFM tool algorithms.

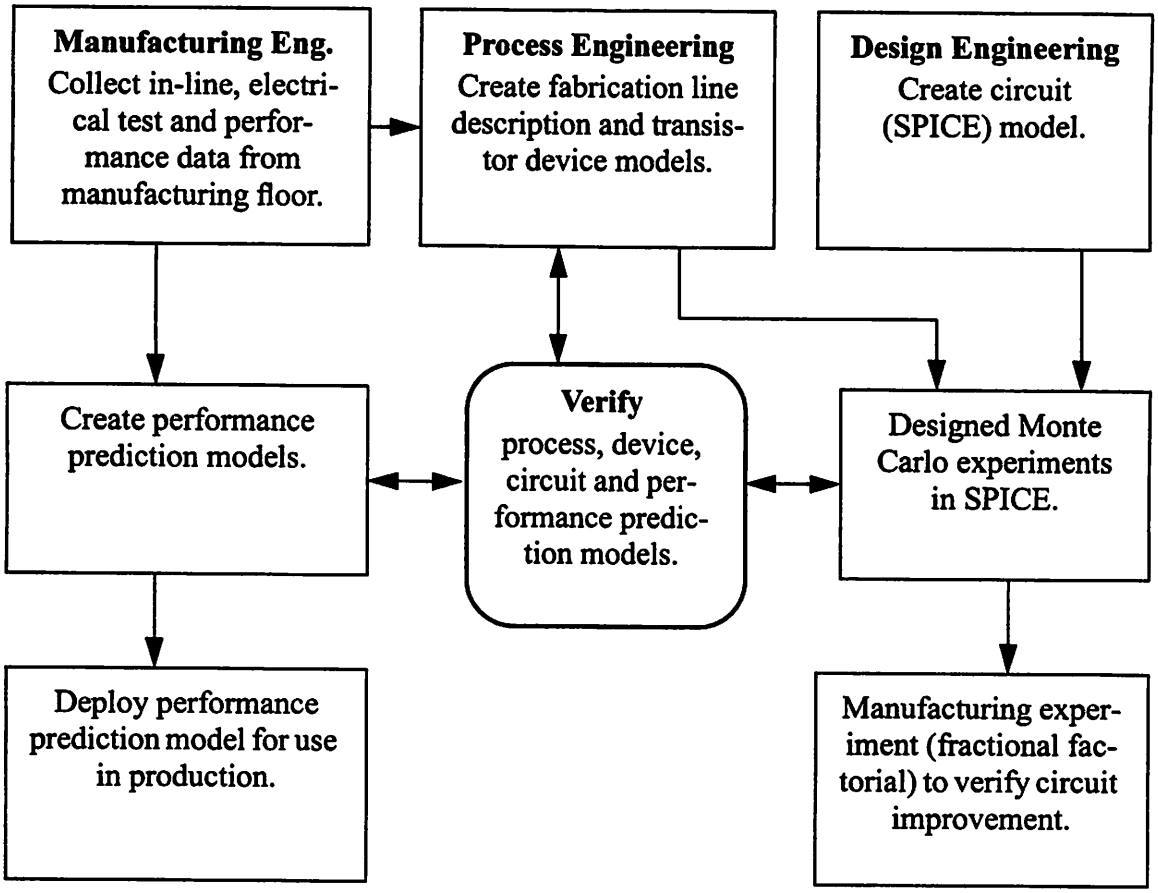


Figure 4-7 DFM Methodology

4.4 Circuit Manufacturability Improvement Method

This section will describe the activities outlined in Figure 4-7, which shows an overview of the DFM methodology as applied to the improvement in the manufacturability of an IC product. Remember, the goal of DFM is to improve a circuit design to decrease its sensitivity to manufacturing variation. Note the parallel, coordinated activities in the manufacturing, process and design engineering sectors. Figure 4-7 shows the methodology applied in the situation where both a new process and a new circuit design are being developed. If an existing product design or process is used, some of the necessary data and models have already been generated.

4.4.1 Fabrication Line Description and Device Modeling

The process engineering organization creates the fabrication line description and device models. The objective is to support computer-based experiments which relate process parameters to product performance. These experiments will be used to identify the sensitivity of IC designs to process parameter variation. The fabrication line description and the device modeling methodology have been discussed in Chapter 3.

4.4.2 Circuit Model

The circuit design engineer generates the model of the circuit for use in simulation. This model contains the transistors and any other devices comprising the circuit, as well as the connectivity. In addition, the circuit model must include descriptions of the parasitic loading in the circuit and any special structures in the product. The descriptions for these structures must include models which reflect their parametric variation over the process range. For example, in a memory circuit, the models of the word and bit lines must include the changes in their resistive and capacitive load caused by variation in linewidth and oxide thickness.

4.4.3 Manufacturing Data Collection

Simple integration with manufacturing is the key to this DFM methodology. Manufacturing data forms the basis for the fabrication line description. Electrical and final test data allow the cross-checks which verify all the models and the simulation results. The manufacturing engineering organization collects in-line and electrical test data from the manufacturing floor and performance data after packaging and final test. This data contains the mean value and range of the key process parameters which describe the fabrication line, measured transistor currents to validate the device models, and performance data which validates both the performance prediction models and the circuit sensitivities.

4.4.4 Verification

The verification of the models and circuit improvements occurs at the device, circuit and product levels. At the device level, the Monte Carlo simulation results are used to compare the mean value, range and sensitivity of transistor currents with the manufacturing data. At the circuit level, sensitivities to the variation of a specific process parameter should appear in both simulated and manufacturing data. Finally, manufacturing data are used in the performance prediction model built from simulation to verify the correctness at the product level [43].

Note that the device model is verified by checking the range and sensitivity rather than distributions or yield. For example, when examining circuit sensitivity, plots of performance versus a process parameter for manufacturing and simulation results should show the same slope rather than the same correlation coefficient. In a process commonly referred to as "tuning," previous work in the area of statistical circuit design devoted time and energy to adjusting the mean and variance of the input process parameters to match the distributions of the output parameters from simulation and manufacturing [44]. Although tuning matches a snapshot of the process, a modern IC process is constantly undergoing evolutionary changes making the snapshot obsolete. Further, matching output distributions does not guarantee that the mapping between the process and performance spaces is correct. By looking at sensitivities across a broad range of the process using physically based device models, instead of adjusting input parameter distributions, the correct relationships between the process parameters, the device characteristics and the circuit performance must be pursued.

When a design improvement is made to increase circuit manufacturability, an experiment should be run to verify the results before putting the new design into production. Typically, a fractional factorial experiment [39] will be used, where a few of the critical process parameters are varied at two or three levels. A fractional factorial is the appropri-

ate experimental design in this case because it covers enough of the process space to verify the improvement, yet is small enough to be run using a single lot split, reducing the cost of the verification run.

4.5 CIM Infrastructure for DFM

A factory-wide Computer Integrated Manufacturing (CIM) infrastructure is critical to the success of this DFM methodology [11]. An overview of such a system is shown in Figure 4-8. The CIM system[3] provides the in-line data, such as linewidth and oxide thickness for a CMOS process. Similarly, the electrical test database provides the electrical test data, such as device currents and threshold voltages. The performance data comes from the results of final test. In the most advanced factory-wide CIM systems, the in-line, electrical test and final test data are all stored in a single relational database. DFM software provides an interface with the circuit designer to run the Monte Carlo simulations. To support the statistical characterization of the manufacturing data, the building of response surface models and the analysis of the Monte Carlo simulation, statistical analysis software is required.

This work used the Workstream [45] CIM System and the RS/1 [46] statistical analysis package. Both the in-line and electrical test databases were managed by custom software, including the interface into RS/1 for analysis [47][48]. The software to run the DFM simulations was automated through custom software written by the author. The HSPICE [49] circuit simulation software was used for circuit simulation.

4.6 Summary

This chapter presented the theory behind the DFM applications. Three applications were presented, including a performance prediction model, a control chart for product per-

3. Specifically, the CIM system is responsible for running the factory, including scheduling the equipment, recipe control, in-line metrology management, etc. In general, the concept of a CIM system is extended to include the functions shown in Figure 4-8.

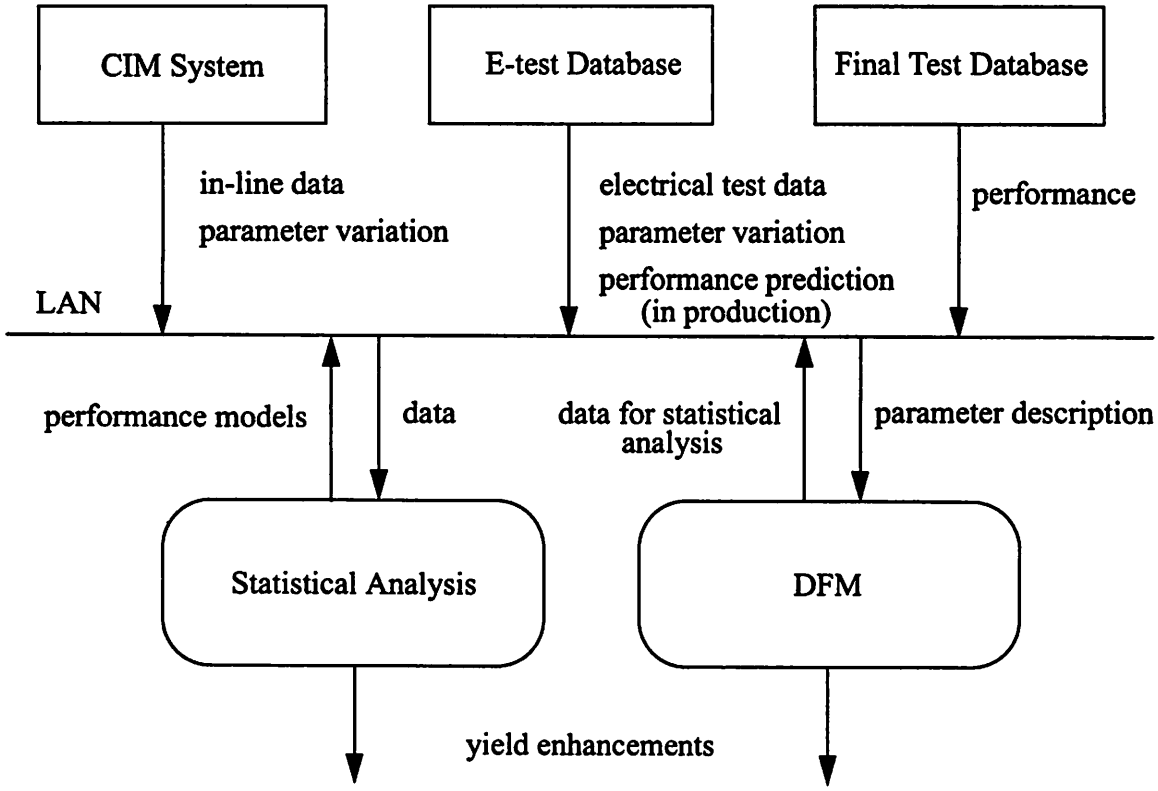


Figure 4-8 CIM Infrastructure for DFM

formance, and IC DFM. These applications were built on a common set of models which described the fabrication line and the devices. Most importantly, all three applications strive to integrate the process, manufacturing and design activities in the company.

Chapter 4 concludes the theory behind manufacturability modeling. Next, in Chapter 5, process and device models will be built for a production 1.2μm CMOS process, and then, Chapter 6 will present the results from developing these applications for the manufacturing line.

Chapter 5

Fabrication Line and Device Models for a Production CMOS Process

This chapter applies the methodology, developed in Chapter 3, for building a fabrication line description and extracting transistor device model parameters for circuit simulation to a production 1.2 μm CMOS EPROM process. This chapter first develops the fabrication line description, focusing on the set of process parameters which describes the spread of performances seen on the fabrication line. Next, the process characterization techniques are summarized, as a key element of this methodology is that all of the parameters are measurable using production line capable techniques. The chapter then presents the device model parameter extraction for a SPICE Level 3 MOS transistor model. Finally, the product specific EPROM cell model is discussed.

5.1 Symbols

The symbols used in this chapter are listed in Table 5-1, except for the SPICE model parameters, which are listed in Table 5-4.

Table 5-1 List of Symbols Used in Chapter 5

Symbol	Description	Units
C_{gd}	gate to drain capacitance	F/cm^2
C_{gs}	gate to source capacitance	F/cm^2
C_{poly}	EPROM cell inter-poly capacitance	F/cm^2
C_{ox}	channel capacitance	F/cm^2
I_D	drain current	A
L^a	channel length	cm
Nsub	surface doping level	cm^{-3}
q	electronic charge	C
T_{ox}	gate oxide thickness	cm
T_{poly}	EPROM cell inter-poly oxide thickness	cm
V_D	drain voltage	V
V_G	gate voltage	V
V_{SB}	source to body voltage	V
V_T	threshold voltage	V
V_{T0}	threshold voltage without body bias	V
W^b	channel width	cm
ε_{ox}	oxide permittivity	F/cm
ε_{si}	silicon permittivity	F/cm
γ	body effect parameter	$V^{1/2}$
ψ_s	surface potential	V
μ_{eff}	effective mobility	$cm^2/V\text{-}s$
μ_0	low field mobility	$cm^2/V\text{-}s$

a. The variations ΔL, L_{drawn}, L_{eff}, L_{poly} and XL are explained in the text.
b. The variations ΔW, W_{eff} and W_{extend} are explained in the text.

5.2 Fabrication Line Description

As explained in Chapter 3, the fabrication line description consists of several components. First, the set of process parameters which explain the performance variation seen on the manufacturing line are identified. Second, the means, variances and correlation matrix (or equivalently, the means and the variance-covariance matrix) of the parameters are determined.

5.2.1 The Process Parameters

For this CMOS process, the key process parameters which effect transistor performance are the variations in channel length[1] (ΔL), channel width (ΔW), gate oxide thickness (T_{ox}), threshold voltage (V_T) and channel surface doping level (N_{sub}). Although this parameter set is similar to previous process models for statistical design [1], these parameters differ in the use of V_T instead of the flatband voltage, and the inclusion of N_{sub} to model body effect, which is important for analog design. These parameters are meaningful to the process, circuit and manufacturing engineering groups and support both the manufacturing applications and the physical device modeling methodology employed in this work.

Four of the parameters, ΔL, ΔW, V_T and N_{sub}, are specified for each transistor type (i.e. n-channel, p-channel, EPROM cell, etc.)[2]. There are three important gate oxide thicknesses in this EPROM process, two separate gate oxides and the inter-poly oxide in the EPROM cell. Polysilicon line resistance is included as a product specific parameter because polysilicon is used as the wordline of the memory and its resistance is important to EPROM performance. Thus, a total of 28 parameters were included in the fabrication line description, specifically, ΔL, ΔW, V_T and N_{sub} for each of the six transistor types,

1. The variation in channel length (and similarly, width) is best described and incorporated into the SPICE transistor model by varying ΔL, which is the difference between the drawn and effective channel length. This will be described in detail in Section 5.3.2.

2. In addition to the 3 devices listed, there are 3 additional devices which have adjustments to the threshold voltage implant.

three oxide thicknesses and the polysilicon line resistance. The mean value, operational spread (variance) and correlation structure of these parameters was found from a combination of several months of historical characterization data for the process and detailed characterization of a few samples. The variance-covariance matrix was extended slightly to cover the process even when it is not in control. The techniques used for characterization will be discussed in Section 5.3.

5.2.2 Process Correlation Structure

An important component of the fabrication line description is the process parameter correlation structure. Table 5-2 shows the correlation structure among the 5 process parameters for each of the n- and p-channel devices. Table 5-2 only includes non-zero correlations which are expected from first principles. Although other correlations were significant in certain data sets, they were not included in the fabrication line description to simplify and generalize the description.

Table 5-2 Correlation Structure for a 1.2 μm CMOS EPROM Process[ab]

	ΔLn	ΔWn	Toxn	V_{Tn}	Nsubn	ΔLp	ΔWp	Toxp	V_{Tp}	Nsubp
ΔLn	1					0.75				
ΔWn		1					0.8			
Toxn			1	0.8				1		
V_{Tn}				1						
Nsubn					1					
ΔLp						1				
ΔWp							1			
Toxp								1	-0.8	
V_{Tp}									1	
Nsubp										1

a. The lower left half of this matrix is not shown due to symmetry.
b. Blank entries are not significantly different from 0.

Note that the significance of an estimated correlation coefficient, $\hat{\rho}$, can be found by transforming it using:

$$z = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \tag{24}$$

where z is approximately normally distributed around:

$$m = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \tag{25}$$

with variance:

$$V = \frac{1}{n-3} \tag{26}$$

where n is the number of samples [50]. To test if a correlation coefficient is significant, the null hypothesis is H_0: $\rho = 0$, in which case $m = 0$. For example, if $\hat{\rho}$ is calculated from 100 data points, the variance of the estimate is 0.01, or a standard deviation 0.1, so that any correlation coefficient less than 0.3 is equivalent to 0 at the 3 sigma level of significance.

5.3 Process Characterization Techniques

As this work is oriented towards the development of manufacturing applications, it is important that the parameters used in characterizing the process can be measured on the manufacturing line. The techniques to characterize the underlying process parameters described in Section 5.2 will be summarized here. The extraction of the parameters to complete the device model, such as mobility reduction, will require additional structures and measurements, which will be described in Section 5.4. The process characterization software used in this work is a modification of MOSTCAP [51].

5.3.1 Test Pattern for Process Characterization

A simple test pattern capable of electrically measuring the necessary parameters in high volume production is shown in Table 5-3. The electrical measurements to characterize the process are done after the processing is complete. The algorithms used to charac-

terize the process using this set of structures will be described in the subsections below. For each transistor type, the six transistor parameters ΔL, ΔW, V_T and N_{sub} can be measured electrically using three transistors: one long, one short and one narrow channel.

Table 5-3 Test Pattern for EPROM Process Characterization.

Parameter Measured	Test Structure
ΔL, ΔW, V_T, N_{sub}	Three sizes for each transistor type: (W/L = 20/20, 20/2, 2/20)
Gate and Inter-poly Oxides	70 x 70 μm MOS Capacitors
Wordline Resistivity	800 x 2 μm Poly Line

The characterization techniques rely on the basic equations underlying transistor operation. In general, the measurements bias the device in the linear region. The SPICE Level 3 model for the drain current, I_D, in the linear region is given by:

$$I_D = \mu_{eff} C_{ox} \frac{W_{eff}}{L_{eff}} \left[(V_G - V_T) V_D - \frac{V_D^2}{2} \right] \tag{27}$$

where μ_{eff} is the effective mobility, C_{ox} is the channel capacitance, W_{eff} is the effective channel width, L_{eff} is the effective channel length, V_T is the threshold voltage, V_G is the voltage applied at the gate and V_D is the voltage applied at the drain. The effective mobility is defined as:

$$\mu_{eff} = \frac{\mu_0}{1 + \theta (V_G - V_T)} \tag{28}$$

where μ_0 is the low field mobility and θ is the mobility reduction coefficient.

The characterization techniques presented here bias the transistor at very low V_D (in the 50 - 100 mV range) so that the V_D^2 term in equation (27) can be ignored.

5.3.2 MOS Channel Length and Width Characterization

Variation in channel length is the single most important variable in explaining product performance variation. This variation in channel length is caused by changes in the patterning resist, the polysilicon etch and the diffusion of the source and drain underneath the gate. In this characterization of channel length, the source and drain diffusion under the gate, LD, is held constant, and changes in effective channel length are attributed to changes in the polysilicon gate dimension, XL, that is:

$$L_{eff} = L_{drawn} - \Delta L \tag{29}$$

$$\Delta L = XL + 2LD \tag{30}$$

$$L_{poly} = L_{eff} + 2LD \tag{31}$$

where L_{poly} is the final size of the polysilicon gate, also known as the critical dimension, CD. These relationships are illustrated in Figure 5-2. Note that for a given L_{drawn}, changes in L_{eff} from wafer to wafer are physically described as change in XL, and hence ΔL through the SPICE transistor model. Equivalent definitions hold for the channel width, W.

The channel length shrink is extracted using Chern's method [52], which calculates ΔL and the external channel resistance from a large and short channel transistor. The channel resistance of a MOSFET from equation (27) under low drain voltage is given by:

$$R_{chan} = \frac{V_D}{I_D} = \frac{L_{eff}}{\mu_{eff} C_{ox} W_{eff} (V_G - V_T)} \tag{32}$$

When measuring the channel resistance, the source and drain resistance as well as the probe resistance are also measured, that is:

$$R_m = R_{ext} + R_{chan} = R_{ext} + B(L_{drawn} - \Delta L) \tag{33}$$

where:

$$B = [\mu_{eff} C_{ox} W_{eff} (V_G - V_T)]^{-1} \tag{34}$$

Therefore, if B is held constant (that is, for a specific value of $V_G - V_T$), a plot of L_{drawn} vs. R_m for transistors with different L_{drawn} will result in a straight line. If lines are plotted for several gate voltages, the lines intersect at the point (R_{ext}, ΔL), generating the desired result.

Chern's method is an industry standard approach and was used successfully in this project. However, there are several problems with the method. First, μ_{eff} is function of V_G, causing the relationship between L_{drawn} and R_m to be non-linear. In addition, especially for Lightly Doped Drain (LDD) devices, the source and drain resistance are strong functions of V_G, causing that component of R_{ext} to vary. Newer methods address these issues, but the accurate measurement of ΔL for small devices is still an active subject of research [53][54].

The channel width shrink is extracted using an approach analogous to channel length, based on the relationship between conductance and channel width. A newer, more accurate approach to measuring channel width is given in [55]. Channel width is especially important in circuits such as memories which use a large number of minimum width transistors.

5.3.3 Threshold Voltage and Doping Profile

The linear extrapolation approach was used to measure the threshold voltage. This method extrapolates the maximum slope of the drain current vs. gate voltage curve to the intercept of the gate voltage axis [56]. The maximum slope is used because mobility reduction increases with V_G, creating a non-linear relationship between I_D and V_G. A common technique for finding the point of maximum slope is to plot the transconductance (g_m) against V_G, and then use the value of V_G at the point of maximum transconductance [57].

The surface doping level, N_{sub}, is calculated through measurement of the body effect. The body effect, γ, describes the change in threshold voltage with applied substrate bias, V_{SB}, that is:

$$V_T = V_{T0} + \gamma\sqrt{V_{SB} + \Psi_S} \tag{35}$$

where V_{T0} is the unbiased threshold voltage, Ψ_S is the surface potential under strong inversion, and γ is given by:

$$\gamma = \sqrt{\frac{2\varepsilon_{si} q N_{sub}}{C_{ox}}} \tag{36}$$

[56]. To calculate N_{sub}, V_T is measured for several values of V_{SB}, and Nsub is calculated from the value of γ.

Alternative techniques for calculating doping levels often involve the addition of an operational amplifier in the test setup [58]. Although more accurate and detailed profile information is obtained, the additional hardware increases the complexity for use in high volume production.

5.3.4 Oxide Thickness Measurement

There are three critical oxide thicknesses in the EPROM, including the two separate gate oxides and the inter-poly oxide in the EPROM cell. The thickness of the oxide layers can be measured on large MOS devices (capacitors) using capacitance-voltage (CV) techniques, wherein a DC voltage is applied to the gate of the device to bias it in accumulation and the small-signal capacitance is measured [56]. A large device is used so that fringing fields and changes in the dimensions of the device can be ignored. The oxide thickness is calculated using:

$$T_{ox} = \frac{\varepsilon_{ox} A}{C_{ox}} \tag{37}$$

where A is the area of the device.

5.3.5 Extension to Simulation of Interconnect Variation

Previous work in the field of statistical circuit design has focused on transistor parameters as the cause of performance variation [1][2][19]. Performance in current IC technologies, however, is also strongly affected by the interconnect parasitics. In fact, accurate modeling of interconnect parasitics is critical for the high speed operation of ICs and is an increasingly active field of research. The wordline resistance plays a key role in the performance of the EPROM circuit discussed here. Therefore, this resistance is measured by a polysilicon line in the test structure and its variation is included in the process model.

Although one interconnect test structure is sufficient for this EPROM circuit, in general, the performance of analog and logic circuits will be sensitive to changes in the resistances and capacitances associated with routing on all the polysilicon and metal interconnect layers. A test structure capable of determining these parasitics is described in [59].

5.3.6 In-line Measurements

In-line measurements refer to process characterization completed during the IC manufacturing process. Although this section has focused on electrical characterization completed after processing, it is desirable to make measurements as early as possible in the manufacturing process, for both early prediction of the performance of the fabricated circuits and early detection of process variation. As it is difficult to make electrical contact with test structures before the deposition of the metal layers, in-line measurements tend to depend upon optical techniques, which are less accurate than electrical techniques. Optical techniques can be used for the measurement of linewidths and thicknesses. In particular, the polysilicon linewidth and gate oxide thickness can be measured in-line.

In-line linewidth measurements are made by scanning a light or laser bean across a line and measuring the intensity of the reflected light. The difference in reflected intensity between the interconnect layer and the underlying layer can be detected and used to calcu-

late the width of the drawn line. The accuracy of this technique is limited by the sensitivity of the detection algorithm to the slope and smoothness of the transition in the reflected intensity of the two layers at the edge of the line [57].

The gate oxide thickness can be measured using ellipsometry. An ellipsometer shines a beam of monochromatic, polarized light at a wafer and calculates the thickness of the top layer from the angles of minimum and maximum reflection intensity [57]. Ellipsometry requires a fairly large area sample. In the case of a thermally grown gate oxide, a blank wafer is included in the furnace for characterization. The monitor wafer has a different doping profile than the actual products, and therefore will have a different oxide thickness. However, the monitor wafer thickness will be highly correlated with the gate oxide in the product wafers, making these measurements suitable for performance prediction and Statistical Process Control.

5.4 Device Models

In this section, the physically based statistical SPICE level 3 MOSFET model parameter extraction is described, which completes the mapping of the fabrication line parameters to device characteristics. The model parameters are separated into three groups, as shown in Table 5-4 [13]. The first group contains physical constants. The second group includes the measurable process parameters, such as gate oxide thickness. It is this second group which has been described up to this point in Chapter 5. The third group consists of the fitting parameters, which will be either extracted or globally optimized, and possibly modeled on the measurable process parameters.

Table 5-4 SPICE Model Parameters

Category	Parameter	Description	Units
Constants (from the process)	μ_0	low field mobility	$cm^2/V\text{-}s$
	x_j	junction depth	cm
	NFS	number of fast surface states	$cm^{-2}V^{-1}$
	LD	source/drain diffusion under gate	cm
Directly Measured Physical Process Parameters	XL	poly gate length shrink	cm
	N_{sub}	surface doping	cm^{-3}
	T_{ox}	gate oxide thickness	cm
	V_T	large device threshold voltage	V
	XW	channel width shrink	cm
Fitting Parameters (adjusted to fit I-V calculation to measurements)	Theta	mobility degradation	V^{-1}
	Eta	short channel effect	
	Delta	narrow channel effect	
	Kappa	channel length modulation	V^{-1}
	Vmax	saturation velocity	cm/s

5.4.1 Overview of the Parameter Extraction Method

The physically based parameter extraction requires I-V curves measured on devices covering a wide range of the variability seen for the process. The following steps are taken to complete the extraction.

1. The parameters representing the constants of the process are assigned their measured values.
2. The physical process parameters were measured on these devices, and the corresponding model parameters were set accordingly. This is a key step in physically based parameter extraction, as actual measured values are used in the corresponding SPICE model parameter, for example, the T_{ox} parameter would be set equal to the measured

oxide thickness. These values are not allowed to vary during subsequent optimization phases.

3. The extracted fitting parameters are assigned their measured values. Regression models are built for any fitting parameters which are a function of the process parameters.
4. The remaining fitting parameters are determined by a global optimization.
5. The results are checked against the original I-V curves to verify the extraction process.

To subsequently determine parameter values for a transistor at any point in the process space, the process parameters must be measured. Then, the corresponding model parameters are set accordingly, and values for fitting parameters dependent on the process are calculated. The constants and optimized fitting parameters are set to their global value.

The remainder of this section will describe the techniques for characterizing the constants of the process and determining the fitting parameters. Also, extraction of the capacitance related parameters is discussed. The characterization of the directly measured physical process parameters has already been described in Section 5.3.

5.4.2 Constants of the Process

The constants of the process include the junction depth, source and drain diffusion under the gate and low field mobility. The junction depth is found by taking a cross-section of a device and staining the junction [57]. The extraction of mobility will be explained later in this section along with the fitting parameter mobility reduction. The source and drain diffusion under the gate, LD, is calculated from the overlap capacitance, C_{gd}, on a large device with a known width and oxide thickness, using:

$$LD = \frac{C_{gd} T_{ox}}{W \varepsilon_{ox}} \tag{38}$$

The measurement of C_{gd} will include both the parallel plate and fringing field capacitance, and therefore this calculation will overestimate LD. The correction factor to compensate for the fringing field, which is dependent on the oxide thickness, is derived in [60].

5.4.3 Fitting Parameters

Given values for the constants of the process and the physical process parameters, the values for the fitting parameters are determined next. Fitting parameters can be found through extraction or optimization, and they may be a function of the physical process parameters. This subsection will focus on the extraction of mobility and the mobility degradation coefficient. Mobility degradation is both extracted and functionally dependent on the oxide thickness, so it is a good illustration of the physically based extraction methodology. The proper modeling of mobility and mobility degradation is also critical for achieving a good fit across the process space. Low field mobility is extracted from the same data, so it is also discussed here.

5.4.3.1 Extraction of Mobility and the Mobility Degradation Coefficient

Figure 5-1 shows the data curves used to extract low field mobility, μ_0, and mobility degradation, θ. This data, taken from the large 20/20 n-channel device, plots the reciprocal of effective mobility against V_G - V_T for small V_D, where the effective mobility is calculated, using Equation (27), as:

$$\frac{1}{\mu_{eff}} = \frac{C_{ox}}{I_D}\frac{W}{L}(V_G - V_t)\,V_D \tag{39}$$

The y-intercept is the reciprocal of low field mobility and the slope of the line is θ/μ_0, which can be seen by rewriting Equation (28) as:

$$\frac{1}{\mu_{eff}} = \frac{1}{\mu_0} + \frac{\theta}{\mu_0}(V_G - V_T) \tag{40}$$

Although in this case the data shows little deviation from a straight line, mobility degradation will often increase with increasing gate voltage. SPICE Level 3 models only have a linear mobility reduction term and cannot account for this effect. Improved models such as BSIM3 model are needed to correctly account for mobility degradation [61].

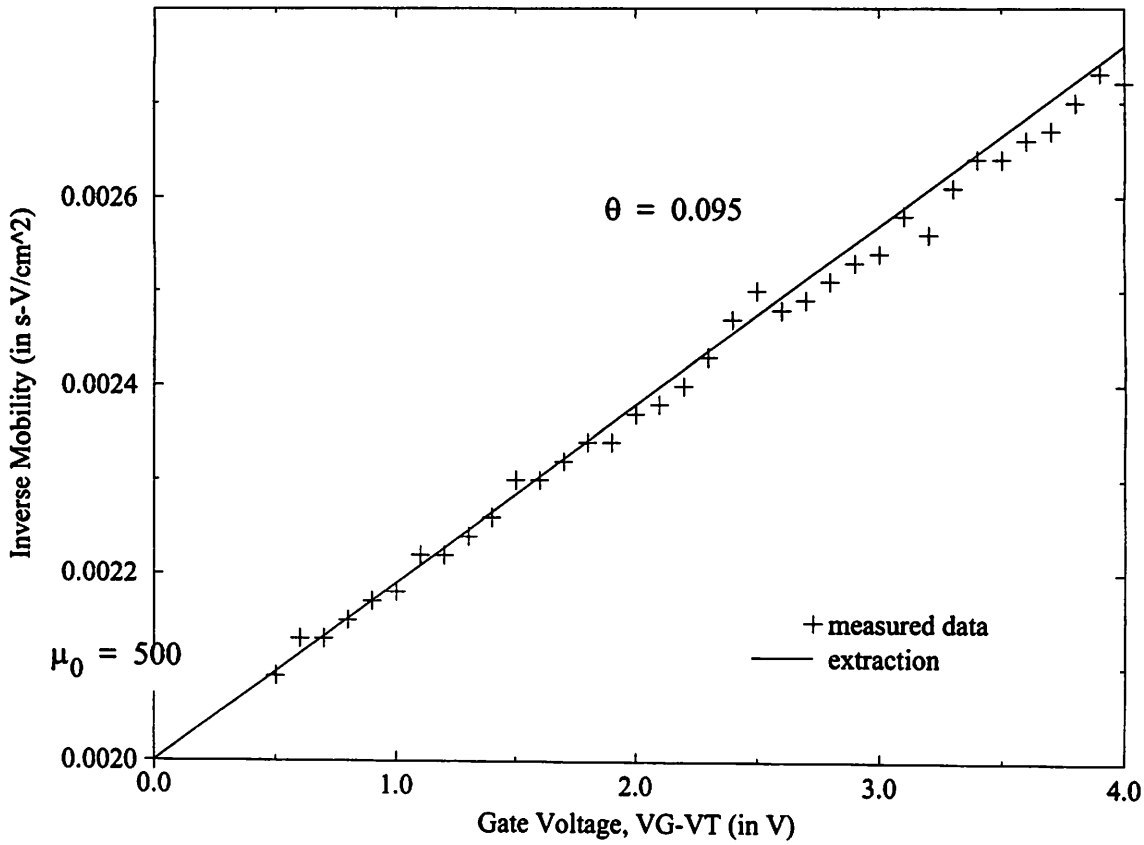


Figure 5-1 Extraction of Mobility and Mobility Degradation

There are several important advantages to extracting μ_0 and θ in this manner. The use of a large device for extraction correctly separates the bulk MOS mobility properties from any geometric effects. In addition, the physically correct extraction of these terms from actual device measurements, as opposed to simply optimized parameter values, allows the physics in the device model to calculate geometry and temperature effects correctly. Otherwise, if geometric effects are confounded with bulk properties or optimized parameters are used, the model prediction will show much greater deviation from actual device characteristics when geometry and temperature are varied.

5.4.3.2 The Dependence of Mobility Degradation on Oxide Thickness

Using the physical parameters in conjunction with single values for the fitting parameters across the process space is not sufficient for accurate modeling across the entire

range of variation. Therefore, to obtain a SPICE model which fits across the process, the device fitting parameters must have a functional dependence on the physical parameters. As was previously explained in Chapter 3, the functional dependence can be derived from device physics or statistical regression techniques. Here, regression models for the device fitting parameters in terms of the measured process parameters were used.

The most important effect is that as the oxide thickness decreases, the saturation current decreases from that predicted by the Level 3 model due to velocity saturation effects [62]. Therefore, the mobility reduction parameter, θ, was made a function of the oxide thickness. A statistical linear regression equation was developed for θ in terms of the oxide thickness. For one device in this process, with oxide thickness *Tox* (in Angstroms), θ is given by:

$$\theta = 0.25 - 0.0005 \times Tox \tag{41}$$

5.4.3.3 Global Optimization

Using the I-V curves collected from across the process space, global optimization was used to find a single value for the other fitting parameters. A numerical optimizer [49] produced a value for each fitting parameter to obtain the best fit across the process range.

5.4.4 Capacitance Device Model

The parameterized Modified Meyer gate capacitance model was used [49][3]. The gate to channel capacitance (C_{ox}), and the drain and source overlap capacitances (CGSO, CGDO) were the appropriate function of Tox. The effect of transistor width variation on the capacitances is handled internally by HSPICE [49]. Note that CGSO and CGDO had no variation due to overlap distance because LD is a constant for the process.

3. In HSPICE, set CAPOP=2.

Junction capacitance is modeled as an abrupt junction and the sidewall capacitance as a linear graded junction. The junction and sidewall diffusion capacitances were held constant, consistent with the view that the source and drain diffusions show little variation.

5.4.5 Final Comments

There are several advantages to the physically based modeling method. First, this model allows accurate circuit simulation at any point in process space, not just at process corners. The resulting SPICE model had a worst case drain current fit of about 8% across geometry and the 3 sigma process range of the key parameters. Second, the method utilizes measured data directly in the compact transistor model, without the need of a device and process simulator. This makes the method particularly attractive to foundry users. These advantages give the user the ability to select a wafer from the fab, measure the necessary parameters, and produce the SPICE model for that wafer. As was explained in Chapter 3, the physically based models connect manufactured wafers and simulation models, enabling the manufacturing applications.

5.5 Product Specific Models

There are two product specific models, the EPROM cell model and the interconnect models. The important point is not these particular models, as many designs will include unique structures which require a product specific model. Rather, the author contends that any structure can be handled using a physically based model, defined at the appropriate level of abstraction. It is important to emphasize that the structure be defined in terms of measurable parameters.

5.5.1 EPROM Cell Model

An extension of the device model is used to describe the EPROM cell. The extension is based on the capacitive coupling of the four MOSFET terminals to the floating gate, as

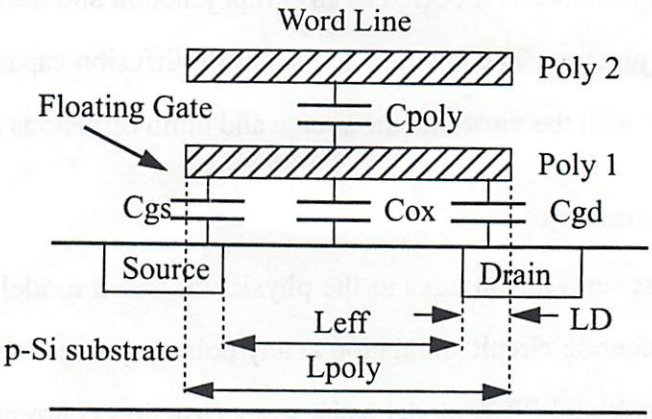


Figure 5-2 EPROM Cross-Section and Floating Gate Capacitance Model

shown in Figure 5-2. The EPROM cell is modeled as an n-channel device, with the floating gate (Poly1) acting as the gate terminal of the device. The voltage on the floating gate is determined from the voltages on the gate and drain, and the capacitive coupling ratios, which act as a capacitive voltage divider [63]:

$$GateCoupling = \frac{C_{poly}}{C_{gd} + C_{gs} + C_{ox} + C_{poly}} \tag{42}$$

$$DrainCoupling = \frac{C_{gd}}{C_{gd} + C_{gs} + C_{ox} + C_{poly}} \tag{43}$$

The capacitance values are calculated from the device dimensions and oxide thicknesses, as simple parallel plate capacitors:

$$C_{poly} = (W_{eff} + 2W_{extend}) L_{poly} \frac{\varepsilon_{ox}}{T_{poly}} \tag{44}$$

$$C_{ox} = W_{eff} L_{eff} \frac{\varepsilon_{ox}}{T_{ox}} \tag{45}$$

$$C_{gd} = C_{gs} = W_{eff} LD \frac{\varepsilon_{ox}}{T_{ox}} \tag{46}$$

The dimensions of the plates are functions of L_{eff}, L_{poly}, LD (all shown in Figure 5-2), W_{eff} and W_{extend} (the extension of the floating gate beyond the electrical width). Currently, L_{eff} and W_{eff} are measured electrically, while LD and W_{extend} are assumed constant. Future models will be based on the in-line measurement of the polysilicon dimensions. Again, a small set of measurable parameters allows the creation of a simulation model over the range of the process for the EPROM cell with similar accuracy to the n- and p-channel device types.

5.5.2 Interconnect Models

The word and bit lines of the EPROM require detailed models due to their large contribution to the memory access time. The resistance of the wordline is significant, and therefore it is measured in the test structure. The capacitance of the wordline is dominated by the large number of EPROM gates connected to it. This gate capacitance is accounted for by the measurement of the two EPROM oxide thicknesses. Therefore, a lumped RC parameter model is used for the word lines, whose resistance values are based on the measured resistivity and capacitance values are functions of the oxide thicknesses.

The resistance of the metal bit line is small. The capacitance of the bit line has components from both the EPROM cell drain junctions attached to it and the metal line. Therefore, the bit line is modeled as a combination of a large fixed capacitor and diodes which represent the junctions.

5.6 Temperature Effects

Accuracy over temperature is a very important feature of a transistor model, as circuits are specified to operate over a wide temperature range. Given the proper physical specification of the bulk silicon properties, the SPICE Level 3 model will account for temperature changes through default parameter temperature dependencies. The most important temperature dependent parameters are the low field mobility, threshold voltage and

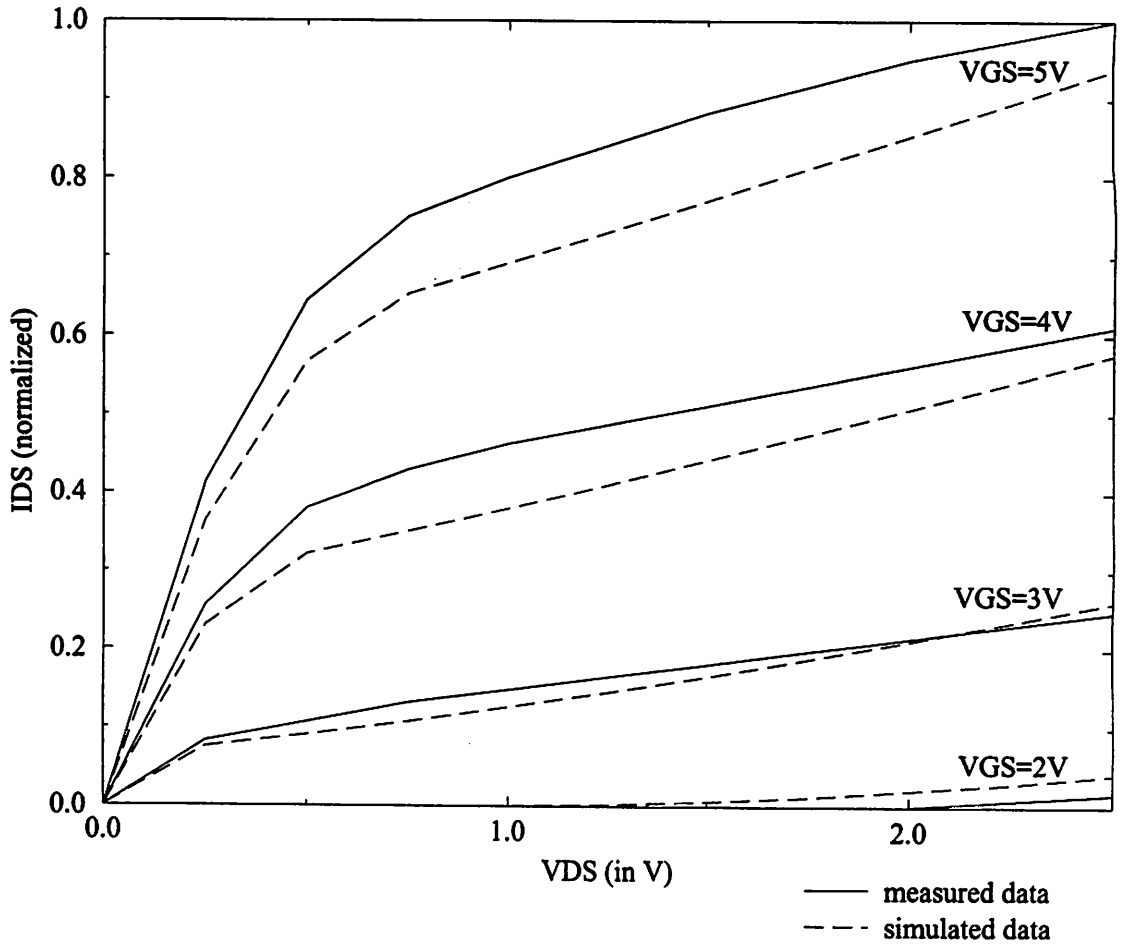


Figure 5-3 Low Temperature Cell Characterization

the saturation velocity. However, the adjustment of the saturation velocity is based on the VMAX parameter, whose value was strictly optimized for the room temperature I-V curves in this work.

The device model was evaluated at 85 °C and -40 °C. The simulated and measured waveforms for -40 °C are shown in Figure 5-3. The model, using parameters extracted at room temperature, gives results with only 2% additional inaccuracy over temperature. Again, there were no adjustments to any parameters other than the Level 3 internal temperature dependencies, and there were no temperature related fitting parameters specified. The physical modeling method along with the straightforward approach to temperature modeling in the Level 3 model is very accurate.

More complex SPICE models force the user to tune temperature related fitting parameters to achieve this level of fit. Although this tuning will improve the fit over temperature, it does require additional work. In [64], physical and statistical models for the SPICE temperature parameters are combined to achieve an accurate fit over temperature.

5.7 Summary

This chapter presented the fabrication line description for a CMOS EPROM process and completed the parameter extraction for a SPICE Level 3 transistor model. The focus of the chapter was the techniques used to measure or characterize the model parameters. In particular, mobility and mobility degradation, and the EPROM cell model, were used as examples of the physically based modeling methodology. Next, Chapter 6 will discuss building the manufacturability applications from the models developed in this chapter.

Chapter 6

Design For Manufacturability Application Examples

Chapter 5 completed the discussion of the fabrication line description and device models which underlie manufacturability applications. This chapter now reports on the development of the applications presented in Chapter 4, a performance prediction model, a control procedure for performance, and an IC Design For Manufacturability (DFM) study. The statistical device models, the parameter distributions from the fabrication line characterization and the circuit simulation model were used together, often in designed experiments, to build the applications. These applications are all developed on a 1 Mbit EPROM product. The applications were built on several generations of a CMOS process, but each generation was characterized using the methodology shown in Chapter 5.

6.1 EPROM Architecture and Circuit Sensitivity

Before developing the applications, the EPROM used as the example circuit will be described. First, the architecture of the 1 Mbit EPROM circuit will be discussed. Second, Monte Carlo results are presented which show the sensitivity of EPROM circuit performance to process variation.

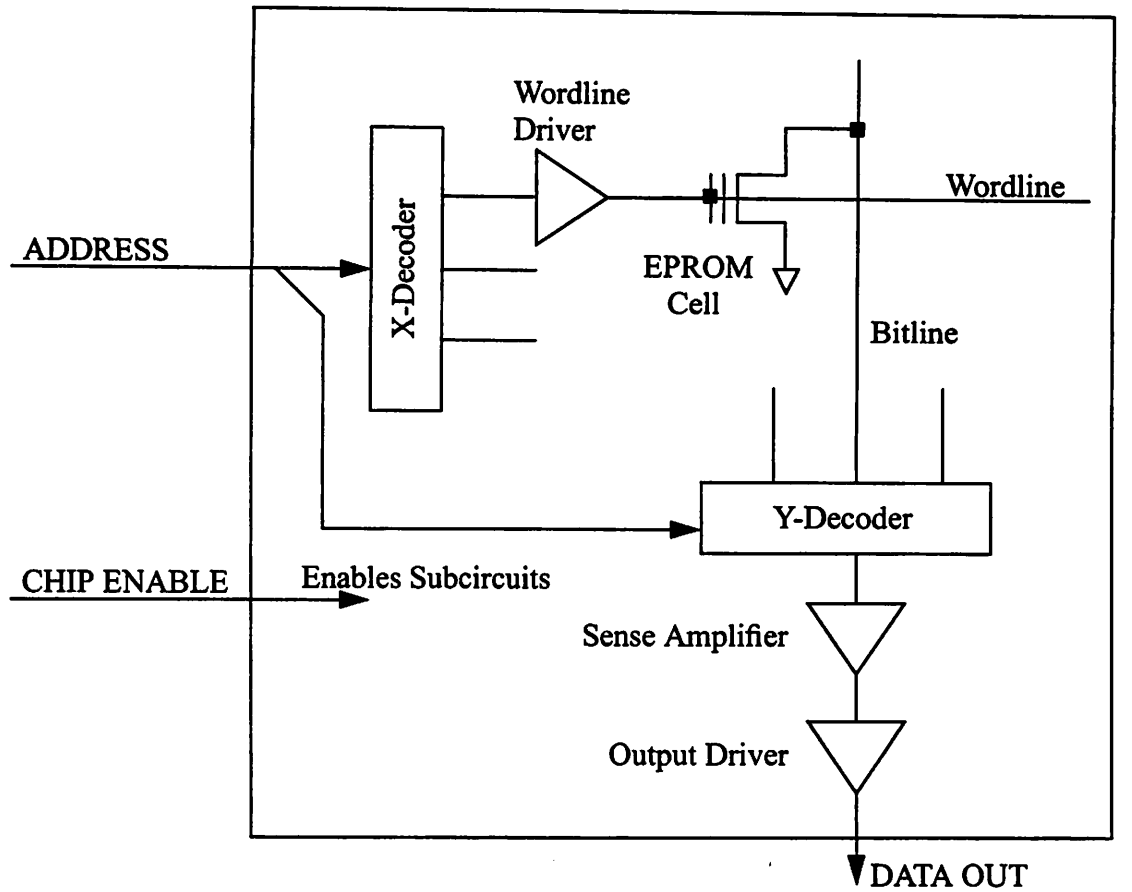


Figure 6-1 Schematic Overview of an EPROM

6.1.1 EPROM Architecture

An overview of the EPROM architecture is shown in Figure 6-1.[1] The incoming memory address is split to control the two decoder blocks. The x-decoder selects one of the wordlines, and each EPROM cell on the wordline puts a "1" or a "0" onto a bitline, depending on whether the cell is programmed or unprogrammed. The y-decoder selects one of the bit lines, and that value is sensed by the sense amplifier and driven off chip by the output driver.

1. Figure 6-1 shows the read path for the memory, because the read access time is the focus of the DFM work. Writing the non-volatile EPROM cell is a slow, one time operation.

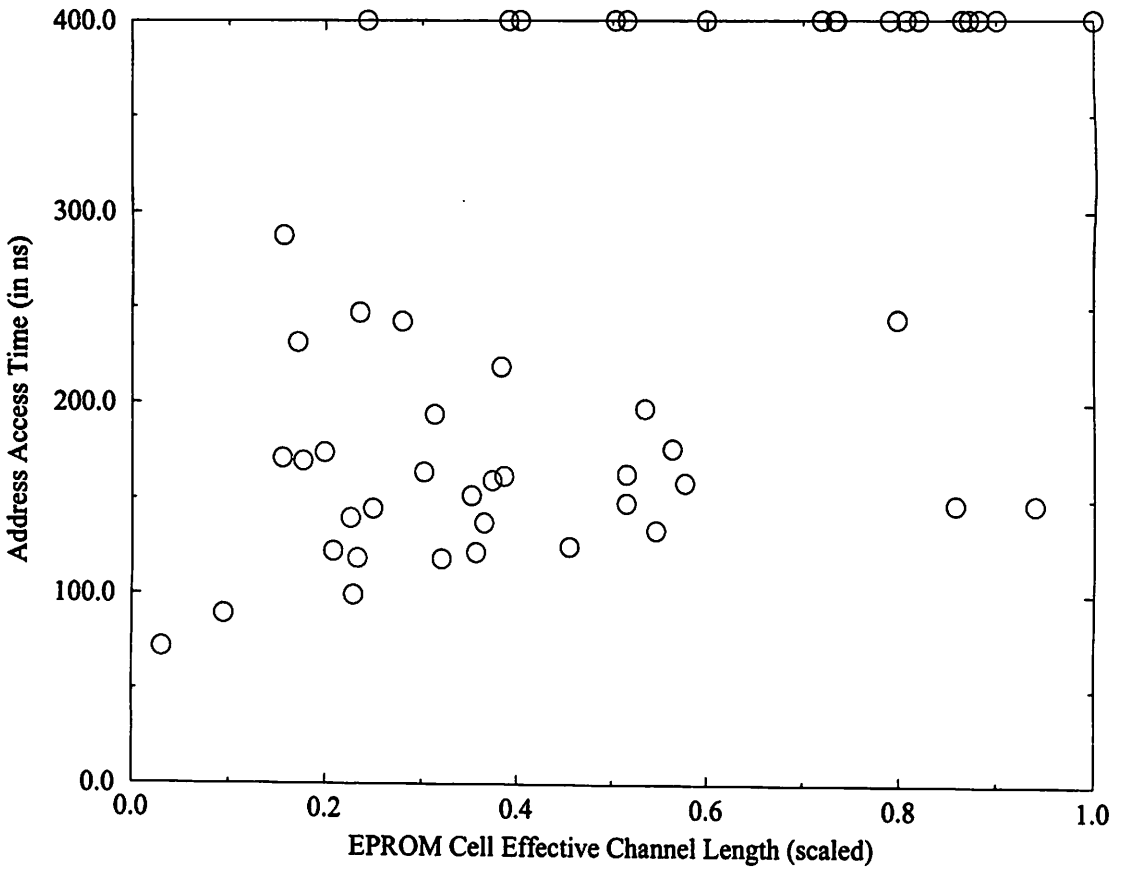


Figure 6-2 Monte Carlo Result: Address Access Time versus Effective Channel Length for 1 Mbit EPROM

The circuit performance modeled in this chapter is the *address access time* (denoted "tacc") of the EPROM, which is defined as the time between a valid address at the EPROM inputs and the data becoming valid at the DATA OUT pin. Typically, EPROMs are binned based on access time, which varies with the process, with the other performance specifications, such as the input and output DC levels, being fixed. Therefore, the access time was chosen as the performance metric to model in the EPROM circuit.

6.1.2 EPROM Sensitivity to Process Variation

The first step to understanding the manufacturability of a circuit is to establish the sensitivity of circuit performance to process parameters variation. Figure 6-2 shows a scat-

terplot of address access time versus the effective channel length of the EPROM cell for a 50 point Monte Carlo simulation of the EPROM. The Monte Carlo shown in Figure 6-2 includes the variation of all process parameters, but the results are plotted against one specific parameter, the effective channel length of the EPROM cell. As the value of effective channel length increases, the EPROM gets slower and starts to fail.[2] In other words, performance yield decreases when the process is producing longer channels. However, there is a lot of variance in the access time even at shorter channel lengths, due to the effects of other process parameters, specifically, the oxide thicknesses and hence the threshold voltages of the transistors.

To more clearly show the effect of the different process parameters in the results of Figure 6-2, a Monte Carlo simulation was run varying only the effective channel length parameters. Figure 6-3 shows the results of the second Monte Carlo. Note that the address access time varies at a given value for the channel length of the EPROM cell because of the variation in the channel lengths of the other transistor types, which are highly, but not perfectly correlated with the channel length of the cell. Figure 6-3 clearly illustrates the strong dependence of performance on channel length variation. Also note that the variance in performance is greatly reduced, showing how the variation of the other process parameters, such as oxide thickness and threshold voltage, also effect circuit performance.

In light of these results, the goal of the performance prediction model is to predict product performance across the range of the process from the measurement of the process parameters. In contrast, the goal of the DFM changes is to reduce the sensitivity of performance to these variations in the process.

2. Parts with a 400 ns access time in Figure 6-2 are either very slow parts or parts which fail to switch.

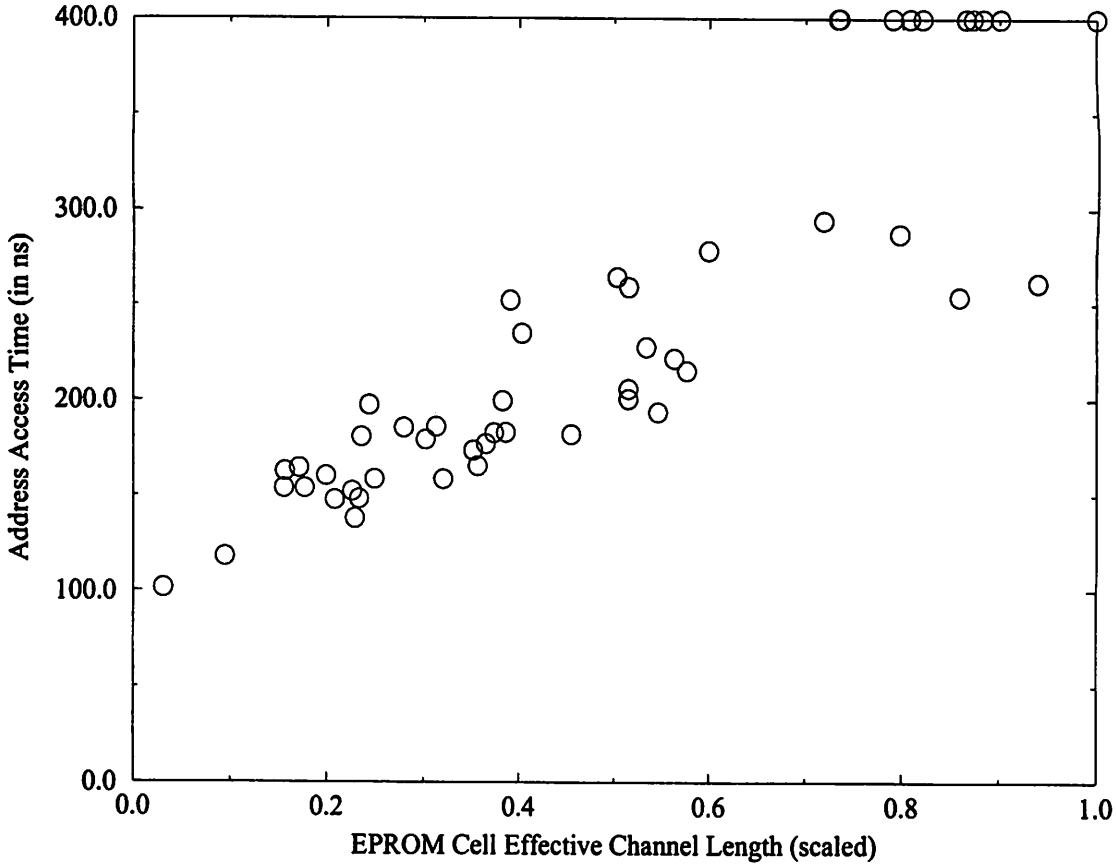


Figure 6-3 Monte Carlo Result Varying only Channel Length: Address Access Time versus Effective Channel Length for 1 Mbit EPROM

6.2 Development of a Performance Prediction Model for Manufacturing

The first application is a model that predicts the performance of manufactured circuits, as was shown in Figure 4-1. A Monte Carlo experiment was used as a computer-based experiment to develop a model for performance in terms of the measurable process parameters. A manufacturing line experiment was run to verify the results.

6.2.1 Monte Carlo Simulation

To cover the range of process variation seen in the manufacturing line, a100 point Monte Carlo simulation experiment was run on the 1 Mbit EPROM. Uniform distributions

were used for the process parameters, and parameter correlation was included to limit the process space of the model. For each of the 100 points, correlated process parameters were generated for each of the six transistor types (n channel, p channel, EPROM cell, etc.) in the circuit. SPICE Level 3 transistor models were generated from the process parameters, as described in Section 5.4. Combining the device models with the circuit netlist for the EPROM, 100 circuit simulations were run and then the access times were extracted from the results.

The simulations were run on a SUN SPARCstation 690 server with four processors. Two simulations, rather than four, were run simultaneously, to allow other users access to CPU time. Each simulation took approximately four minutes, so the Monte Carlo required an elapsed time of 3 hours and 20 minutes. Many circuit designers have even more processors available for running circuit simulation, which will linearly decrease the elapsed time required. Monte Carlo simulation is an excellent application to parallelize, as each simulation can run on an individual processor.

6.2.2 Principal Component Transformation and Regression on Simulation

Using the process parameters generated for the Monte Carlo as the independent variables and the access time predicted by HSPICE as the dependent variable, a linear regression model was built to predict the access time of the EPROM from the measured parameters. To aid in the model building, a Principal Component rotation of the correlated input parameters was employed. The 28 correlated process parameters required seven Principal Components to explain 80% of their variance, as shown in Table 6-1.

Table 6-1 Table of Principal Component Variance

PC	1	2	3	4	5	6	7
% Variance Explained	22.2	20.8	17.6	7.3	4.7	4.0	3.8
Cumulative Variance Explained	22.2	43.0	60.6	67.9	72.6	76.6	80.4

Using step-wise regression, three Principal Components (numbered 1, 2 and 4 in Table 6-1) were chosen to obtain a highly significant model for the address access time of the EPROM. The fact that a significant model was achieved with only 50% of the variance of the process parameters shows the importance of effective channel length in determining circuit performance. It also shows that large portions of the observed variance did not correlate with the product performance. However, as each Principal Component is a linear combination of the 28 process parameters, this model encompasses all the original variables.

The ANOVA table is summarized in Table 6-2. The model explains the variation in performance, as is shown by its high significance. The model has an excellent R-squared of 0.95 and a standard deviation of regression of 1.6 ns, or just over 1% of the 120 ns access time of the part.

Table 6-2 ANOVA Table for Performance Prediction Model.

Error Source	Sum of Sq.	DF	F	Sig.
Regression	4249.4	3	561.7	0.0
Residual	242.1	96		
R-Squared	0.95			
Std Dev of Regression	1.6 ns			

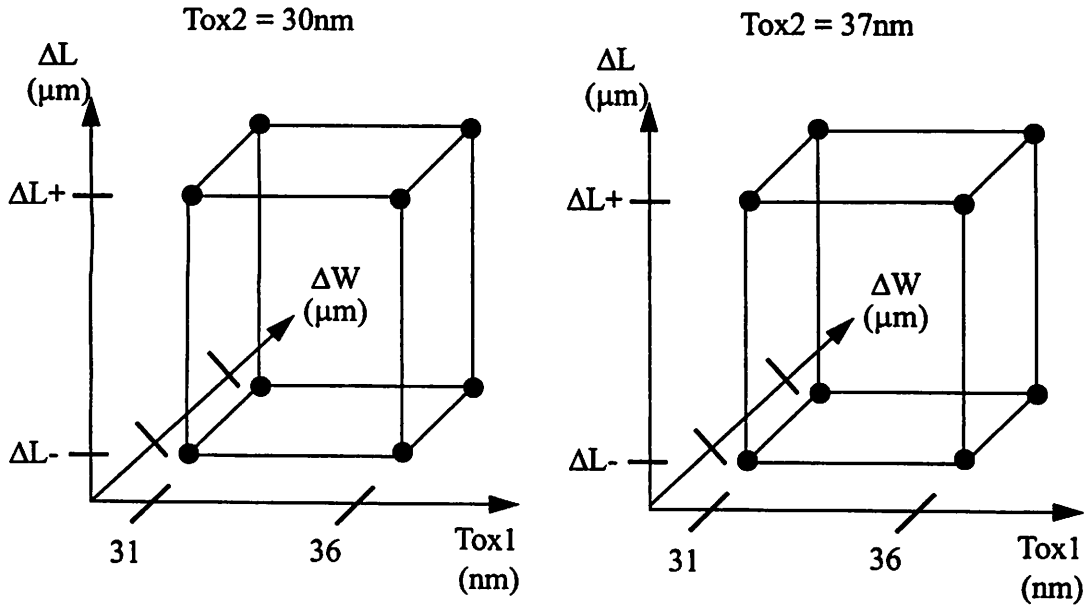


Figure 6-4 Factorial Design for Model Verification

6.2.3 Manufacturing Verification

So far, we have described the statistical circuit model, and how we used a computer-based Monte Carlo experiment to build the performance prediction model from simulation. To complete the integration with manufacturing, an experiment was designed and executed on the manufacturing line. The experiment explicitly varied the process parameters on production wafers. After manufacturing, the process parameters were measured and used as inputs to the performance prediction model developed from simulation. The correct prediction of manufactured product performance verified the manufacturability models and the methodology.

The experimental design, as shown in Figure 6-4, was a full factorial experiment with center point replication in the four most important parameters: ΔL, ΔW, and two of the oxide thicknesses [21][3]. The process flow accounts for the fact that only two out of the

3. In fact, the experiment originally sought to build such a performance prediction model directly from manufacturing data. Combining simulation and manufacturing data is clearly more effective.

three gate oxide thicknesses are independently varied. The minimum and maximum values used in the factorial design were chosen to represent the true excursions of the process over an extended length of production, thus verifying the model over a wide range of conditions.

After the experiment was run through the manufacturing line, the 28 process parameters were measured at electrical test at four sites on each of the wafers. Measurement outliers, including bad data, were replaced by the appropriate average, that is, single points on a wafer were replaced by the average for that wafer and a missing wafer average was replaced by the split or global mean. Rather than the usual performance testing, where the die are placed into a bin based on their performance, detailed histograms of the address access time of the EPROMs from each wafer were recorded after packaging.

The performance test results showed that each wafer has a deterministic spread in speed of approximately 8 ns, independent of the absolute performance of the wafer. Therefore, it must be caused by equipment specific, non-random radial variation on the wafers [26]. As this deterministic spread is consistent, it does not affect the prediction capability of the model, which predicts the average performance of the wafer. The average performance for each wafer is a function of the non-deterministic, but measurable variation of the key process parameters in manufacturing.

The measured parameters were put through the Principal Component transformation determined in the development of the model from simulation. Recall that the Principal Component transformation is based on the correlation structure of the data, as was discussed in Chapter 4. Since the simulated data is based on the correlations found in manufacturing, applying the same transformation to the manufacturing data is appropriate. The principal components were then used in the model built from simulation data. The performance model was used to predict the average access time for each wafer from its average electrical test values.

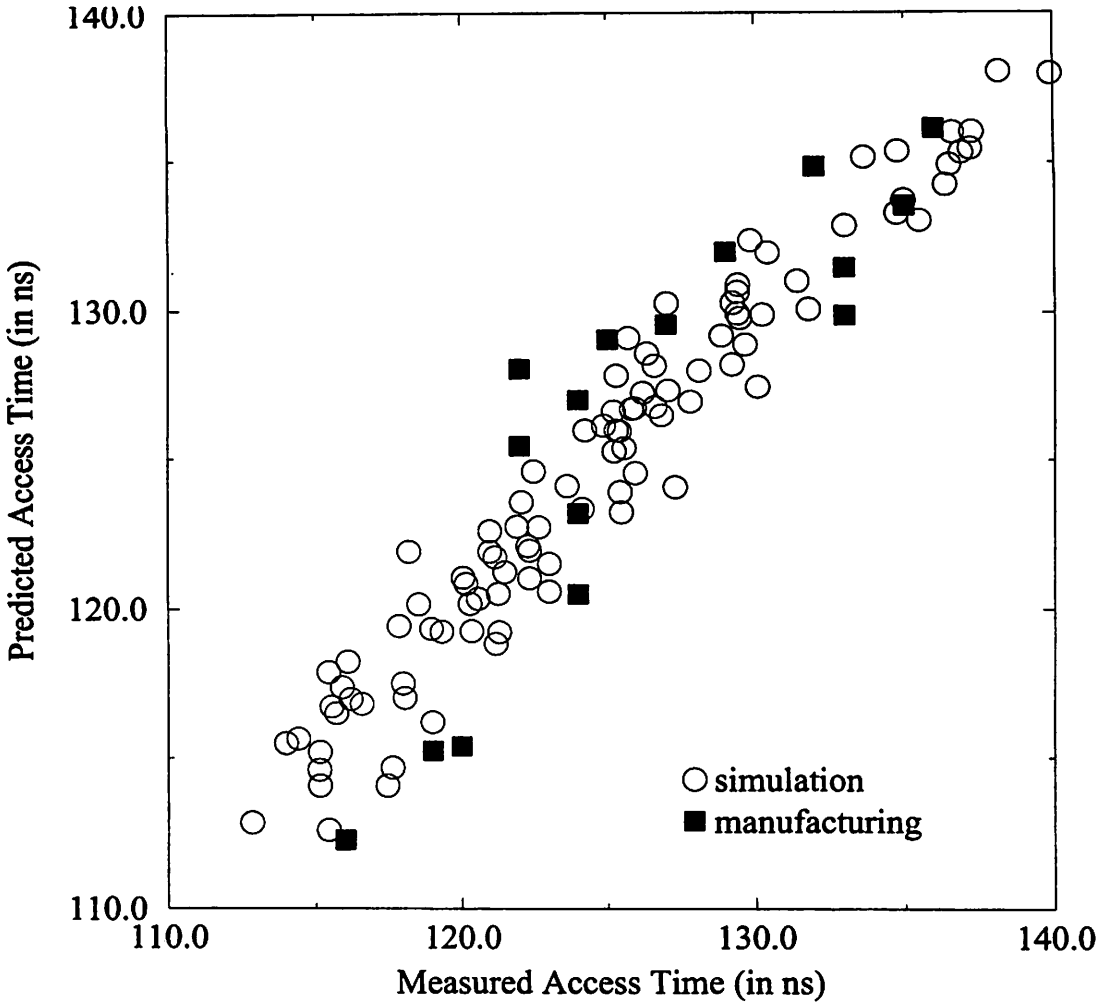


Figure 6-5 Predicted vs. Measured Access Time, Simulation and Manufacturing Data

To account for the cumulative inaccuracy of the simulator, the tester, etc., we allow a change in the constant term of the model in predicting the performances of the manufactured wafers. In this case, the shift in the constant term was 15 ns. A significant portion of this shift was attributed to the redundancy circuits, which are not included in the circuit simulation model but are active in the manufactured parts. Future work on performance prediction for memories will require improvements to correctly handle the implications of redundancy circuits.

To graphically display the goodness of fit, the predicted performance versus the actual average performance for these wafers is shown against the simulation results in Figure 6-

5. Figure 6-5 shows the access time predicted by the linear model plotted against the "measured" access time for these wafers. The "measured" access time is extracted from SPICE for the 100 Monte Carlo points, and is measured at performance test on the manufactured wafers. Again, the manufacturing data points represent the mean access time for the die on that wafer. The predicted access time is the prediction of the model from the process parameters, using different constant terms for the simulation and manufacturing data. A good fit can be seen over the entire range of performance. It is important to emphasize that the manufacturing data was not used to build the model (except the constant term), and therefore the good fit is an indication of both the accurate predictive capability of the model and that the statistical device model and the performance prediction model are valid across the range of process variation.

6.2.4 Analysis of Residuals

The regression residuals are often good indications of relationships between the dependent and independent variables which have not been properly included in the model [65]. Figure 6-6 plots the residuals from the model for both the simulation and manufacturing data against the predicted access time[4]. Although the residuals from the simulation data appear to be identically, independently and normally distributed (IIND), the residuals from the manufacturing data show a decreasing trend as the predicted access time increases.

From a statistical modeling point of view, trends in residuals are most likely indications that additional terms should be added to the model. Although adding non-linear terms would improve the distribution of the residuals, this is not the root cause, because this model results from both simulation and a statistical regression model building procedure. In fact, this trend in the residuals was more pronounced in the early stages of this

4. Residual plots against the dependent variable should always plot against the fitted results, rather than the actual results, as there is an expected relationship between the residuals and the actual results used to build the model.

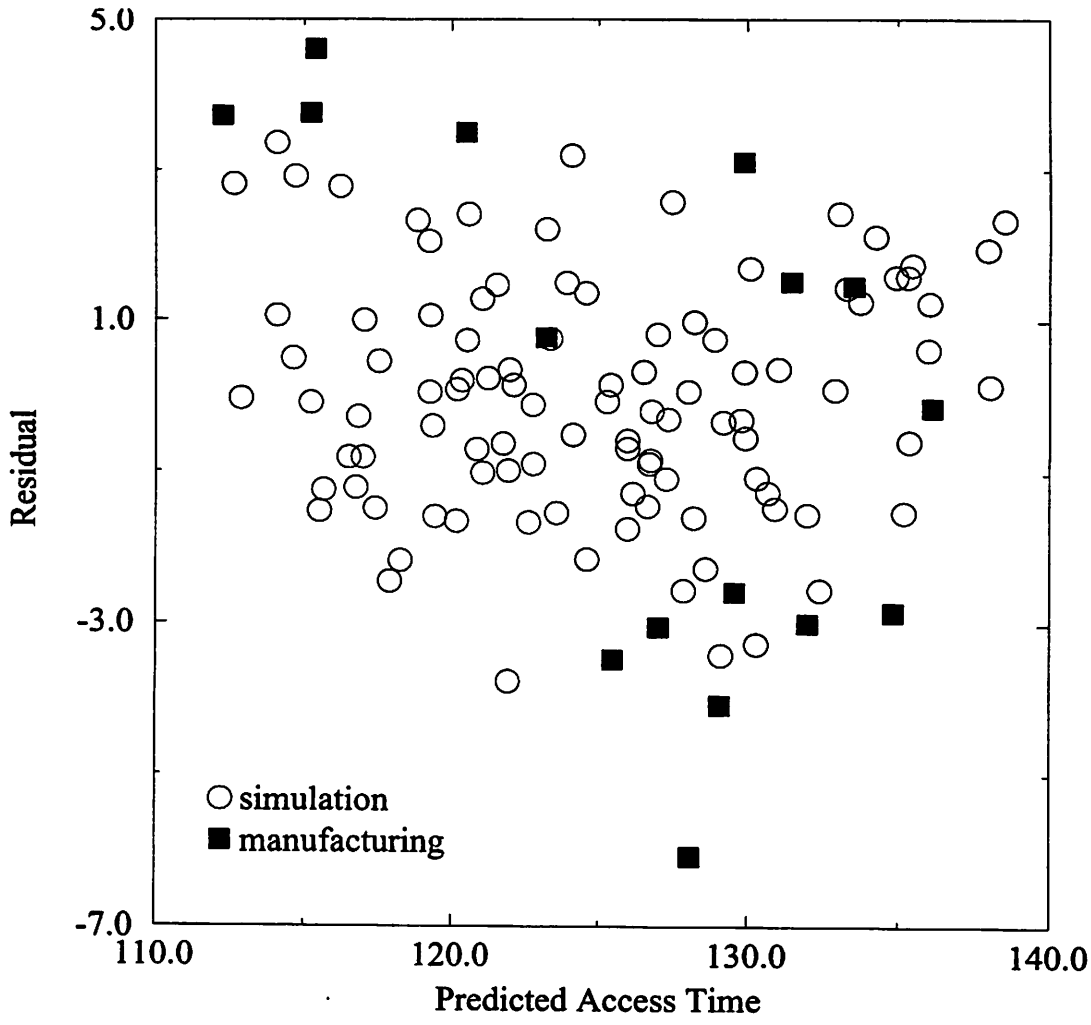


Figure 6-6 Performance Model Residuals versus Fitted Access Time

project, and also showed in a plot of residuals against oxide thickness. The problem was reduced by improving the handling of mobility degradation (θ) in the device model. The residual trend in Figure 6-6 is most likely an indication that further improvements to the mobility degradation model are needed.

In general, the addition of non-linear effects to improve the model was not pursued in this project. This project focused on the issues related to applying manufacturability modeling to an actual fabrication line. The author believes that other issues, such as measurement capability, device modeling and CIM infrastructure, need further consideration before the lack of fit of the model becomes a priority issue. In addition, non-linear terms

can potentially amplify measurement noise, rendering the predictions meaningless. Work in this area which has included nonlinear effects can be found in [6] and [66], however, these works make limited use of actual manufacturing data.

6.3 Statistical Process Control for Product Performance

The second application developed was a model based control chart for product performance, which utilized a performance prediction model to predict the fraction of a lot which should occur in each bin. The model used in this application was based simply on device currents, rather than the process parameters, as the currents for each lot were more readily available through the CIM system. Table 6-3 shows the bin specifications for the access time of the 1 Mbit EPROM.

Table 6-3 Bin Specifications for the 1 Mbit EPROM

Bin	Access Time (in ns)
1	120
2	150
3	170

A total of 7 lots were put through the control procedure. Table 6-4 shows the actual test results and the upper and lower limits for the fraction occurring in each bin, as calculated by equations (19) and (20) with $Z_{\alpha} = 2$.

Table 6-4 Control Procedure Results[a]

Lot	Lower Control Limit			Actual Bin Fractions			Upper Control Limit		
	Bin 1	Bin 2	Bin 3	Bin 1	Bin 2	Bin 3	Bin 1	Bin 2	Bin 3
1	0.0	0.4	0.0	0.02	0.94	0.04	0.6	1.0	0.0
2	0.0	0.0	0.0	0.0	0.96	0.04	1.0	1.0	0.0
3	0.0	1.0	0.0	0.0	0.98	0.02	0.4	1.0	0.0
4	0.0	1.0	0.0	0.0	0.02	0.98	1.0	1.0	0.0

Table 6-4 Control Procedure Results[a]

Lot	Lower Control Limit			Actual Bin Fractions			Upper Control Limit		
	Bin 1	Bin 2	Bin 3	Bin 1	Bin 2	Bin 3	Bin 1	Bin 2	Bin 3
5	0.0	0.4	0.0	0.0	0.99	0.01	0.6	1.0	0.0
6	0.0	0.4	0.0	0.0	0.95	0.05	0.6	1.0	0.0
7	0.0	0.0	0.0	0.0	0.97	0.03	1.0	1.0	0.0

a. All entries in this table denote fractions of the lot in a bin.

Lot 4 generated an alarm because the actual test result was slower than the prediction of the model. As these lots were all shipped immediately to customers, the assignable cause of the alarm could not be determined. Lot 4 had the smallest CD and the largest currents of all the lots in the control chart, therefore, it is likely that the lot was incorrectly tested. However, it is possible that noise induced by high current switching actually reduced the performance of this lot.

The other lots all had small fractions of product in Bin 3. Although the control chart limit for Bin 3 was always 0 because the process was producing faster parts during this time, the small fraction for Bin 3 found in actual testing should not generate an alarm. There is often a long tail on the actual performance distribution, because a small number of cells in a lot will have low currents and produce slow parts. This is not a symptom that the process is out of control.

These results illustrate the value of instituting a model based control procedure for product performance. On most final test floors, although severe drops in yield will generate an alarm, there is generally no mechanism for detecting when a lot has been incorrectly tested. However, given a model between electrical test and product performance, final test results can easily be verified by the CIM system. Increased sensitivity to test problems and process shifts represents the important benefits of process control.

6.4 DFM Application for Low Voltage EPROM

The final application to be discussed is an improvement to the manufacturability of an IC product. The circuit chosen to demonstrate DFM was a commercial 1 Mbit EPROM. This product was designed to operate with a 5V supply, however, market demand for ICs which operate at lower voltages is increasing, driven by the need for portable, battery powered devices [67]. The goal of this project was to improve the manufacturability when the EPROM was operated from a 3V supply. Currently, the EPROM design operates at 3V over a certain range of the process, and a subset of production selected by an appropriate performance prediction model can be shipped to customers for operation at 3V. However, the product performance at this lower supply voltage is sensitive to process variation. Therefore, the 1 Mbit EPROM was selected to test the DFM methodology developed in this work.

This situation is typical of a common industrial decision - given the demand for a product to operate at a lower supply voltage or higher performance level, can the design be modified to meet this need, or is a re-design required? A complete re-design will result in a longer time to market for the product, however, making only minor modifications to the design brings the risk of only slightly increasing the performance yield. The DFM methodology presented here is intended to address theses concerns by improving the re-design process. DFM will allow the circuit designer to complete the task faster and more effectively, lowering the risk of the re-design.

Memory circuit performance is critically dependent on the characteristics of the memory cell. However, it is very expensive to change the fabrication process to alter the cell characteristics and then verify reliable product operation. Therefore, the DFM effort focused on the circuits surrounding the cell, including the wordline driver and the sense amplifier, as was shown in Figure 6-1. Limited circuit details will be presented due to the

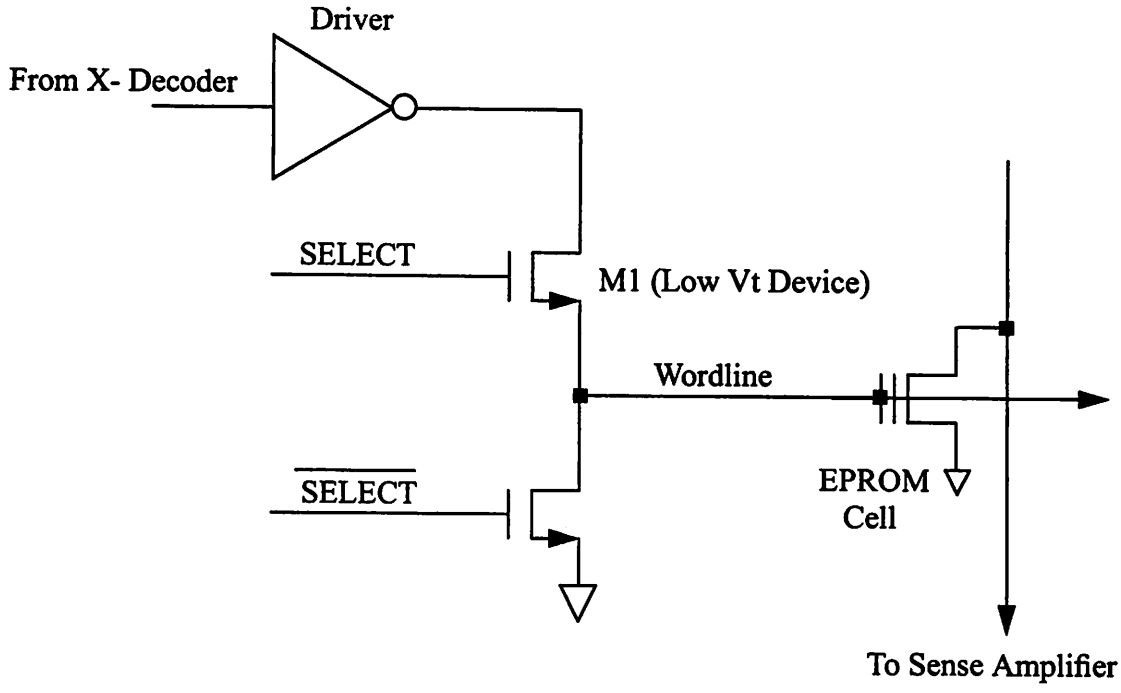


Figure 6-7 EPROM Word Line Driver

proprietary nature of these production circuits, however, the results of the methodology are independent of the circuit details.

This section will present a circuit overview, the DFM analysis, circuit modifications and verification results for the wordline driver and sense amplifier circuits. This section will conclude with a discussion of this DFM methodology.

6.4.1 Wordline Driver Circuit

Given the architecture and sensitivities of the EPROM, the first circuit chosen for analysis was the wordline driver. The wordline driver circuit is shown in Figure 6-7. The wordline is driven through an n-channel device with a lower threshold device, which acts as the final stage of decoding. The gates of the EPROM cells are connected to the word-line.

6.4.1.1 Word Line Driver Circuit Analysis

The voltage on the word line is applied directly to the gate of the EPROM cell. Therefore, the wordline voltage should be as close to the supply voltage as possible. For products specified to operate at 3V with +/- 10% operating margins, the worst-case supply voltage for simulation and testing is 2.7V. The cell has a high threshold voltage to protect against breakdown during high voltage programming, and the EPROM model showed that the threshold voltage for the cell was over 2V under the worst-case process conditions. The Monte Carlo simulations showed circuit failures, as small V_G - V_T values for the cell were causing low cell currents. The goal of the circuit change to the wordline driver is to increase the voltage on the gate of the EPROM cells, and thereby increase the cell current and improve the circuit performance for all process conditions.

6.4.1.2 Word Line Driver Circuit Modifications

Examining the circuitry driving the word line, the voltage on the word line is limited to the supply voltage minus the threshold voltage drop required to turn on transistor M1 in Figure 6-7. Although this threshold voltage drop is not a problem with a 5V supply, this voltage drop becomes critical for low voltage operation. To improve the word line voltage level, the threshold voltage of device M1 was lowered. However, lowering the device threshold voltage results in excessive leakage in the case where the output of the word line driver is high and the final decoding stage is not selected (SELECT is low and $\overline{\text{SELECT}}$ is high). This leakage caused the power to exceed the specification. The leakage problem was solved by inserting an additional transistor between the n-channel device and ground.

6.4.1.3 Word Line Driver Verification Results

A single split lot experiment was run through the fabrication line to verify the design change. One half of the wafers contained the original design, and the other half contained the design improvement, including the lower threshold voltage on device M1. The wafers

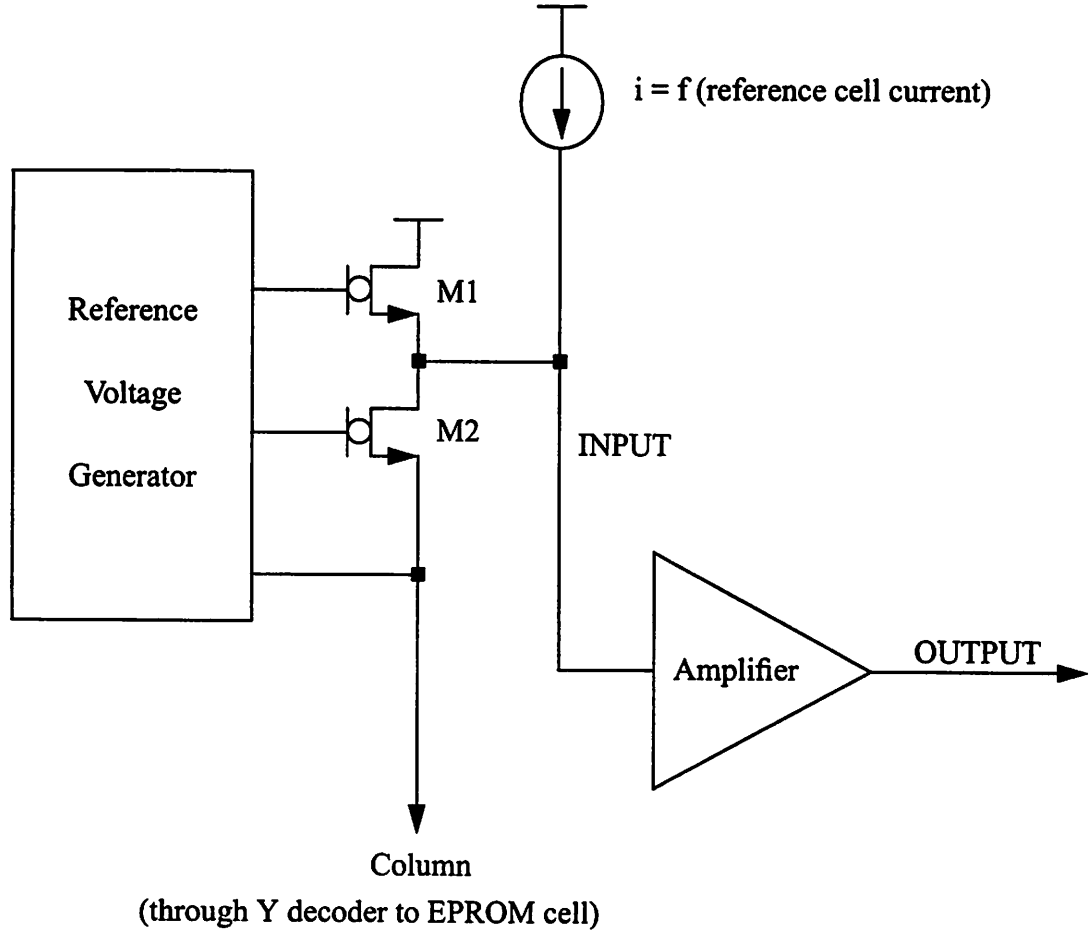


Figure 6-8 EPROM Sense Amplifier

with the new design had an improvement in the lowest power supply level for fully functional operation equal to the threshold voltage difference of the two designs. This increase in the power supply margin will improve the yield of lots across a wide process range for 3V operation.

6.4.2 Sense Amplifier Circuit

The next circuit to be discussed is the sense amplifier, shown in Figure 6-8. This is a current sensing circuit, wherein the presence of cell current (an unprogrammed cell) pulls

the node labeled "INPUT" to a low voltage level and the absence of cell current (a programmed cell) allows the current source and transistor M1 to pull the node "INPUT" to a high voltage level. The amplifier is made differential and compensated for changes in cell current across the array through the use of an unprogrammed reference cell, whose current is multiplied to provide a reference current to the sense amplifier.

The sense amplifier is characterized by three voltages, all related to the node labeled "INPUT" in Figure 6-8. These voltages are the low and high DC levels and the level which switches the amplifier stage. The levels are called the lower window, the upper window and the trip point, respectively. The final amplifier stage is designed to set the trip point in the middle of the window voltages.

Monte Carlo simulations were run, including all process parameters across the range of process variation, to determine the lower window, trip point, upper window and access time for the sense amplifier. The results for the window and trip point levels for the original design are shown in Figure 6-9. Each vertical column of three points represents the results from a set of simulations run at a specific point in the process. For improved visualization, the results are sorted in the order of increasing cell current. There is a decreasing trend in the level of the upper window and trip point, showing that these levels are set by circuits sensitive to overall current levels. The lower window is flat over the process, with fluctuations caused by changes in the threshold voltage of the n-channel transistors. Figure 6-9 also shows a reduction in margins at lower cell currents.

Section 6.4.2 and its subsections, which cover the modification, analysis and verification of the sense amplifier, compare two circuit implementations of the sense amplifier. The first implementation, which has already been presented, is the original design. The second is a minor modification which will be named the new design. At the time the new design was proposed, a problem with the DFM software made it appear that the new design would improve the performance yield of the circuit. In fact, this was not the case.

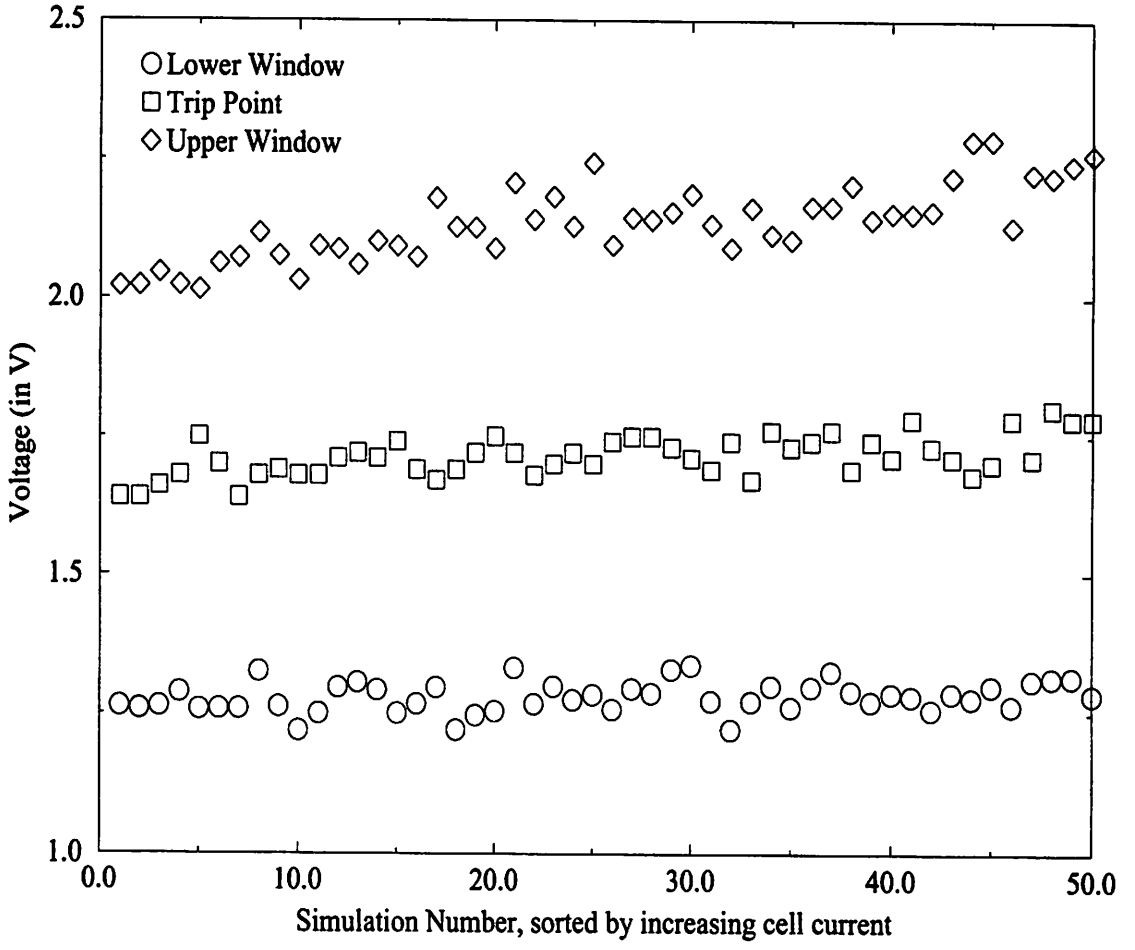


Figure 6-9 Monte Carlo: Window Levels for Sense Amplifier - Original Design

All the simulation results reflect the corrected software. This problem with the DFM software will be discussed in Section 6.4.3.

6.4.2.1 Sense Amplifier Circuit Modifications

The EPROM product under consideration is a high volume, low cost product. To limit the cost involved in the design improvement, circuit changes were allowed only in the polysilicon mask, which limited changes to the drawn length and width of the transistors. The circuit designer selected 4 transistors in the sense amplifier for size modification, including the zero threshold devices labeled M1 in Figure 6-8 and 3 transistors in the voltage generator. The goal of the design changes were to modify the DC voltage levels of the

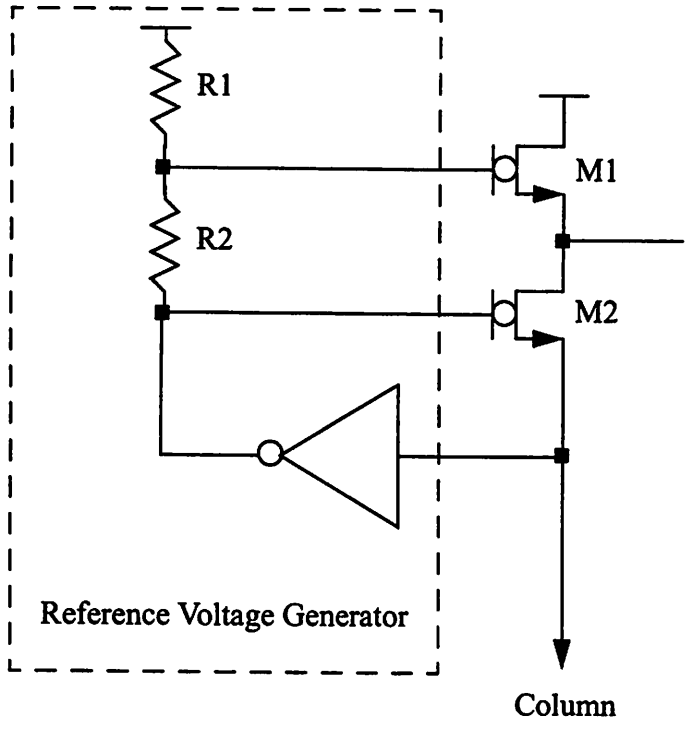


Figure 6-10 Reference Voltage Generator Equivalent Circuit

windows depicted in Figure 6-9. The exact sizes were chosen with the aid of the optimizer in HSPICE.

To explain the circuit changes in the reference voltage generator, a simplified equivalent circuit is shown in Figure 6-10. The voltage bias levels at the gates of transistors M1 and M2 are set by the voltage drops across controlled impedances, shown schematically as R1 and R2. The amplifier is part of a circuit which clamps the voltage on the bit line, as any voltage swing on the bit line would slow down the access time due to the large capacitance on the bit line. The circuit changes in the reference voltage generator were to decrease the resistance of R1 and R2 and lower the current flowing through them. This reduces the effect on the reference voltage levels of the small changes in the output voltage of the amplifier caused by variation in the threshold voltage of the transistors. Another effect of these changes was to raise the DC level of the upper and lower windows.

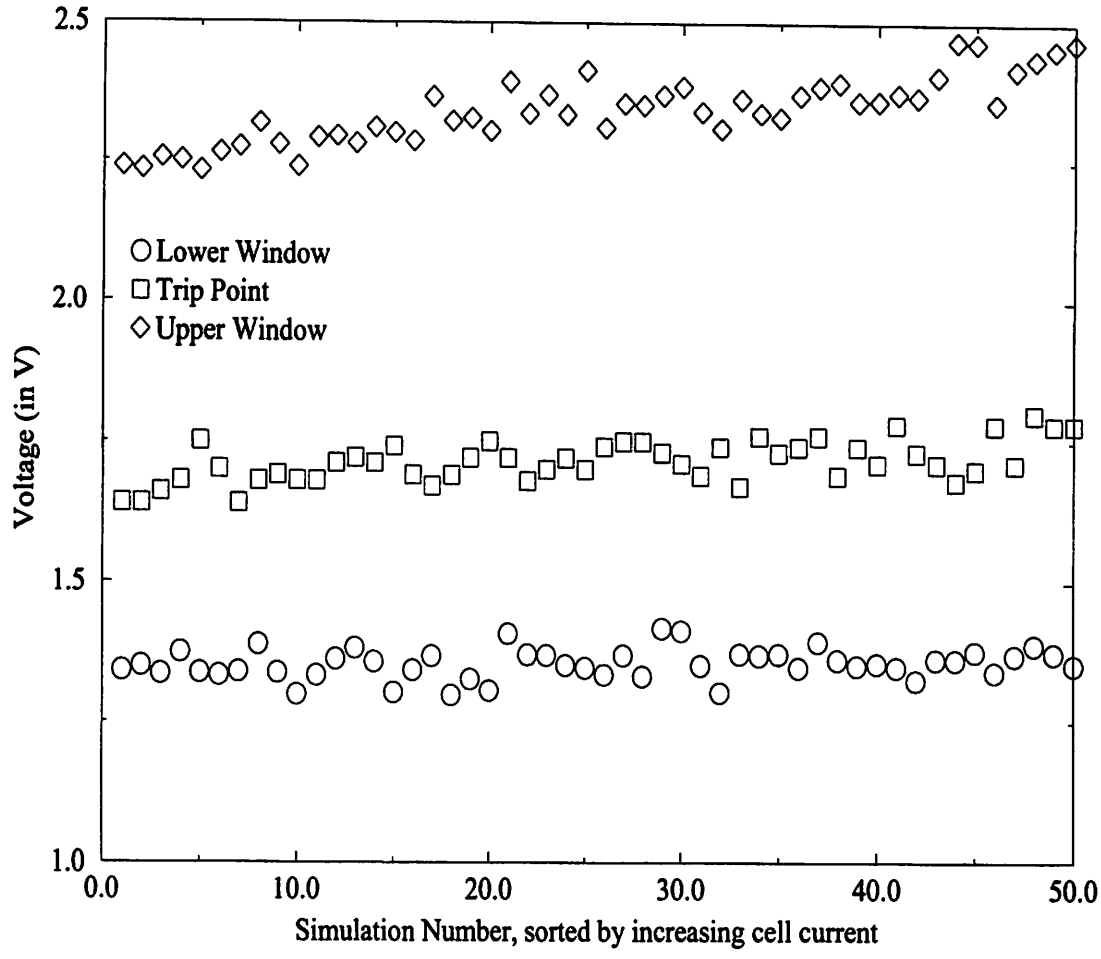


Figure 6-11 Monte Carlo: Window Levels for Sense Amplifier - New Design

The other circuit change was to reduce the size of the load transistor M1. As can be seen in Figure 6-8, the current source and transistor M1 are both supplying current to the INPUT node for the case when the EPROM cell is unprogrammed (i.e. the cell transistor is on). In the case of low cell currents, the low impedance of the load slows the pulldown of the INPUT node, so the impedance of M1 was raised to increase the switching speed.

The Monte Carlo results for the windows and trip point of the new design are shown in Figure 6-11. The circuit changes raised the upper window level by about 250mV and the lower window raised by about 125 mV, with the trip point unchanged. Further, the variability of the upper window has been reduced. This design change has moved the trip

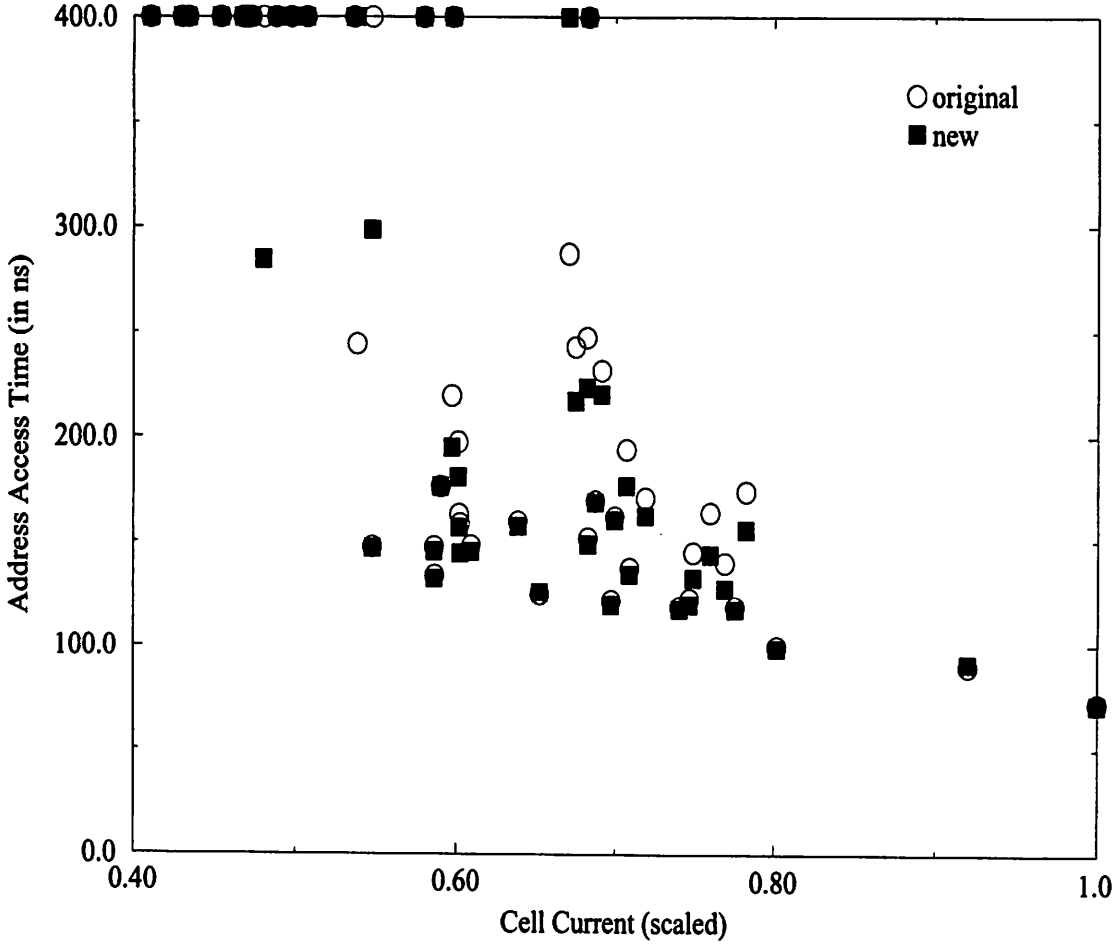


Figure 6-12 Monte Carlo: Read Access Time versus Cell Current - Original and New Designs

point off of center, which will have the effect of slowing the transition from the upper to the lower window with respect to the transition from the lower to the upper window, that is, the transition from the lower to the upper window will reach the trip point faster.

6.4.2.2 Sense Amplifier AC Analysis

To further analyze the sensitivity of the original and new circuit designs, Monte Carlo simulations were run, again over all process parameters, to determine the address access time of the original and new designs. The results are shown in Figure 6-12, which plots the address access time of the EPROM against the cell current. Cell current is used as the x-axis in this graph, instead of effective channel length as was used in Figure 6-2 and Figure

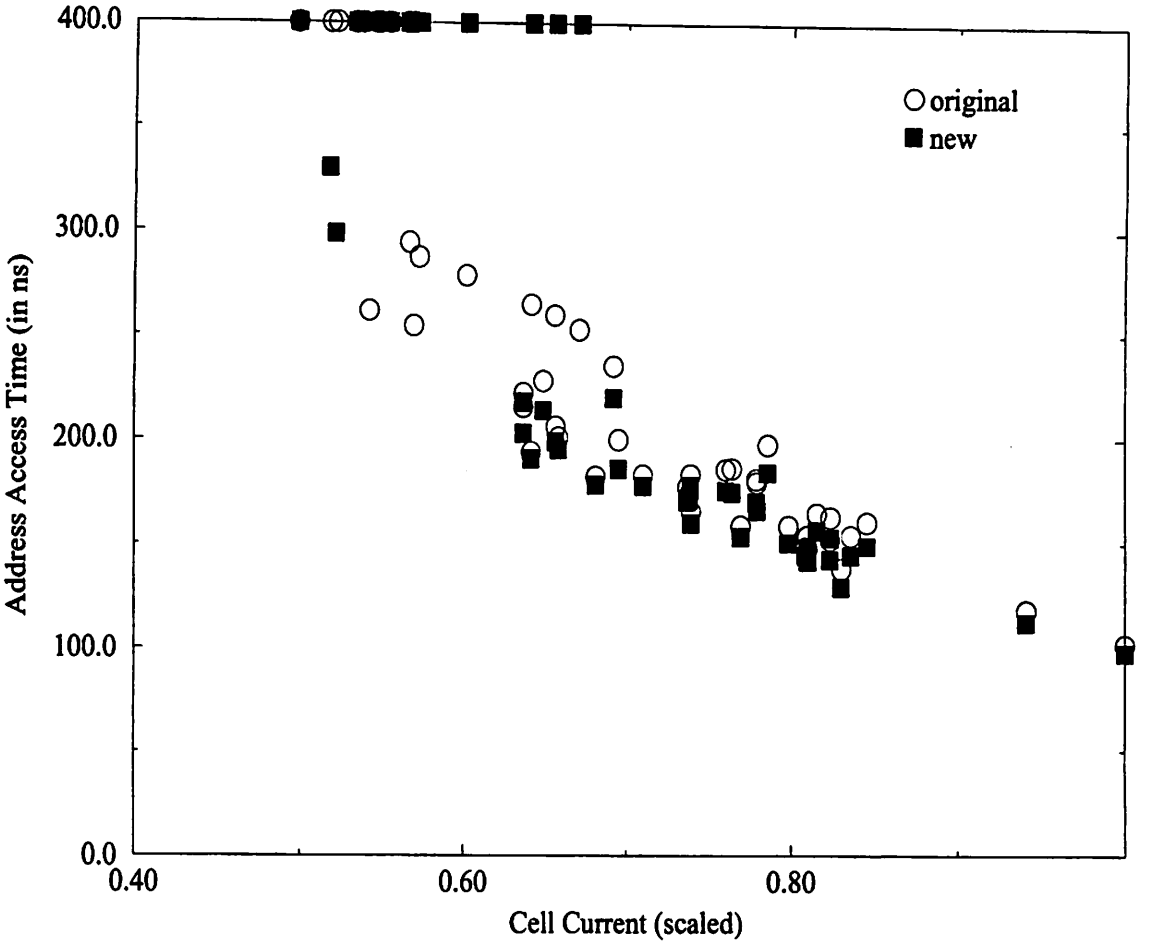


Figure 6-13 Monte Carlo Varying Only Channel Length: Read Access Time versus Cell Current - Original and New Designs

6-3, because cell current represents the combined effect of all process parameters. Figure 6-12 shows that the average value and the variance of the EPROM access time increase as the cell current decreases. Figure 6-12 does not indicate that there is a significant difference in the performance of the original and the new designs.

However, Figure 6-13 gives additional insight into the new design. Figure 6-13 plots the address access time against cell current for the new and original designs, for a Monte Carlo simulation which varied only the channel length of the transistors. Figure 6-13 shows that the new design actually has an increased sensitivity to channel length variation. For EPROM circuits having cell currents with scaled values between 0.55 and 0.65, there

is a group of 8 simulations which operate at approximately a 275 ns access time for the original circuit design, but these simulations are either slow or fail completely for the new design.

6.4.2.3 Sense Amplifier Verification Results

To verify the simulation results for the original and new sense amplifier circuit designs, a fractional factorial experiment was run on the manufacturing line. The experiment, summarized in Table 6-5, varied the polysilicon linewidth at three levels to produce samples with slightly longer, nominal and slightly shorter channel lengths. These three levels produced transistors with low, medium and high device currents. Both the original and new designs were fabricated, resulting in a total of six splits, which were processed within a single lot. The access time for approximately 80 samples[5] from each range of cell current was characterized through detailed performance testing to examine the sensitivity of the two designs to process variation.

Table 6-5 Sense Amplifier Verification Experiment

Factor	Number of Levels	Description of Level		
		1	2	3
Design	2	original	new	
Cell Current[a]	3	high	nominal	low

a. The cell current was varied by process changes which effected the polysilicon linewidth. This change effected all transistor types.

The results of the verification experiment are summarized in Table 6-6. The average address access time for the original design increases from 108 ns to 135 ns, and its standard deviation increases, over the range of cell current. The shows good agreement with the simulation results shown in Figure 6-12, as the range of cell current variation seen in

5. The entire lot was not performance tested due to a mixing of the splits during packaging. Instead, the cell currents in the packaged EPROMs were measured and separated into ranges for testing.

the experiment corresponds to the scaled range from 0.7 to 0.9 in Figure 6-12. However, the access times for the new design are 15 ns slower than the original design.

Table 6-6 Sense Amplifier Verification Results

Cell Current Range	Performance Mean (in ns)		Standard Deviation (in ns)	
	Original Design	New Design	Original Design	New Design
High	108.0	122.3	9.4	12.6
Nominal	118.6	129.3	9.6	11.3
Low	135.1	151.3	16.1	16.6

The increase in access time for the new design is explained by the simulation shown in Figure 6-13. The Monte Carlo simulation shown in Figure 6-13 represents the process space of the experiment, as channel length was varied over a broad range with no variation in oxide thickness or threshold voltage. Because there is a distribution of cell currents in a 1 Mbit memory, manufactured parts in the higher cell current ranges still have some cells with lower cell current [68]. Therefore, even higher cell current parts have cells which enter the region of high performance sensitivity, producing the lower performance shown in Figure 6-13. In other words, performance is dominated at all current ranges by the lowest current memory cells in the part.

Table 6-6 also shows that the standard deviation of the new design increased with respect to the original. The change in variance can be tested for statistical significance by calculating the ratio of the variances, which follows the F distribution. Therefore, for the high cell current case:

$$\frac{\hat{\sigma}^2_{new}}{\hat{\sigma}^2_{current}} = \frac{(12.6)^2}{(9.4)^2} = 1.8 > F_{0.05,\,79,\,79} = 1.5 \qquad (47)$$

Based on this we conclude that the variance has indeed increased. We cannot reach this conclusion for the cases of medium and low currents.

6.4.3 Discussion

The analysis of the EPROM circuit, experimental verification and the explanation of the results was completed over the course of one year. There were important insights gained into the EPROM circuit and the DFM process as a result of this project. This section will discuss three issues surrounding the DFM method, including issues with the verification experiment, the process model, the effect of mismatch, and interactions between the DFM tool and the circuit designer.

6.4.3.1 The Verification Experiment

The verification experiment was designed to vary only the channel length parameter and the circuit topology. The oxide thickness was kept constant to keep the cost of the experiment low. In light of the effect of oxide thickness on performance, the experiment should have been extended to include oxide thickness variation, as was done in the experiment to build the performance prediction model. Since the oxide thicknesses in a given has little variation, additional variation should have been introduced to study the design change.

6.4.3.2 The Process Model

The process model used in this work extends to cover the range of variation seen in production, and uniform distributions were used to ensure even coverage over this entire space. The issues with this model focus on how well it represents unlikely combinations of process parameters.

During the initial Monte Carlo simulation of the EPROM, the results showed a larger percentage of failures than were seen on the manufacturing line. The failures in the simulation were traced to instances where there was a large mismatch between the cell and the n-channel device currents. Further, as these device types have highly correlated channel lengths, this mismatch occurred when the two independent oxide thicknesses were at their

opposite extremes. Comparing the variance in the oxide layer thicknesses used in simulation with manufacturing data, it became clear that the range specified by the fabrication line model was larger than what was currently being produced by the manufacturing line. The process and design engineers involved resolved the issue by decreasing the specified variation in the oxide thickness, and the failure rate seen in simulation was reduced to more closely match manufacturing data.

This ease with which this problem was solved illustrates two advantages of our DFM methodology. First, the use of process parameters with physical meaning allowed the process and design engineers to identify and solve the problem. In contrast, many industrial modeling schemes use oxide thickness as a fitting parameter, which would have made it impossible to identify the problem or communicate so effectively between groups. Also, the integration of the CIM system which made manufacturing data readily available was critical to the rapid identification and solution of this problem.

In addition, this problem is important in light of the failure in the DFM experiment. This incident showed the importance of oxide thickness mismatch in causing the sense amplifier to fail. DFM algorithms must account for how frequently the process exhibits a condition, as there is little value in improving a circuit failure based on a very infrequent process condition. The use of uniform, rather than gaussian distributions, increases the visibility of unlikely process parameter combinations. Although uniform distributions worked well in developing the performance prediction model and were successful in finding circuit sensitivities, they over-emphasize potentially infrequent process conditions. This was seen in the increased failure rate for the EPROM for the potentially infrequent case of oxide thickness mismatch. Gaussian distributions must also be used in the DFM tool, after a sensitivity is identified, to gauge its importance through the use of yield prediction.

One way to solve the problem with these distributions is to use bootstraping methods. To bootstrap the simulation, samples would be drawn from the empirical distribution rather than using a random number generator and a theoretical distribution. In other words, actual samples from the characterized process would be used in the simulation. As long as enough historical data was available, using the bootstrap would cover the process space of the manufacturing line and add importance to each sample because it could be a potential yield limiting point. Further work in this area should explore the trade-offs in the distributions used by the DFM tool.

6.4.3.3 Transistor Mismatch

The problem with the initial simulations of the sense amplifier, mentioned in Section 6.4.2, was a problem with the Monte Carlo software which mistakenly introduced a mismatch between the memory cell being accessed and the reference cell. Recall that the sense amplifier has a current source based on a reference cell which compensates for cell current variation across the array. The circuit requires that the cell being accessed and the reference cell be closely matched in their characteristics.

The Monte Carlo software incorrectly introduced a difference between the reference cell and the accessed memory cell transistor models in a given simulation. Figure 6-14 shows the Monte Carlo simulation result for the sense amplifier window including mismatch. It can be seen that the level of the lower window is sensitive to the mismatch causing the window to collapse at low cell current. Although the mismatch in the simulation of these transistors was larger than what would be seen on the manufacturing line, a problem with mismatch has since been identified in manufacturing, confirming the simulation result.

It was noted in Chapter 3 that this work did not include the mismatch between transistors. Clearly, in light of the problems with mismatch, it is important to characterize mismatch and include its effect in future simulations of the EPROM circuit. Future work

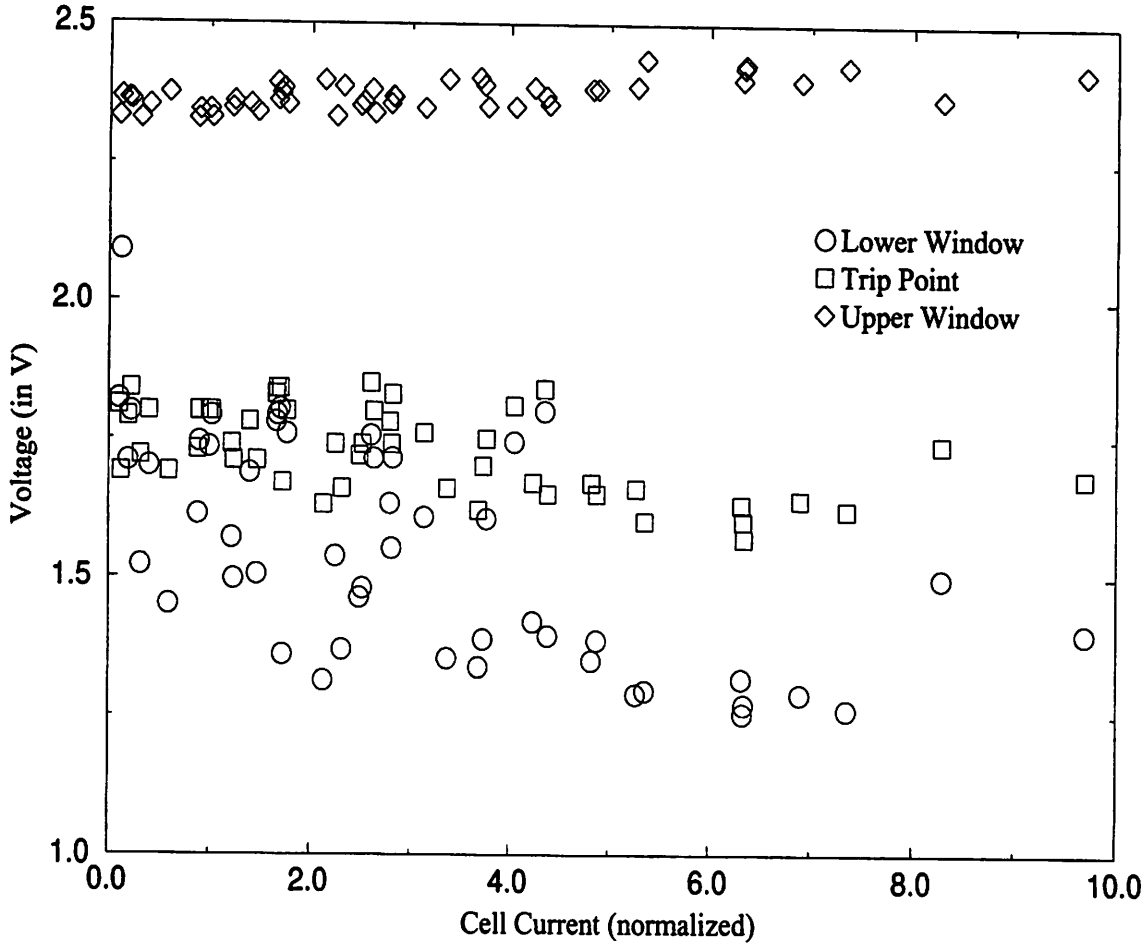


Figure 6-14 Monte Carlo Simulation Including Mismatch: Window Levels for Sense Amplifier - Original Design

should include a characterization of mismatch and an exploration of circuit changes to minimize its impact.

6.4.3.4 Interactions Between the DFM Tool and the Circuit Designer

This section will discuss two problems related to the interaction between the DFM software and the circuit designer. The first problem was that the designer created a new transistor model for the reference cell. This model was identical to the existing cell model but had a new name. As the DFM software keys on the name of the model, the reference cell was not correctly modifying, contributing to the mismatch problem in the software.

Although the physically based DFM methodology presented in this work encourages the circuit designer to understand and vary the SPICE transistor models, this problem indicates that limits should be imposed on the designer. For example, changes to the transistor models should be communicated to the DFM tool support personnel, and the software should warn the user of the presence of an unrecognized model.

The second interaction between the DFM tool and the designer involved the new design changing the correlation structure between the different transitions. Often, after working with a design, the circuit designer establishes a subset of possible transitions which represent the worst-case transitions of the circuit. These transitions become the focus for improving the design. Especially when running Monte Carlo simulations, the number of cases must be limited to decrease the simulation time. In this experiment, the improved design slowed down a transition which was not being analyzed. The complexity surrounding the DFM software prevented the designer from seeing the assumptions underlying the tool, and therefore this problem was not detected until after the verification experiment. This issue needs to be addressed in future work.

6.5 Summary

This chapter developed three manufacturability applications - a performance prediction model, a control chart for product performance and a circuit DFM analysis. The first two applications can be used to improve the capability of the manufacturing line. The DFM analysis contributed to a greater understanding of the effect of the EPROM cell characteristics on low voltage operation, leading to a circuit improvement in the word line driver. Although this particular change might have been obtained through the use of worst case models, the methodology presented here led us directly to the problem. However, the insight into the sense amplifier could not have been achieved without the Monte Carlo simulation. The circuit sensitivity to oxide variation would never have been seen with only "fast" and "slow" simulations.

The initial analysis of the circuits took approximately three months and the manufacturing of the experiment required another three months. While the experiment was being fabricated, the circuit designer started work on a complete re-design of the EPROM for low voltage operation. After testing the circuit change, it was clear that the DFM modifications would not bring enough performance yield improvement. Even though this was a negative result, there was a large amount of experience gained in the DFM process which was applied to the new design. In addition, it was important to verify the limits of improvements which could be made in small modifications of the sense amplifier, justifying the time required for the re-design process.

Chapter 7

Conclusions and Future Work

This research has developed a methodology for modeling the manufacturability of IC products. The method utilizes a fabrication line description based on a few key measurable process parameters to explain the performance variation seen in IC manufacturing. These process parameters are used for a physically based device model parameter extraction, which describes transistor performance over the range of the process. The combination of the fabrication line description and device model enables circuit simulation which maps the process space into the performance space, supporting the construction of DFM applications. This fabrication line and device modeling methodology was applied to a high volume 1.2 μm CMOS EPROM process, and results were presented.

Given these underlying models, three applications were developed. First a performance prediction model was presented which uses electrical test measurements to predict the performance of manufactured parts. The model built for the access time of the 1 Mbit EPROM had a standard deviation of regression of less than 2% of the access time of the product. The model can use manufacturing line measurements to predict performance early in the manufacturing cycle.

This model was utilized in a second application, a model-based Statistical Process Control for product performance. This control chart can detect shifts in the process or problems with the testing of the final product.

The third application addressed the manufacturability of the EPROM under low voltage operation. The product suffered from low parametric yield over a range of the process space. The circuit design was investigated and two critical subcircuits were identified, the wordline driver and the sense amplifier. The product was redesigned and experiments were run to test the new circuits. The design change to the wordline driver resulted in improved margins.

This chapter will discuss several important directions for future work. These areas are improvements in fabrication line measurement and characterization capability, device modeling, and the software infrastructure for DFM.

7.1 Process Characterization

Measurement capability, both in-line and at electrical test, is an area of ongoing research. Accurate in-line measurement capability will allow the performance prediction model to be used much earlier in the production cycle. Today, optical thickness measurements suitable for gate oxides are available. However, optical linewidth measurements are either imprecise or expensive. Given the high correlation between polysilicon linewidth and effective channel length, an accurate in-line measurement for polysilicon would be highly advantageous for in-line binning estimation, allowing early prediction of product performance for fabrication line scheduling and packaging decisions.

At electrical test, further investigation is needed into the measurement of the effective channel length and the source and drain resistance of submicron LDD devices. In particular, improved methods must be able to work through a probe card on a production line and correlate well with device currents. The accurate measurement of channel length is important for effective physical device models.

Another important area for continuing research is the characterization of mismatch. The final analysis of the sense amplifier clearly shows the importance of mismatch in analog circuits. Memory arrays are particularly sensitive to mismatch because reference devices are located at relatively large distances from the matched device. In general, the methods for characterizing mismatch have required detailed characterization which is not suitable for production line monitoring. Future work in the area of mismatch characterization should focus on test patterns and analysis techniques capable of monitoring mismatch on the manufacturing line.

7.2 Device Modeling

There are currently inaccuracies in the SPICE MOSFET models which must be addressed. One issue of primary concern to analog designers is accurate small signal parameters [69], such as transconductance (g_m) and output resistance (r_O). These problems have started to be addressed in the new BSIM3 model [61]. In addition, current research on a methodology for BSIM3 model parameter extraction is addressing the physical mapping of process parameters to the device fitting parameters such as mobility reduction and channel length modulation [70]. The addition of physically based parameter extraction methods to widely used device modeling tools is critical for the applicability of this DFM methodology.

Although the characterization described in this chapter was sufficient for this process, a submicron process with LDD MOSFETS will require additional parameters to adequately account for the variation in performance of products built using those processes. First, due to the shrinking of the device size and the addition of the resistive LDD region, the resistance of the source and drain regions requires improved characterization. The low doping level in the LDD regions implies not only higher resistance, but also that this resistance is a strong function of the applied gate voltage as the region accumulates charge.

The changes in source and drain resistance with V_G are best described within the device model. In addition, their statistical variation must be measured and accounted for.

Another problem is related to the fact that current SPICE transistor models do not accurately model subthreshold current. In an effort to maintain current capability with low gate voltage, IC manufacturers will be pushed to lower V_T, hence increasing the importance of subthreshold current. Accurate subthreshold modeling is required for two reasons. First, circuit designers are biasing MOS devices in this region to take advantage of the higher gain in this region. Second, subthreshold current describes one important component of leakage current, which must be properly accounted for to maintain low power consumption in an IC.

An extension of the statistical characterization pursued in this work would be to develop statistical macromodeling for subcircuits. The IC industry is making increased use of standard subcircuits, such as standard cells for digital applications and pre-defined operational amplifiers, comparators, etc., for analog applications. These subcircuits are combined, often times in automatic synthesis tools, to produce specific functionality [71][72]. It would very useful to statistically characterize and verify the performance of these subcircuits. First, this would improve simulation time, as the subcircuits would not require detailed simulation. More importantly, manufacturing verification of subcircuit performance would potentially reduce worst-case performances which result from unlikely combinations of process parameters.

7.3 Software Infrastructure for DFM

This project utilized several manufacturing databases, relied on schematics generated from the circuit CAD tools, used the HSPICE circuit simulator and RS/1 statistical package and spanned several computer systems. The DFM software developed in this project interfaced with all the necessary packages. However, the DFM system required user familiarity with several subsystems, each with its own interface. Further integration is neces-

sary to provide a user friendly system, including providing common interfaces and moving the algorithms inside the existing applications. Software integration and ease of use is the most important area for improvement in future research.

7.3.1 Process Simulation for DFM

Process simulation was not applied because of the difficulty in tuning the simulator to the specific manufacturing process used in this work. However, the potential advantages of process simulation include allowing the manufacturability modeling to begin before the process is running on the fabrication line, and a large reduction in cost by reducing the number of wafers manufactured to characterize the process and verify DFM results.

Simulation and analysis tools, such as PDFAB [73], are beginning to address the problems with the process and device simulation described in Chapter 2. For example, these tools integrates process and device simulation results with manufacturing data, making it easier to match an actual process. These analysis tools also perform physically based parameter extraction. The so-called "virtual wafer fabrication line" allows a user to easily simulate a process across a wide of parameter variation and extract transistor models for use in circuit simulation.

In addition, process simulation frameworks are starting to emphasize manufacturing applications and support statistical methods such as Principal Component Analysis and linear regression. Examples of the applications include support for the development of performance prediction models and producing diagnostic information from the analysis of electrical test data [73]. These applications will extend the use of process simulators into product engineering, in addition to their use in process development and device modeling groups.

7.3.2 Infrastructure for Performance Prediction Modeling

Performance prediction modeling could also be further integrated with statistical analysis software. Integrating such software with manufacturing databases would provide the user with simple but powerful commands to combine in-line, electrical test and performance data for a set of lots. The resulting datasets could then be analyzed using statistical techniques such as Principal Component Analysis and linear regression. Also, rather than separate design and manufacturing interfaces, a single interface would most directly provide the ease of use necessary to support both the manufacturing and design communities.

7.3.3 Software Tools for Circuit DFM

There were several problems with the software tools developed for this project. Firstly, the software was difficult to use, requiring the user to be an expert in both the programs utilized to perform DFM and the algorithms inside the DFM software. Secondly, the tool requires intelligent input from the circuit designer to interpret the Monte Carlo results, find the process sensitivities, identify critical transistors or subcircuits, and finally modify the design and verify the changes. Thirdly, unless the user has established SPC to automatically verify the mapping from process space to performance space, these complex relationships must be verified by other, manual methods.

To alleviate these problems, the DFM software algorithms should be implemented inside a circuit CAD software system. An integrated CAD/DFM system should support the fabrication line description and transistor models, run Monte Carlo simulations and allow the user to view the results. This integration must include intelligent validation steps to separate the designer from the algorithms and methods inside the DFM tool.

In addition, the tools must bring a better interface with the circuit designer. The software developed in this work suffered from a complex user interface, and did not give the user enough feedback on which process parameters or transistors were causing the performance sensitivities. Daoud [16] presents an improved interface for DFM, and Hocevar

[74] presents a good example of an improved user interface for a circuit optimizer. A simple, usable interface for these complex tools is critical for user acceptance of DFM. The designer should be able to easily specify circuit performance metrics to examine, and the tool automatically extract the performance from the simulation output.

Given a sensitivity in circuit performance, the DFM software tool should automatically identify the process parameters and the transistors or subcircuits which underlie this sensitivity. Research continuing from this project intends to use the Monte Carlo circuit simulation results to accept or reject hypotheses about the design. For example, the designer could hypothesize that a certain process parameters was the cause of a circuit sensitivity, or that a specific transistor should be modified for improved manufacturability, and the DFM software would accept or reject this hypothesis. The goal is for the DFM tool to automatically find the important process parameters and circuit elements, rather than rely on the expertise of the circuit designer.

References

[1] P. Yang, D. Hocevar, P. Cox, C. Machala and P. Chatterjee, "An Integrated and Efficient Approach for MOS VLSI Statistical Circuit Design," *IEEE Trans. on CAD,* Vol. CAD-5, pp 5-14, Jan. 1986.

[2] T.K. Yu, S.M. Kang, J. Sacks, W.J. Welch, "Parametric Yield Optimization of MOS Integrated Circuits by Statistical Modeling of Circuit Performances," *Technical Report No. 27*, Dept. of Statistics, University of Illinois, July, 1989.

[3] S.R. Nassif, A.J. Strojwas, S.W. Director, "FABRICS II: A Statistically Based IC Fabrication Process Simulator, *IEEE Trans. on CAD*, Vol. 3, No. 1, pp 40-47, January 1984.

[4] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

[5] W. Maly, A.J. Strojwas and S.W. Director, "VLSI Yield Prediction and Estimation: A Unified Framework," *IEEE Trans. on CAD*, Vol. 5, No. 1, pp 134-150, January 1986.

[6] K.K. Low and S.W. Director, "A New Methodology for the Design Centering of IC Fabrication Processes," *IEEE Trans. on CAD*, Vol 10, No. 7, pp. 895-903, July 1991.

[7] W. Maly, "Computer-Aided Design for VLSI Circuit Manufacturability," *Proceedings of the IEEE*, Vol. 78, No. 2, pp. 356-392, Feb. 1990.

[8] A.J. Strojwas, editor, *Selected Papers on Statistical Design of Integrated Circuits*, IEEE Press, 1987.

[9] M. Rencher, "Analog Statistical Simulation," *IEEE 1991 CICC*, pp 29.2.1-29.2.4, 1991.

[10] M.J.B Bolt, J. Engel, C.L.M. v.d Klauw, M. Rocchi, "Statistical Parameter Control

for Optimum Design and Manufacturability of VLSI Circuits," *Proc. ISMSS*, Burlingame, pp 99-106, 1990.

[11] J.P. Spoto, W.T. Coston, C.P. Hernandez, "Statistical Integrated Circuit Design and Characterization," *IEEE Trans. on CAD*, Vol. CAD-5, No. 1, pp. 90-103, January, 1986.

[12] P.K. Mozumder, *Statistical Quality Control for VLSIC Fabrication Control*, Ph.D. thesis, Carnegie Mellon University, Report No. CMUCAD-89-44, July 1989.

[13] W. Davis and R. Ida, "Statistical I.C. Simulation Based on Independent Wafer Extracted Process Parameters and Experimental Designs," IEEE BTCM-89, 262-265.

[14] H.P. Tuinhout, S. Swaving, J.J.M. Joosten, "A Fully Analytical MOSFET Model Parameter Extraction Approach," *Proc. ICMTS*, pp 79-84, Feb. 1988.

[15] D.E. Hocevar, P. Cox and P. Yang, "Parametric Yield Optimization for MOS Circuit Blocks," *IEEE Trans. on CAD*, Vol. 7, No. 6, pp. 645-658, June, 1988.

[16] Z. Daoud, *DORIC: Design of Optimized and Robust Integrated Circuits*, Master's Thesis, UC Berkeley, ESRC 93-22/CSM-06, December 1993.

[17] T.K. Yu, S.M. Kang, I.N.Hajj and T.N. Trick, "iEdison: An Interactive Statistical Design Tool for MOS VLSI Circuits," *Proc. ICCAD*, pp. 20-23, Nov. 1988.

[18] J.C. Zhang and M.A. Styblinski, "Design of Experiments Approach to Gradient Estimation and its Application to CMOS Circuit Stochastic Optimization," *Proc. ISCAS*, Singapore, pp 3098-3101, 1991.

[19] J.B. Brockman and S.W. Director, "Predictive Subset Testing: Optimizing IC Parametric Performance Testing for Quality, Cost, and Yield," *IEEE Trans. on Semiconductor Manufacturing*, Vol. 2, No. 3, August 1989, pp 104-113.

[20] L. Milor and A. Sangiovanni-Vincentelli, "Optimal Test Set Design for Analog Circuits," *Proc. ICCAD*, pp 116-119, 1990.

[21] E.D. Boskin, C.J. Spanos, G. Korsh, "IC Performance Prediction from Electrical Test Measurements," *Proc. ISMSS*, pp 13-17, June 1992.

[22] S.P. Cunningham, et al, "Use of In-line Yield Estimates in Production Control," ORSA Conference, Phoenix, AZ, Nov. 1993.

[23] W. Davis, presentation at ISSCC Discussion Session "Statistical Design: Competitive

Weapon or Designer's Nightmare?," February 1992.

[24] M.J.M Pelgrom, A.C.J. Duinmaijer and A.P.G. Welbers, "Matching Properties of MOS Transistors," *IEEE Journal of Solid-State Circuits*, Vol. 24, No. 5, pp 1433-1440, October 1989.

[25] C. Michael and M. Ismail, "Statistical Modeling of Device Mismatch for Analog MOS Integrated Circuits," *IEEE Journal of Solid-State Circuits*, Vol. 27, No. 2, pp 154-166, February, 1992.

[26] J.K. Kibarian, *Statistical Diagnosis of IC Process Faults*, Ph.D. thesis, Carnegie Mellon University, Report No. CMUCAD-90-52, December 1990.

[27] C. Yu, *et al*, "Use of Short-Loop Electrical Measurements for Yield Improvement," accepted for publication in *IEEE Trans. on Semiconductor Manufacturing*.

[28] TMA SUPREM III User's Manual, TMA Inc., August 1990

[29] S. Selberherr, A. Schutz, H.W. Potzl, "MINIMOS - a Two-Dimensional MOS Transistor Analyzer," *IEEE Trans. on Electron Devices*, ED-27, pp 1540-1550, 1980.

[30] J.A. Power, A. Mathewson and W.A. Lane, "MOSFET Statistical Parameter Extraction Using Multivariate Statistics," *Proc. ICMTS*, pp 209-214, March 1991.

[31] J.A. Power, A. Mathewson and W.A. Lane, "An Approach for Relating Model Parameter Variabilities to Process Fluctuations," *Proc. ICMTS*, pp 63-68, March 1993.

[32] E. Boskin, R. Chen, Z. Daoud, H. Liu, C. Spanos, *IC Design for Manufacturability I*, Memorandum No. UCB/ERL M92/17, February 1992.

[33] S. Liu and K. Singhal, "A Statistical Model for MOSFETs," *Proc. ICCAD*, Santa Clara, pp 78-80, 1985.

[34] L. Milor and A. Sangiovanni, "Optimal Test Set Design for Analog Circuits," *Proc. ICCAD*, pp 116-119, 1990.

[35] R.J. Harris, *A Primer of Multivariate Statistics*, Academic Press, 1975.

[36] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, Methuen and Co. Ltd., London, 1964.

[37] N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley and Sons, Inc., 2nd edition, 1981.

[38] G.A.F. Seber, *Linear Regression Analysis*, John Wiley and Sons, Inc., 1977.

[39] G.E.P. Box, W.G. Hunter and J.S. Hunter, *Statistics for Experimenters*, John Wiley and Sons, Inc., 1978.

[40] S. Leang, *Supervisory Control System for a Photolithographic Workcell*, Master's thesis, University of California, Berkeley, Memorandum No. UCB/ERL M92/70, July 1992.

[41] P.K. Mozumder and G.G. Barna, "Statistical Feedback Control of a Plasma Etch Process," *IEEE Trans. on Semiconductor Manufacturing*, Vol. 7, No. 1, pp 1-11, Feb. 1994.

[42] D.C. Montgomery, *Introduction to Statistical Quality Control*, John Wiley & Sons, Second Edition, 1991.

[43] E.D. Boskin, C.J. Spanos, G.Korsh, "A Method for Modeling the Manufacturability of IC Designs," *Proc. ICMTS*, pp 241-246, March 1993.

[44] C.J. Spanos, *Statistical parameter extraction for IC process characterization*, Ph.D. thesis, Carnegie Mellon University, Research Report No. CMUCAD-85-46, 1985.

[45] *Workstream Users Manual*, Consilium, Inc., 1989.

[46] *RS/1 Users Guide*, BBN Software Products Corporation, 1990.

[47] N. Ingersoll, *KEITHLEY Data Extraction Program*, ATMEL Corporation, 1992.

[48] N. Ingersoll, *EDE Extract Program Documentation*, ATMEL Corporation, 1993.

[49] *HSPICE User's Manual*, Meta-Software, Inc., 1990.

[50] K.A. Brownlee, *Statistical Theory and Methodology In Science and Engineering*, John Wiley & Sons, Second Edition, 1965.

[51] G.S. May, *MOSTCAP - An MOS Transistor Characterization and Analysis Program*, Master's Thesis, University of California, Berkeley, January 1988.

[52] J.G.J. Chern, P. Chang, R.F. Motta and N. Godinho, "A New Method To Determine MOSFET Channel Length," *IEEE Electron Device Letters*, Vol. EDL-1, No. 9, pp. 170-173, Sept., 1980.

[53] J. Ida, A, Kita, and F. Ichikawa, "A New Extraction Method for Effective Channel Length on Lightly Doped Drain MOSFET's," *Proc. ICMTS*, pp 117-122, March 1990.

[54] S. Nakanishi, *et. al.*, "Effective Channel Length Determination Using Punchthrough Voltage," *Proc. ICMTS*, pp 73-77, March 1992.

[55] R.A. Ashton, P.A. Layman and C.C. McAndrew, "Modeling and Characterization of MOSFET Width Dependencies," *Proc. ICMTS*, pp 133-137, March 1993.

[56] R. Muller and T. Kamins, *Device Electronics for Integrated Circuits*, John Wiley & Sons, 1986.

[57] D.K. Schroder, *Semiconductor Material and Device Characterization*, John Wiley & Sons, 1990.

[58] D.W. Felbaumer and D.K. Schroder, "MOSFET Doping Profiling," *IEEE Trans. on Electron Devices*, Vol. 38, No. 1, pp 135-140, January 1991.

[59] W. de Lange, "Accurate Determination of CMOS Capacitance Parameters using Multilayer Structures," *Proc. ICMTS*, pp 57-61, March 1992.

[60] R. Shrivastava and K. Fitzpatrick, "A Simple Model for the Overlap Capacitance of a VLSI MOS Device," *IEEE Trans. on Electron Devices*, Vol. ED-29, No. 12, pp 1870-1875, December 1982.

[61] J.H. Huang, Z.H. Liu, M.C. Jeng, K. Hui, M. Chan, P.K. Ko and C. Hu, *BSIM3 Manual*, Version 2.0, March 1994.

[62] P.K. Ko, "Chapter 1: Approaches to Scaling," *Advanced MOS Device Physics, VLSI Electronics: Microstructure Science, Vol. 18*, Academic Press, 1989.

[63] A. Kolodny, S.T.K. Nieh, B. Eitan, J. Shappir, "Analysis and Modeling of Floating-Gate EEPROM Cells," *IEEE Trans. on Electron Devices*, Vol. ED-33, No.6, pp 835-844, June 1986.

[64] J.A. Power, R. Clancy, W.A. Wall, A. Mathewson and W.A. Lane, "An Investigation of MOSFET Statistical and Temperature Effects," *Proc. ICMTS*, pp 202-207, March 1992.

[65] J.M.Chambers, W.S. Cleveland, B. Kleiner and P.A. Tukey, *Graphical Methods for Data Analysis*, Wadsworth & Brooks, 1983.

[66] C.Y. Chao and L. Milor, "Performance Modeling of Analog Circuits Using Additive

Regression Splines," *Proc. CICC*, pp. 301-304, 1994.

[67] A.P. Chandrakasan, *Low Power Digital CMOS Design*, Ph.D. thesis, University of California, Berkeley, Memorandum No. UCB/ERL M94/65.

[68] G. Korsh, ATMEL Corporation, personal communication.

[69] Y.P. Tsividis and K. Suyama, "MOSFET Modeling for Analog Circuit CAD: Problems and Prospects," *IEEE Journal of Solid State Circuits*, Vol. 29, No. 3, pp 210-216, March 1994.

[70] J. Chen, University of California, Berkeley, personal communication.

[71] J.P. Harvey, M.I. Elmasry and B. Leung, "STAIC: An Interactive Framework for Synthesizing CMOS and BiCMOS Analog Circuits," *IEEE Trans. on CAD*, Vol. 11, No. 11, pp 1402-1417, November 1992.

[72] H.Y. Koh, C.H. Sequin and P.R. Gray, "OPASYN: A Compiler for CMOS Operational Amplifiers," *IEEE Trans. on CAD*, Vol. 9, No. 2, pp 113-125, February 1990.

[73] *PDFAB Users Manual*, PDF Solutions, Pittsburgh, PA, 1992-1994.

[74] D.E. Hocevar, et. al., "A Usable Circuit Optimizer for Designers," *Proc. ICCAD*, pp 290-293, Nov. 1990.

Appendix A

Monte Carlo Software Manual

MC - the Monte Carlo Circuit Simulation Package

Version 2.0

A.1 Introduction

MC is a UNIX software package, written in shell scripts and the C programming language, which generates, runs, and analyzes the results for a Monte Carlo circuit simulation. The Monte Carlo simulation varies the process parameters in the transistor models, allowing the user to see the change in the performance of the circuit over the fabrication line variation seen during manufacturing. MC must be customized to account for the specifics of a given circuit. This document will describe both use and the customization of the MC software package.

MC utilizes HSPICE, and therefore must be run on a computer with an HSPICE license. The transistor models must be physically based transistor models, so that changing the values of the physical parameters will accurately represent the changes in the device characteristics. In general, MC will be run on a circuit towards the end of the design phase, to ensure it behaves properly over range of variation it will see in the fab.

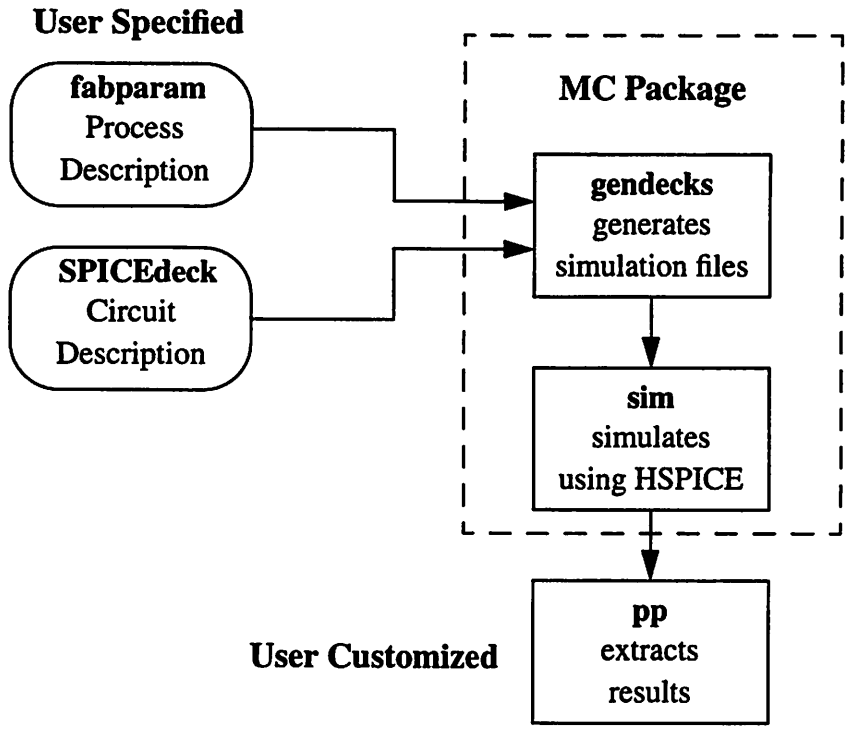


Figure A-1 MC Software Modules

Monte Carlo simulation is not meant to be used early in the design phase, when many different signals must be examined.

The general flow of the Monte Carlo software system is shown in Figure A-1. First, the specified number of HSPICE input files are generated from the template file. Specific items such as the device models and specific circuit elements are modified to simulate process variation. Next, the simulations are run, generating performance results. Finally, the results of the simulation are summarized from the HSPICE output. After running MC, the results can be analyzed graphically or using a statistical package.

A.2 User Manual for the MC Software Package

This section will describe how to run the MC software package.

A.2.1 Preparing to run the Simulation

MC should be run with a separate directory for each run. As MC uses fixed filenames for the large number of simulation files, separate directories are necessary. Monte Carlo simulations can use up a large amount of disk space, so the user should delete simulations after they have been superceded or are no longer needed.

MC requires an input file named "SPICEdeck." This file must be a "flat" representation of the circuit including all subcircuits and models. A flat representation implies only that there are no "included" files. It does NOT imply that there are no subcircuits, which are acceptable. If the circuit is normally split among several files, simply combine all the files into one file and comment out the include statements. The transistor models must be in this single file, and not be brought in with a ".lib" command in HSPICE. It does not matter which model (i.e. slow, fast, nominal) is brought into the deck, as the MC program will alter the model parameters.

Before running the Monte Carlo, the SPICEdeck should be edited to set the temperature desired for the simulations. Statements which generate large output files should be commented out, for example, ".option post=1" statements, which generate the graphical output files. If graphics are desired, a few of the Monte Carlo simulations which represent interesting cases should be run with this option enabled, in order to get a detailed look at signals. Whenever possible, ".measure" statements should be used to produce the specific desired result, and make it easy to obtain the desired specification from the simulation.

In addition, it will often be appropriate to modify the input signal transitions and length of a transient analysis, in order to produce only the worst case specification or transition, rather than all the transitions normally examined during design. The key thing to remember is that many simulations will be run which generate a lot of data - try and make it manageable.

A.2.2 Running the Monte Carlo

The Monte Carlo is run using the MC command. MC calls a series of programs which generate the spice decks for simulation, run the HSPICE simulations, and then post-process the results. The command line format is:

mc <-s sample count> <-i id> <-m max jobs> <-p post processor (pp) string>

MC can be run with any combination, or none of the options. Below are the arguments to MC.

-s ***sample count***

Specifies the sample count, that is, how many simulations to run in the Monte Carlo. Sample count should be an integer. As rules of thumb, 20 is a small number, 100 is a large number. The default is 50.

-i ***id***

Specifies the id, which is the seed for the random number generator. The ability to specify the seed will allow you to run several simulations at exactly the same points in the process, so you can directly compare the results. If no id is specified, MC will pick one for you. Typically, the first simulation will be run with no id specified and MC will generate and print the id chosen. Further simulations will be run with the id specified as the value printed in the first simulation. The id should be an integer between 1 and 30000.

-m ***maxjobs***

Specifies the maximum number of jobs to run in parallel on the HSPICE server. The default is 2. If MC is running on a multiple processor machine, it is not advisable to run more jobs than the number of processors at a time. Running less than that is advisable to be courteous to your fellow HSPICE users.

-p ***post-processor string***

Specifies the post-processor to extract results from the simulation output file. MC expects the custom post-processor to have the name pp***string***, i.e. pp1, pp2, ppfalltime, *etc.*, it is this number which is specified here.

A.3 Customization of the Package

MC was created to be flexible, however, most circuits will require customization of the code to properly simulate the effect of process variation on your circuit.

Important Note: Rather than customize the MC package, the reader should consider utilizing the Monte Carlo features within HSPICE. The ".param" statement can be used to both create samples of process parameters from gaussian or uniform distribution, and also to specify other variables, such as coupling ratios or fitting parameters, as functions of those parameters. The outputs from ".measure" statements can then be graphed in HSPLOT.

There are four main categories of changes which may be required. Each category will be discussed in detail in the subsections below.

1. *Process variation* - MC requires the means and standard deviations of the critical process parameters to generate the Monte Carlo simulation files.
2. *Transistor model names* - MC modifies the transistor models based on the model names. Currently, MC uses six transistor names from an EPROM process.
3. *Circuit elements* - In addition to varying the transistor models, there are other circuit elements which need to vary to reflect process variation. Examples include the coupling ratios used to model the EPROM cell, and lumped capacitors which model the effect of many MOS gate capacitors on an address, bitline, wordline or data line.
4. *Performance extraction* - After the Monte Carlo simulations are completed, the circuit performance of interest, for example access time or D.C. operating point, will be con-

tained within the HSPICE output files. The user must customize a simple shell script which extracts the performance metric from the output file.

A.3.1 Process Variation

The default process parameters (and the corresponding HSPICE MOS Level 3 model parameter names) varied by MC are change in channel length (xl), change in channel width (xw), threshold voltage (vt0) and oxide thickness (tox). The mean and standard deviation of each of these parameters for each transistor type is specified in a file. The MC software simulates +/- 3 sigma from the mean, covering the process space with a linear distribution of these parameters. The process variation file is currently named "fabparam.18[1]" and is included in the source file inputs.c. The program inputs.c contains an include statement which must be modified to refer to the proper filename.

A.3.2 Transistor Models and Circuit Elements

This is the most difficult part of the package to customize. Again, the user should investigate the use of ".param" statements to include this variation in the HSPICE circuit or device models, which may simplify this customization.

The names of the transistor models are declared in the file "gendecks.c." The source code must be modified to recognize the models used in the circuit simulation input file and be consistent with the names in the process description file.

To handle the variation of circuit elements, C language code must be written. Examples of circuit elements which are functions of the process parameters are lumped resistors and capacitors, like word or bit line models, and coupling ratios in floating gate subcircuits. The basic algorithm behind changing circuit elements is that the program recognizes these elements, and replaces the value, such a resistance, with a number that the program

1. The "18" is meant to be an abbreviation for the process, for example, fabparam.18 specifies the ATMEL 18000 process. This is just notation, any filename will work.

calculates. In order for the user to customize MC, the user must write a segment of code which performs this function.

As a historical note, the coupling ratios in the EPROM cell model used to be circuit elements calculated in the program, as described above. However, in the current cell models[2], the coupling ratios are automatically calculated within the HSPICE file through the use of ".param" statements. In fact, the ".param" statements are powerful enough to capture more of the modeling functions. The user should consider using this method to calculate other circuit element values directly in the SPICE deck.

A.3.3 Performance Extraction

After running HSPICE, the performance parameter of interest, such as access time or the DC sense amp window levels, are contained in the HSPICE output files. Usually, the best way to convey the manufacturability information is through plots graphing the performance of interest against a process or circuit parameter, such as oxide thickness or cell current. Therefore, a simple script should be written which extracts the parameter from the HSPICE output file.

This script will most likely be written using the UNIX utilities awk, grep and sed, rather than the C programming language. This script is run as a post-processor to the HSPICE runs, and hence the existing performance extraction routines are called "ppn," where n is a number. This naming convention also simplifies running the post-processor, as the -p switch to MC specifies the number, n, of the post-processor to run. The existing post-processors in should serve as sufficient examples for writing new performance extraction scripts.

---

2. The models were produced by T. Randazzo.

A.3.4 Compiling the Software

During the customization of the MC software, the C language files "inputs.c" and "gendecks.c," and the include file "fabparam.xx" were likely modified, and therefore the software must be re-compiled. The "fabparam.XX" and "inputs.c" are in one directory and "gendecks.c" in a second. There are "Makefiles" in each directory, therefore, type "make" in the "inputs.c" directory, then type "make" in the "gendecks.c" to re-compile. Even if "gendecks.c" was not modified, you must also type "make" in the gendecks directory to rebuild the high level executable. Finally, copy the "gendecks" executable into the "bin" directory. The rest of the MC software package consists of UNIX scripts which do not require compilation.

A.4 MC Utilities

There are several utilities available which perform support functions for the Monte Carlo analysis.

graphmaker

Graphmaker takes columns from several files and combines them into a single file of columnar data for plotting. Graphmaker prompts the user for the name of the file, and then the column to take from the file. If the filename given is "inputs," the program assumes it is the file of process parameters produced by MC, and the user is prompted for device type and/or process parameter name, and maps the name to the appropriate column in the file. Graphmaker terminates when an empty filename is given (just hit return), or when the limit of four columns is reached. The output is in the file named "plot.out."

Known bug: all files must be in the current directory.

xvgr

xvgr is free software which plots columnar data files.

A.5 Site Specific Details

The program source, executable programs and example are contained on olympus under the directory ~eboskin/spice/code. In this directory, there are four subdirectories: bin, doc, gendecks, and inputs. The bin directory contains the executable programs and should be added to your unix path. The doc directory contains documentation, such as this file. The gendecks and inputs directories contain the source code.

The directory ~eboskin/spice/code/bin also contains the other shell scripts which from the MC software package, and example post-processors. The directory ~glen/bin4 contains the xvgr plotting utility, and many other useful packages.

Appendix B

Performance Modeling Software Manual

Electrical Test, In-Line and Performance Data Collection and Analysis

Version 2.1

B.1 Introduction to the Performance Analysis System

A set of programs has been developed to tabulate and analyze in-line, electrical test and performance data collected during the manufacturing of products. The main purpose of the analysis is to generate models which predict the performance of manufactured parts from the in-line and electrical test data collected during the manufacture of the product. The software system utilizes the RS/1 statistical package and programs which access the manufacturing databases. The databases required for performance analysis include the electrical test, in-line data and final test databases.

This manual describes the important software modules needed, and includes the step-by-step procedure for creating a performance prediction model from in-line and electrical test data. In general, the user will be generating tables of data in RS/1, and then using RS/1 to analyze the data and build the performance prediction model. This manual, after several introductory sections, will discuss how to build a table of in-line data, how to build a table of electrical test data, how to then combine the tables with performance data and use

RS/1 to build the model. There are several sections at the end of this Appendix which briefly summarize some useful VMS (the VAX operating system) and RS/1 commands.

B.2 Computer Access at ATMEL

This manual is specific to the software and hardware used in the ATMEL CIM system and is being included as an Appendix to this thesis for completeness. The general reader can use this Appendix to get an idea of the implementation of a performance analysis system, but few specific details will apply outside of ATMEL.

At ATMEL, the performance modeling software runs on the VAX named "CEDAR" in Colorado springs. The program used to access the electrical test database is the Keithly Database Extractor (KDE), which extracts data from the Keithly Electrical Test Database. The Engineering Database Extractor (EDE) allows users to access the database produced by the Workstream software system running in the fab, which contains the in-line data. Both KDE and EDE were written and supported by Nelson Ingersoll. The manuals for these programs should be obtained before using the programs.

The performance prediction model is created using data and programs on the VAX. To access the data and run the programs, one must have an account on the VAX and access to the SJ_ENG, KDE and EDE RS/1 group homes.[1] Mary Zawacki is the person to contact to obtain an account and the group home access privileges. A manager's approval is required to obtain the account.

To access CEDAR, you can use a lat port on a PC (there is a "public" lat port in the PC area using the PC Print Station 2), or, you can connect to CEDAR from a window on a UNIX workstation. To get to the VAX from the lat port, use the program ST240 to "connect" to RS/1. (Connecting to RS/1 through the lat port logs the user onto CEDAR.) From UNIX, you can rlogin to CEDAR. Enter your user name and password at the CEDAR LOGIN prompt, and you'll enter a list of menu choices to access the utilities on CEDAR.

1. The group home is a collection of RS/1 procedures and tables that can be accessed by members of the group.

Type VMS at the list of commands to get to VMS, where you can run all the necessary software.

B.3 Conventions Used in this Manual

In the command summaries given in this Appendix, commands following a "$" are entered at the VMS prompt, while commands beginning with "#" are entered in RS/1. Within RS/1, group home procedures begin with the '#' character. Commands beginning with ">" are entered in KDE or EDE. Capital letters will be used to denote computer programs, files and keywords.

B.4 Collecting In-Line Data

In-line data is collected periodically from Workstream and entered into RS/1 tables. To access this data and supporting software, you need the VMS privileges to access the EDC group home. The in-line data is accessed through an RS/1 procedure called autopilot. Autopilot uses a template to pull the in-line data from the engineering database, by calling EDE. Finally, you will use the RS/1 procedure AP2TABLE, which will convert the output of autopilot to the form needed for building the prediction model. You will collect 5 key in-line parameters: the two polysilicon CDs, the active CD and two gate oxide thicknesses.

The step by step instructions are given below.

1. Select the EDE group home and enter RS/1.

   Command Summary: $ setrs edc
   $ rs1

2. In order to collect in-line data, you must create an autopilot template. This template must be created only once, and then can be copied or modified for all subsequent use. Given an autopilot template, the autopilot procedure can be run. First, the lot numbers must be specified, and then autopilot will be started to collect the in-line measurements for

those lots. Within autopilot, use "Modify Entry" (command 3) to specify the desired lots. Select the row of the autopilot template file (if there is only 1 template, it will be row 1) and then modify the lot number (line 7), specifying the desired lots separated by commas. Finally, EXIT back to the autopilot main menu, and Start the Autochart (command 6).

Command Summary: # call #autopilot
Follow the menu within autopilot.

3. Autopilot produces a file with the averages of the in-line data for the specified lots. Next, the averages must be pulled from the autopilot output, and converted to a table which can be easily combined with electrical and performance data. The procedure to generate the table is called AP2TABLE, and it resides in the SJ_ENG group home. To access the procedure, use the following commands within RS/1:

Command Summary: # call $setrs("SJ_ENG")
call #ap2table

This procedure is hard-wired to produce a file called "inline_avg". This file must be renamed or deleted to run AP2TABLE - AP2TABLE will refuse to overwrite this file.

B.5 Collecting Electrical Test Data

Electrical test data, collected at the end of wafer processing, is contained in the Keithly Electrical Test Database. Similar to EDE, the program KDB extracts specific electrical test results for the desired lots from the database. The step by step instructions are listed below.

1. Electrical test data is available in the Keithly Database using KDE. First, obtain the file FOURMEG.KDB. This file will be used to set up KDE with most of the correct defaults. To obtain the file, copy it from Eric Boskin's home directory.

Command Summary: $ copy [eboskin]fourmeg.kdb fourmeg.kdb

2. KDE is run to extract the data. Within KDE the user must read in the defaults from the file FOURMEG.KDB using the "@fourmeg" command. You will need to change the dates

to include the time when your lots were processed. Although this may be done by editing the file FOURMEG.KDB, it is simpler to use the BEFORE command once in KDE. Also, the required lots to be extracted must be selected using the LOT command and the date set. Alternatively, use COUNT DAYS *n*, to look back *n* days from today. Use the LIST command to make sure all data is entered correctly. LIST should print out the lots you entered, with the dates of the lots. Finally, extract the data using the EXTR command. This produces a file names KDBEXTRACT.TMP as a default.

Command summary:
```
$ kde
> @fourmeg
> before 1/20/95
> lot 233012,233056,...
> list
> extr
> exit
```

3. Set the RS/1 group home to KDB, and start up RS/1.

Command Summary:
```
$ setrs kdb
$ rs1
```

4. Next, read the data into RS/1. The program KDBREAD defaults to a VMS input file called KDBEXTRACT.TMP, and produces an output table (within RS/1) called KDBTABLE. Use the default answers to the questions by hitting return after the question.

Command Summary:
```
# call #kdbread
```

Note: you may get a warning that your table contains EMPTY values. This usually means that not all electrical test values were taken at all sites on all wafers and can be ignored.

5. Change to the SJ_ENG RS/1 group home, which will allow you to access a program to average the electrical test data. The program #kdb2rs1 is hard-wired to read an input table called KDBTABLE (the output of KDBREAD) and output a file called AVERAGES.

Command Summary:
```
# call $setrs("sj_eng")
# call #kdb2rs1
```

At this point, you have created a file called AVERAGES, which is the average of each electrical test parameter for each lot.

6. To exit RS/1, type LOGOUT.

B.6 Collecting Performance Data

Next, the performance data from final test should be collected. Currently, the performance file must be entered manually. The file should be created in RS/1 with the MAKE TABLE command and the editor (answer E to create the table using the editor) and must be called PERFORM. Simply, the arrow keys move you around the table, and at any table cell just type in a value to put that value in the table. See the RS/1 manuals for details. Save your table and exit by typing /EXIT.

Important: you must put the lot number in column 1 (not column 0). You must not have any empty rows in this file. (If you do create a blank row during the creation of the table, perhaps the last row will be blank, delete it by positioning the cursor anywhere in the blank row and type /DELETE, then answer ROW and confirm.)

Command Summary: # make table perform

B.7 Combining the Data Files for Analysis

At this point, the performance, in-line, and electrical test data for each lot must be combined into one file. Run the procedure JOINALL in the SJ_ENG group home. This procedure uses the fixed filenames INLINES, AVERAGES and PERFORM, and produces a file named JOINDATA. The tables also have specific columns for the lot number (PERFORM in column 1, INLINES in column 0, and AVERAGES in column 1).

Command Summary: # call #joinall

B.8 RS/1 Analysis

This section will briefly describe five applications which are run within RS/1. See the RS/1 user manual for more details on the use of these commands.

1. To examine the data before regression, it is interesting to look at the correlation structure of the data. RS/1 will calculate the multivariate correlation structure of the table JOINDATA with the $mvcorr procedure.

Command Summary: # call $mvcorr

2. The in-line and electrical test data is a highly correlated data set. Therefore, the user may desire to perform a Principal Component Transformation on the data. The user must specify to write the transformed data into a table (the RS/1 default is to not save the transformed data).

Command Summary: # call $princo

To build a performance prediction model, a stepwise regression should be run. The user will enter variables until the program suggests stopping. After a regression, examining the residuals is usually quite useful.

Command Summary: # fit multiple

The "fit multiple" command will ask you for the table portions for the independent and dependent variables. The independent variables are the electrical test and in-line parameters, the dependent variable is the performance. A model can only be created for one performance at a time. To select table portions, responses should look something like:

independent: cols 4 to 21 of joindata
dependent: col 2 of joindata

The exact column numbers will depend on how many variables you have (that is, the number of performances in the PERFORM table and the number of in-line and electrical test parameters). If a Principal Component transformation has been done, the table with the Principal Component scores should be used instead of JOINDATA.

3. Another useful feature is the ability to make graphs. The easiest way to make a graph is to enter the menu mode, and follow the hierarchy of commands to make a graph. Type

Summary: # delete tablename

7. A table can be edited with the EDIT command. This will allow the user to modify entries in the table, or add rows or columns. In the editor use the arrow keys to move around in the table. The command /HELP will list the other available commands, such as deleting rows or columns and exiting the editor.

B.11 Printing Tables

To print out a table or graph from within RS/1, create a file and then transfer the file to California using kermit or ftp. There are two ways to create a file. To create an ascii file, use the command $ez_writefile. Answer the questions as you see fit. The author usually uses the defaults except:

a. Include row 0 in the printout (column headings).
b. Write using only spaces, not tabs, spaces and commas.

Command Summary: # call $ez_writefile

To retain the format of the table, use $printout, that is, $printout will keep the columns of your tables aligned.

Command Summary: # call $printout

Good choices for printers are postscript, lpt72, lpt80 or lpt132. Postscript generates a postscript file and the lpt printers are generic ascii files of width 72,80 or 132 (for landscape). Do not give a system command or your file will print in Colorado. A file called PRINT.TXT will be produced which you can send transfer to California and then print.

B.12 Transferring files to California

Files on the VAX may be transferred locally in two ways. They can be transferred to UNIX using FTP, or they can be transferred to the PC using KERMIT.

To use FTP, the user must PUT the file. This puts the file onto unix, where you can use lpr to print the file.

Command summary:
$ ftp olympus (it will ask for a password)
ftp> put filename
ftp> bye

To use kermit, type kermit, then type server. After typing server, type ALT-Y, which will bring up a kermit menu. Use the GET command, then type finish followed by an ALT-Y when the transfer is compete.

B.13 Customizing for another Product Line

This software is customized for use on an EPROM in the AT18000 family. There is no configuration file to drive the software, rather, the programs must be edited to work on another product. In general, the specific process numbers and steps must be modified for each product. However, these changes are relatively minor, and all the table forming, averaging and joining will work on any dataset. Specifics are given below.

1. Electrical test data: The file FOURMEG.KDB contains the process type and front end. For the EPROM, it is AT18000. These lines must be changed for another process. Also, the specific electrical test parameters are listed. These could be changed to be any of the over 100 electrical test measurements.
2. In-line data: The autopilot template must be changed to list the proper process and process front end. If other in-line measurements are desired, you must find out the appropriate step name and number.