

# Yield modeling for deep sub-micron IC design

***Citation for published version (APA):***

Simon, P. (2001). *Yield modeling for deep sub-micron IC design*. [Phd Thesis 2 (Research NOT TU/e / Graduation TU/e), Electrical Engineering]. Arts & Boeve Publishers. <https://doi.org/10.6100/IR556679>

***DOI:***

[10.6100/IR556679](https://doi.org/10.6100/IR556679)

***Document status and date:***

Published: 01/01/2001

***Document Version:***

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

***Please check the document version of this publication:***

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

***General rights***

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

***Take down policy***

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

# Yield Modeling for Deep Sub-Micron IC Design

Paul Simon





# **Yield Modeling for Deep Sub-Micron IC Design**

# **Yield Modeling for Deep Sub-Micron IC Design**

Proefschrift

ter verkrijging van de graad van doctor  
aan de Technische Universiteit Eindhoven,  
op gezag van de Rector Magnificus, prof.dr. R.A. van Santen,  
voor een commissie aangewezen door het College voor Promoties  
in het openbaar te verdedigen  
op donderdag 20 december 2001 om 16.00 uur

door

**Paul Simon**

geboren te Voorschoten



Dit proefschrift is goedgekeurd door de promotoren:

prof.ir. M.T.M. Segers

en

prof. W. Maly

Cover design: Elly van Domburg, Van Domburg Ontwerp, Nijmegen

Cover photos: Physical Characterization Group, Crolles and FA group MOS4YOU

The work described in this dissertation was carried out at MOS4YOU waferfab, Philips Semiconductors, Nijmegen, The Netherlands, as part of a joint research program with Carnegie Mellon University Pittsburgh and Technical University Eindhoven

© Copyright 2001 P.L.C. Simon

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission from the copyright owner.

Published by Arts & Boeve Publishers  
P.O. Box 31187  
6503 CD Nijmegen  
The Netherlands

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Simon, Paul  
Yield modeling for deep sub-micron IC design / by Paul Simon.  
Eindhoven: Technische Universiteit Eindhoven, 2001.  
Proefschrift.

Headings: VLSI, semiconductor manufacturing, VLSI design, yield modeling, DfM

ISBN 90-75341-28-8  
NUGI 832

*Voor Loes, Daniel, Nina*

*Jules en Tineke*



---

# Preface

This thesis is the result of almost five years of work done in the area of DfM and yield engineering at the MOS4YOU wafer fab of Philips Semiconductors in Nijmegen, The Netherlands. Next to being a very stimulating environment to work and learn in, MOS4YOU and its people have given me a fantastic time. Although my name appears on the cover of this dissertation, without the help and support of many people, both technical and moral, the book would not have seen daylight. Here I would like to take the opportunity to express to them my deepest appreciation and thanks.

In particular, I would like to thank:

Professor Wojciech Maly, my thesis supervisor and teacher, for his warmth and understanding throughout these years. During his many tiring visits to Nijmegen I was able to dig in the mine of his great experience and, finally, understood the main issues of DfM. His guidance was a great source of inspiration. The time we worked together was a great pleasure and has a lasting value for me.

Dr. Paul van Wijnen, for his companionship, enthusiasm, and for giving me the opportunity to combine the work at MOS4YOU with the preparation of a Ph.D. thesis.

Professor Rene Segers and Professor Jochen Jess of the Technical University of Eindhoven, to whom I am indebted for reading and correcting the manuscript and giving me the opportunity to carry out the project in cooperation with the university.

Dr. Dirk de Vries, and Dr. Jan-Marc Luchies, my ‘paranymphs’, for their company, moral support and many invaluable discussions on yield modeling, plasma damage, and gastronomy. From both I learned a lot.

Ir. Kees Veelenturf, my manager during most of the time I spent at MOS4YOU, for always supporting me in my work, and for defending an academic approach in the sometimes ‘fire fighting’ oriented fab organization.

Drs. ir. Clemens van den Berghe, for his everlasting positivism, encouragement, support and relaxing company during many coffee breaks.

---

ir. Stanley Sprij and ir. Marc van de Pol former roommates, for their pleasant company and jokes.

Klaas Arts for his advice on the layout of this thesis and the help with printing and publishing.

Dr. Diederik fokkema, ir. Paul Volf, ir. Arjan Mels, and ing. Tony Jurg of the DfM group in MOS4YOU for their company, support in development of tools, and preaching the DfM word.

All my co-workers in MOS4YOU and LTG, they made my stay in Nijmegen a very pleasant one.

Ing. Ronald de Bruijn and ing. Roland Antheunis who were of great importance for helping to develop the layout extraction tools and designing test structures respectively.

My wife, Loes for her moral support and understanding during these difficult years and my children Daniel and Nina. To them and to my parents I dedicate this thesis.

Paul Simon  
La Terrasse, September 2001



---

# Contents

Preface.....	vii
<b>1. Introduction .....</b>	<b>1</b>
1.1. Introduction .....	3
1.2. Thesis outline .....	5
<b>2. Yield Modeling Principles .....</b>	<b>7</b>
2.1. Introduction .....	9
2.2. Yield loss: causes, classification and characteristics .....	11
2.3. Yield model overview .....	16
2.3.1. Basic yield models .....	18
2.3.2. Yield models based on design attributes other than area.....	21
2.3.3. Critical area yield model derivatives .....	25
2.3.4. Yield model for any design attribute .....	27
2.3.5. Experimental comparison of yield models .....	29
2.4. Conclusions .....	33
<b>3. Yield Prediction Methodology and Model Parameter Extraction .....</b>	
3.1. Introduction .....	39
3.2. Yield prediction methodology .....	39
3.3. Process parameter extraction.....	40
3.3.1. Test structure based yield model parameter extraction .....	42
3.3.2. Other considerations for test structure implementations.....	50

---

3.3.3. Examples of test structures for parameter extraction .....	51
3.4. Design parameter extraction: structural layout characterization .....	58
3.4.1. Practical extraction techniques .....	59
3.4.2. Extraction toolbox .....	60
3.4.3. Examples of design attribute extraction algorithms.....	66
3.5. Development of a manufacturability assessment environment (MAE) .....	69
3.5.1. Motivation for the development of an MAE .....	70
3.5.2. The Mapex-II system.....	73
3.6. Conclusions .....	78
3.7. References .....	80

## **4. Plasma Process Induced Damage: Physics and Modeling**

4.1. Introduction.....	85
4.2. Charging failure mechanism .....	86
4.2.1. Charge imbalance mechanisms .....	88
4.2.2. Layout dependency of charging .....	93
4.3. Modeling charging induced yield loss .....	95
4.3.1. Charging induced yield loss experiment .....	96
4.3.2. Modeling charging related yield loss .....	99
4.3.3. Conclusions and discussion .....	100
4.4. Plasma process induced damage characterization .....	101
4.4.1. Conventional charging test structures .....	101
4.4.2. MAM test structures .....	102
4.4.3. Characterization of the layout dependence of charging for 0.35 $\mu\text{m}$ and 0.18 $\mu\text{m}$ processes .....	107
4.4.4. Conclusions .....	114
4.5. Robust IC design for charging .....	116
4.5.1. Conventional approach .....	116
4.5.2. Sensitivity index model for plasma damage during metal etch .....	120
4.5.3. Quantifying design sensitivity to charging.....	125
4.6. Conclusions .....	126
4.7. References .....	128



---

<b>5. Design for Manufacturability in VLSI</b>	131
5.1. Introduction	133
5.2. A common language between design, manufacturing and test	134
5.3. DfM methodology	137
5.4. DfM in IP design	139
5.4.1. Standard cells and memories	140
5.4.2. Synthesis for high yield	145
5.4.3. Routing	147
5.5. DfM in the manufacturing environment	153
5.5.1. Process development	153
5.5.2. Understanding product variability	156
5.5.3. Yield prediction and sensitivity analysis	168
5.6. DfM in test development	173
5.7. Current R&D needs for DfM implementation.	173
5.8. Conclusions	175
5.9. References	176
 <b>6. Summary</b>	 177
 <b>7. Samenvatting</b>	 183
 List of abbreviations	 189
List of publications	191
About the author	193

---

# **Chapter 1**

## **Introduction**

# 1

## 1.1 Introduction

Manufacturing yield has always been an important parameter determining the economic viability of any semiconductor company. Due to the present directions the semiconductor market is moving in, this has become even more so [1]. In order to be able to follow Moore's law over the past few decades, the costs associated with developing and manufacturing VLSI products have grown tremendously. Nowadays, modern manufacturing sites are built at very high costs (2-3B\$) and rapid return on investment is essential for the economic viability of the business. The ability to achieve high yield at an extremely fast rate has consequently become a crucial factor that decides whether a company is successful, or risks to go out of business. It can be easily calculated that a small amount of unnecessary yield loss for a modern semiconductor fab easily translates to the loss of hundreds of millions of dollars per year in terms of revenues from manufactured products and lost manufacturing volume.

Another reason for the increasing significance of a fast yield ramp is the change with respect to the need for ever-faster product introductions and shorter product lifetimes. Driven by market needs, more products are developed that tend to have shorter lifetimes resulting in narrow windows of opportunity. Therefore immediate high yield is vital for products that are introduced into a fab. There is no time for costly re-engineering of the design or process. Sometimes even the first lot of a product is not out of the line when the last lot is put in. It is obvious that in such a case smooth product introduction and high, and even more important, predictable yield, are crucial.

Product yield is not only a function of quality of a manufacturing process, but also of the sensitivity of the design to the failure mechanisms that are present at the time the product is going through the manufacturing steps. Along these lines process development engineers tend to argue: "the yield of a product is determined by how well the design fits the manufacturing process". This is of course true. However, a designer's usual reply is: "as long as the product is

designed according to design rules, the yield is determined by how well the process is capable of accommodating all design attributes”.

This example shows well how far design and manufacturing have been drifting apart. Over the past few years both disciplines have increased dramatically in complexity. For both lines of work very differently skilled engineers are needed, and, in addition, activities often take place in different organizations, often situated at distant locations. Therefore communication is very much formalized and, in some cases, over-simplified. As a result, the responsibility for product yield is not shared, nor is the information that is necessary to generate robust designs. Often a simple design manual is the only source of manufacturing information that is used to design a product. This separation of product design and the process development has caused an increased probability that designs do not optimally fit the manufacturing processes.

From the above arguments one can derive the following list of fundamental needs which today's and future IC design and manufacturing technologies should address:

- As feature size will decrease rapidly and more complex processes are being used, not only more failure mechanisms, but also more complex ones will play an increasingly important role. However, conventional test structures and in-line monitoring techniques are no longer adequate tools to characterize all possible failure mechanisms in a process. **For both rapid yield learning and robust IC design a methodology is needed that enables the characterization of the manufacturing process with respect to all possible defects.**
- For each of the identified defect types, it is then necessary to quantify the yield impact on product level so that its importance can be evaluated. Therefore **yield models are needed that take into account both the design sensitivity and defect characteristics of the process.** Without such models it is difficult to quantify the yield impact on product level and consequently to set priorities in yield improvement efforts. Yield models that describe layout sensitivities are also needed in order to design products that are as insensitive as possible to the identified failure mechanism.
- In order to be able to apply yield models to IP design (standard cell, memory, IP blocks or products) and to predict their yield capability, **a methodology for extensive design characterization is needed.**
- The detachment of the design and manufacturing communities over the past few years has led to a communication structure that is inadequate to address the yield problems that are present in modern VLSI technologies. Therefore **a reintegration platform for design and manufacturing activities is necessary.**

- **For robust IC design, methodologies, models, tools, and data are needed in order to be able to assess or influence the yielding capability of a product on *all* levels of design abstraction.**

The above needs were also recognized at MOS4YOU, CMOS waferfab of Philips Semiconductors in Nijmegen, Carnegie Mellon University, Pittsburgh and the Technical University of Eindhoven. Consequently a joint research project was started to address the above needs. This document reports the results that were collected during a period of almost five years. The accomplishments reported here are the result of the cooperation in many teams that brought together engineers from many different disciplines of the above organizations.

## 1.2 Thesis outline

This thesis is structured as follows:

In *Chapter two* an overview and classification of yield loss causes in IC manufacturing is given. Existing yield models are reviewed and the benefits and limitations are described. New yield models are proposed and verified using extensive manufacturing data.

*Chapter three* presents methodologies for yield model parameter extraction. Both process parameter extraction and design characterization methods are discussed. A rapid yield learning and yield prediction methodology are described and the associated costs of yield model calibration are discussed. The development of an industrial product characterization platform is described.

In *Chapter four* the plasma induced damage mechanism and its layout dependence are studied. Plasma damage is used as a case study to show an implementation of the methodologies developed in chapter 3. Special new test structures are developed and their results are presented. A clear relationship between plasma damage and yield is found and subsequently a yield model is derived and verified with experimental data. Also a new method to quantify product charging sensitivity is proposed in order to achieve a methodology for charging robust design.

*Chapter five* shows that for better manufacturability it is necessary and possible to better integrate design, process development, and test. A “common language” between these disciplines is proposed that enables faster yield ramp and robust design. Several examples in both the design and manufacturing environment show the effectiveness of the methodology. Recommendations for further research in this field are given.

Finally *Chapter six* summarizes the research conducted.

### Reference

- [1] W.Maly, “High levels of IC Manufacturability: One of the Necessary Prerequisites of the 1997 SIA Roadmap Vision”. IEDM 1998

---

## **Chapter 2**

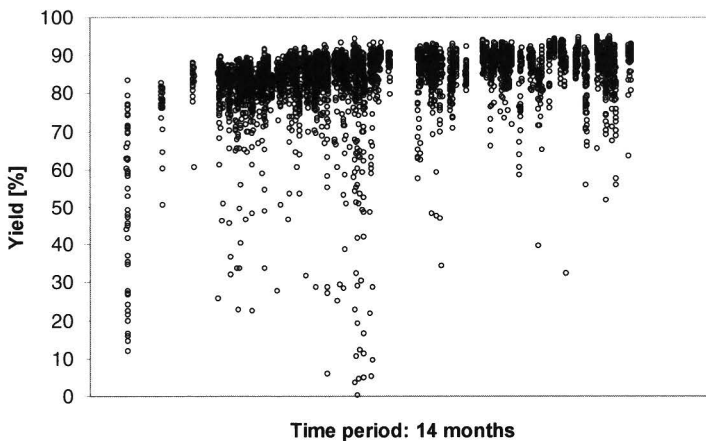
# **Yield Modeling Principles**

# 2

## 2.1 Introduction

Although the total cost of a VLSI product is only partly determined by the silicon manufacturing costs, the level of its profitability is largely determined by the yield that is achieved during the manufacturing process. Increased yield loss results in fewer functional devices at the same manufacturing costs. Sometimes inefficient wafer usage, wafer damage or miss-processing may be important contributors to the yield loss. But, in general the most important contributors to yield loss are failures caused by local unintended product-process interactions. This chapter focuses on product yield, which can be defined by the ratio of the number of working devices and the total number of devices that are tested.

Figure 2.1 shows a characteristic wafer yield trend of a particular product that was produced in a time period of approximately one year. It is clear from this figure that the yield level and its variance change over time. Every wafer is subject to hundreds of different processing steps on many different machines.



**Figure 2.1** Product wafer level yield trend over a period of 14 months



In each IC layer a large number of failure phenomena with different probabilities of occurrence and different levels of “killing potential” may take place. Hence, it is obvious that in today’s complex process technologies it is naïve to assume that yield loss is caused by only one failure mechanism. Therefore the identification of the major failure mechanisms in a manufacturing process is often a complicated task that requires dedication and detailed knowledge on both the process imperfections and the associated product sensitivities. Throughout the period of yield ramping much effort is put into the continuous debugging of the manufacturing process in order to increase the yield and tighten its distribution. It has been well understood that models that describe failure mechanisms and their relevance for product yield are essential to effectively manage continuous yield improvement activities, and the setting of priorities therein [1, 2-4].

Many types of yield models have been derived over the years and have proven to be useful for many purposes such as:

### ***Inside manufacturing environment***

- Yield ramping and setting priorities for corrective actions
- Understanding the impact of different types of failure mechanisms on product yield
- Understanding yield differences between products
- Assessing and predicting the yield impact of process changes or options

### ***Outside manufacturing environment***

- Reporting of process yield capability parameters such as average defect density,  $D_0$ .
- Planning of manufacturing volume
- Assessing the yield impact of decisions that are taken during the design phase of a product

Depending on the application, existing yield models use different (sets of) parameters. However, all yield models contain parameters that characterize both the manufacturing process and the design. Design related parameters may describe, for example, device area, critical area, number of transistors, number of nets, total length of conductors etc. Process parameters may describe defect density, defect size distribution, defect density distribution, defect clustering, line width distribution or layer thickness. In most attempts at yield modeling so far, the focus has been on the yield models themselves and not on the calibration or extraction of the yield model parameters. But, a user of any yield model must always realize that the data that is used to acquire yield model parameters is as important as the accuracy of the model itself. When it is not clear how to obtain the necessary parameters, questions can be raised on the accuracy, reproducibility, and applicability of these models. This is the reason why many yield models are viewed with suspicion and skepticism, even though they may be derived based on sound statistical principles.

This chapter describes the benefits and limitations of existing yield models. In section 2.2 examples of yield loss causes and a yield loss classification is described. In section 2.3 an overview of the most common yield models is given. Yield

models based on device area and other design attributes are discussed. The accuracy of existing and new yield models is compared using extensive amounts of manufacturing data. In section 2.4 conclusions are drawn.

## 2.2 Yield Loss

### Causes, Classification and Characteristics

Wafer productivity loss can have many causes, some of which are obvious and others are more difficult to trace. To begin with, silicon is lost because of *inefficient wafer usage*. Often at the edge of the wafer, parameters such as layer thickness or defectivity are not within process specifications. Also some equipment may use the edge of the wafer for mechanical positioning, damaging certain layers. In addition wafers contain scribe line test structures, alignment markers or structures for in-line monitoring. Such structures require silicon real estate and therefore reduce the wafer productivity.

Secondly, wafers may be damaged during manufacturing due to wafer handling by humans or machines. Often this kind of yield loss is labeled as *line yield loss*. A third category of yield loss occurs when at the end of the manufacturing line the scribe-line test structure results indicate that device parameters are outside manufacturing specifications. In that case product testing costs are reduced by scrapping the entire wafers and the yield loss can be categorized as *PCM (process control module) yield loss*.

Yield loss due to the above reasons is usually much smaller than the *die yield loss* which is the ratio of the number of failing dies and the number of tested dies. Die yield loss is caused by manufacturing imperfections that occur on each wafer. This chapter focuses only on die yield which will simply be referred to as *yield*. In this context it is useful to distinguish between sources of yield loss, events that cause them, defects, failure mechanisms and faults. Figure 2.2 shows the relationships between those terms and shows some typical examples.

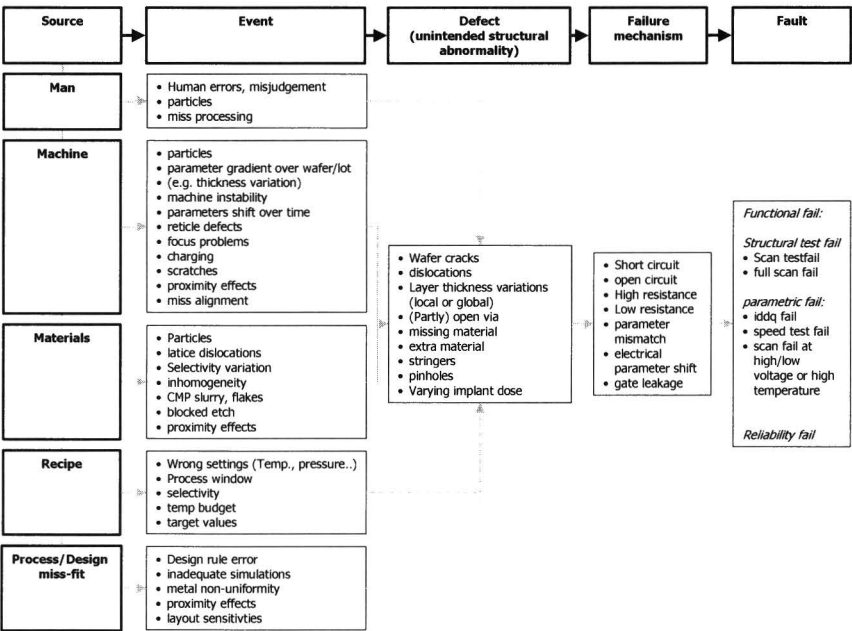
#### Reasons for yield loss

Figure 2.2 shows one of many possible classifications of the relationship between yield loss causes and resulting symptoms. In complex manufacturing processes many sources of yield loss exist. For instance, human errors can never be excluded and will remain an uncontrolled source of yield loss. Yield loss can also be caused by equipment settings that are wrong or drifted because of instabilities. Another important source of yield loss are materials such as the wafers themselves, but also deposited materials such as aluminum or resist may contain particles that cause defects.

Furthermore, the robustness of the process influences the susceptibility to yield loss. Complex interactions between different processing steps and materials sometimes cause a tight “processing window”. In such cases only a minor deviation in any of the relevant parameters or chemical properties (such as selectivity of an etching substance) may cause yield loss.

Due to the increasingly complex manufacturing processes and product designs, subtle product-process interactions are becoming an increasingly important source

of yield loss. Design rule checking no longer guarantees an optimal design-process fit and problems such as cross talk, delay faults and device matching are affecting the yield of VLSI products.



**Figure 2.2** Classification of sources of yield loss, events, defects, failure mechanisms and faults

Below the terms used in the classification of figure 2.2 are discussed in more detail.

### Events

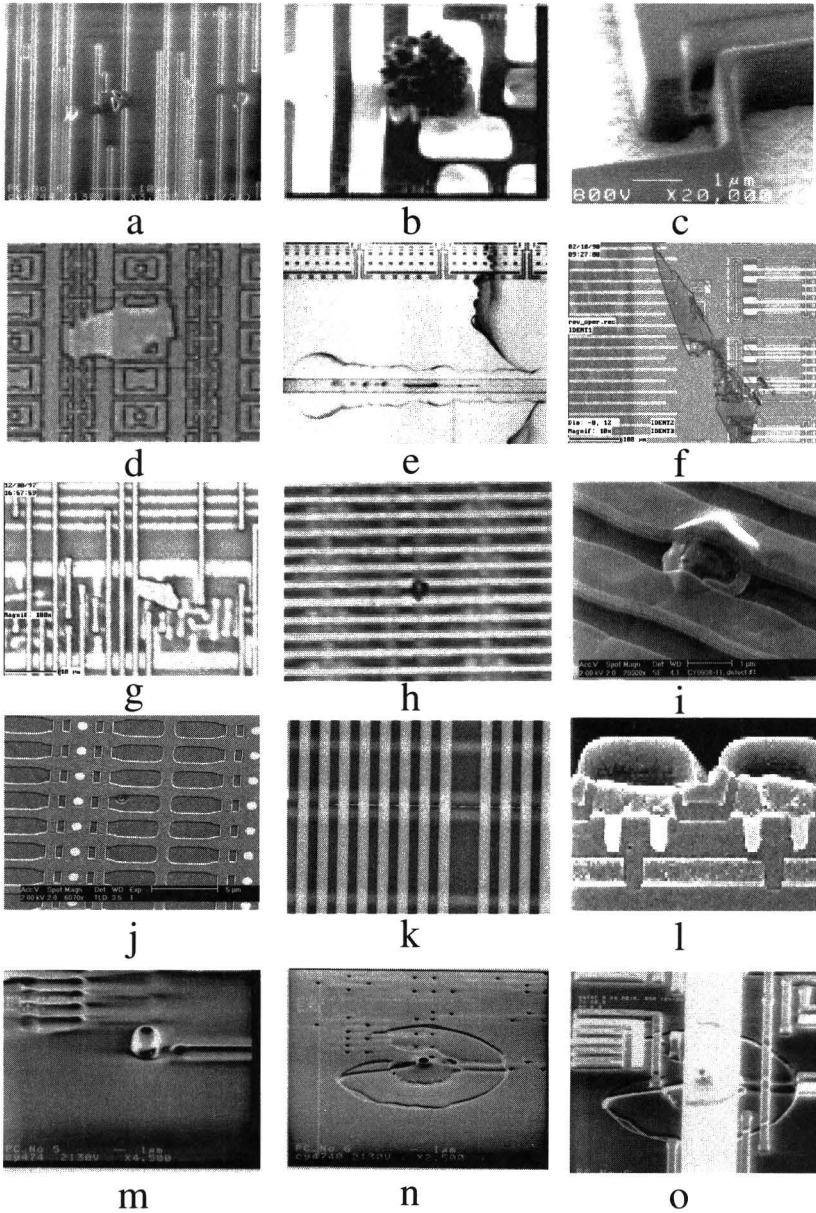
The manufacturing of VLSI products contains hundreds of steps during which many incidents may cause disturbances in the process. An event may occur randomly and its impact may vary from wafer to wafer and from batch to batch. Events or combinations of events lead to unintended structural abnormalities on the wafer. For example the combination of local oxide thickness variation and proximity effects may cause lithographic problems.

From figure 2.2 it also becomes clear that there are much less symptoms that indicate yield loss than there are possible events causing it. Many different events may lead to exactly the same behavior during device testing.

### Defects

In many cases events lead to unintended structural abnormalities on a wafer called defects. Defects may include extra or missing materials that occur predominantly locally. Such defects are therefore often called spot defects. However, global

defects such as thickness or electrical parameter gradients across a batch, wafer or die exist as well. Figure 2.3 shows some examples of defects.



**Figure 2.3** Examples of defects. a: CMP scratch; b: particle; c: bad focus; d: result of a blocked metal etch; e: residual stain; f: flake; g: blocked etch; h: small particle; i: decorated particle; j: gate oxide pinhole; k: stringer; l: open vias; m,n,o: defect integration

**Failure mechanisms**

The failure mechanism describes how a circuit failure results from a defect. For instance, extra or missing material may lead to open or shorted circuits resulting in a scan test fail. The same defect may trigger different failure mechanisms. For example, a gate oxide pinhole may cause gate leakage resulting in Iddq failure. A similar gate oxide pinhole may cause a shift in the threshold voltage of a transistor causing a delay fault.

It is important to realize that *not all defects*, or even the majority of defects, do *result in a failure*. Defects may, for example, occur in areas where they do not affect the structures that determine the functionality of the IC. In some cases ICs are designed in such a way that they are insensitive to certain types of defects by application of for example circuit redundancy or robustness to electrical parameter shift by statistical design methodologies [6,7]. Therefore often the term “killing defects” is used to distinguish defects that have an impact on yield from those that don’t. Often the term “kill ratio” is used to quantify the probability that a defect, which is located on a random location on the die, causes an electrical failure. The kill ratio is therefore determined by defect type, defect size, and product sensitivity.

**Faults**

Whether a failure mechanism leads to a fault, and degrades the yield as such, depends on whether the product test program covers the affected area of the chip. Although scan tests usually have good test coverage, they never test the complete area of the die, nor all the electrical conditions that make the fail observable at the chip’s outputs.

When chips are tested, the yield loss manifests itself in many ways. Therefore yield loss can be classified in many different ways. A few examples are given in the table 2.1.

Classification of yield loss	
Manifestation	<ul style="list-style-type: none"><li>• Functional</li><li>• Parametric / performance</li></ul>
Affected area	<ul style="list-style-type: none"><li>• Local</li><li>• Global</li></ul>
Pattern	<ul style="list-style-type: none"><li>• Random</li><li>• Non random / systematic / gross</li></ul>

**Table 2.1**    *Classification of yield loss*

**Manifestation**

A *functional fault* in an IC is detected when a test vector doesn’t yield the expected result. In some cases the circuit is functional, but does not meet the specifications with respect to accuracy, attainable clock frequency or power due to parametric

variability of the process. In such cases the yield loss can be classified as *performance related yield loss* or *parametric yield loss*. The device may operate at certain operating conditions, but fail tests as a function of a continuous parameter such as supply voltage, current, or temperature.

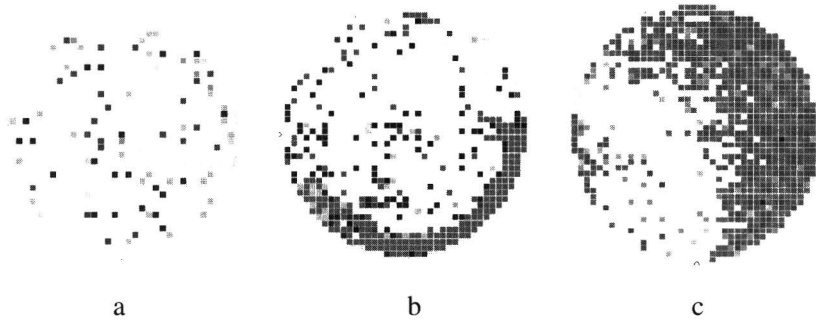
### Affected area

Whereas local yield loss occurs in small areas or points on the wafer, global yield loss is associated with large areas of the wafer such as the wafer center or edge. Local yield loss can for example be caused by random spot defects. Global yield loss can be caused by an electrical parameter gradient over the wafer. See figure 2.4.

### Pattern

Random failures occur anywhere on the wafer in an unsystematic way. Random yield loss can be caused by for example scratches or particles causing shorts or opens. Both local and global yield loss can be random. Non-random failures tend to group or cluster on the wafer.

Systematic yield loss always occurs on the same dies on the wafer and can be caused by for example reticle defects.



**Figure 2.4** Typical wafer maps with random (a) and both random and systematic yield loss (b,c)

## 2.3 Yield Model Overview

Since the beginning of semiconductor manufacturing history yield models have played an important role in solving technical problems related to wafer productivity as well as in predicting yields for strategic decisions related to process development and shrinking. As a result much effort is put in the development and description of yield models over the years. This section shortly summarizes the main yield models that have been reported so far.

If  $A_s$  is the defect-sensitive area of a product, and  $D$  is the average number of killing defects per unit area, then the fault density, often designated as  $\lambda$ , is

$$\lambda = A_s D \quad (2.1)$$

and, if defect density is low, defects are distributed randomly across the wafer, and each failing die is killed by exactly one defect, the yield can be written as

$$Y = 1 - \lambda \quad (2.2)$$

However, in practice multiple killing defects may fall into one product resulting in a better overall yield than predicted by this simple linear model. Under the assumption that defect density is constant and distributed randomly across the wafer, Poisson statistics can be used and the product yield can be derived as [8]:

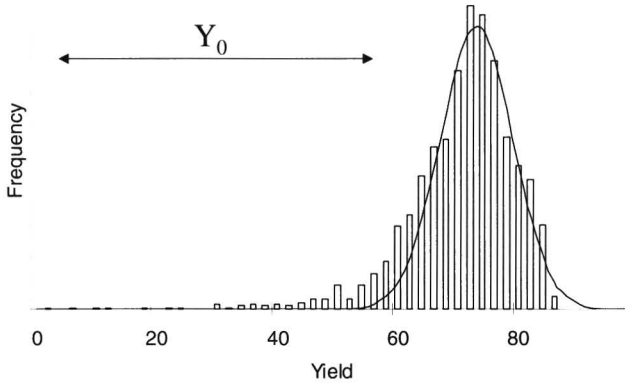
$$Y = e^{-\lambda} \quad (2.3)$$

In the case of non-random or systematic yield loss defects are not distributed randomly across the wafer. In such cases the yield loss can be translated into the loss of a fraction of the wafer area. See for example figure 2.4. Therefore this kind of yield loss can be accounted for in a yield model simply by adding a factor  $Y_0$  which is often referred to as the gross yield loss factor [9,10].

$$Y = Y_0 \cdot Y_{random} \quad (2.4)$$

Systematic yield loss affects the 'tail' of the lower part of the yield distribution as is indicated in figure 2.5 that shows a typical wafer level yield distribution of one product for a large number of wafers. High  $Y_0$  values are usually caused by global defects that originate from for example sensitivity of the circuit to electrical parameter gradients over the wafer. Yield excursions on certain lots or wafers may also cause higher  $Y_0$  values. Such low yielding wafers often pose a reliability risk and therefore manufacturers usually set a cutoff limit for low yielding wafers at which they are to be rejected.





**Figure 2.5** Example of a wafer level yield distribution for approximately 4000 wafers of one product

Whereas  $A_s$  is determined by design attributes such as die area, structure density, minimum design rules etc.,  $D$  is determined by the defect characteristics of the manufacturing process. Therefore yield is always expressed in terms of IC design parameters and manufacturing process parameters:

$$Y = f(\text{Manufacturing process}, \text{IC design}) \quad (2.5)$$

The first yield models were developed for memory applications where the sensitivity to defects is largely determined by the design of the memory cell which is distributed homogeneously over the total chip area [11]. In that case the die area was an adequate design parameter to differentiate different products with respect to sensitivity to defects. Yield modeling was therefore dominated by the search for an accurate statistical description of the defect density distribution on the wafer and from wafer to wafer. This eventually led to the basic yield models of which the benefits and limitations are described in section 2.3.1.

When yield models are used for digital, analog, or mixed signal products with embedded memories, the differences in product sensitivity between products can get quite significant. In such cases other design parameters than just IC area must be considered to predict the yield. Some of the existing models that are based on design attributes are described in section 2.3.2.

### 2.3.1 Basic yield models

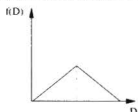
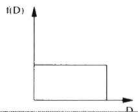
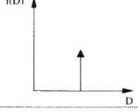
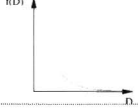

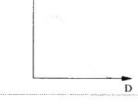
The defect density  $D$ , in a manufacturing process is not constant, but tends to vary from lot to lot, from wafer to wafer and even across the wafer. Many yield models have been proposed over the years that are based on different defect density distributions. To account for the defect density variability in these yield models,  $D$  is summed over all wafers using the following model [12]

$$Y = \int_0^{\infty} f(D) e^{-AD} dD \quad (2.6)$$

Where  $f(D)$  is the defect density distribution and

$$\int_0^{\infty} f(D) dD = 1 \quad (2.7)$$

An overview of the most generally used defect density distributions and the resulting yield models is given in table 2.2.

Model	Compounder function	Yield Formula	References
Murphy (triangular)		$Y = \left( \frac{1 - e^{-\lambda}}{\lambda} \right)^2$	12
Murphy (rectangular)		$Y = \frac{1 - e^{-2\lambda}}{2\lambda}$	12
Poisson		$Y = e^{-\lambda}$	2
Seeds		$Y = \frac{1}{1 + \lambda}$	13,14
Gamma		$Y = \left( 1 + \frac{\lambda}{\alpha} \right)^{-\alpha}$	15
Truncated Gaussian		$Y = \exp \left( \frac{A^2 \gamma^2}{2} - AD_p \right) \frac{1 + \operatorname{erf} \left( \frac{A \gamma - D_p}{\sqrt{2} \gamma \sqrt{2}} \right)}{1 + \operatorname{erf} \left( \frac{D_p}{\gamma \sqrt{2}} \right)}$	16

**Table 2.2** Basic yield models and their defect density distributions.  
 $\lambda$  = fault density,  $f(D)$ =defect density distribution,  $\alpha$  =clustering factor,  
 $D_p$  = peak of distribution

### Murphy's yield model

Murphy was the first to propose a non-constant defect density in manufacturing processes. Murphy's model was the base for the development of many yield models, each assuming a different defect density distribution. To demonstrate the effect that the defect density distribution has on the predicted yield, Murphy tried a triangular and uniform defect density distribution. See also table 2.2.

### Seeds yield model

Because Seeds observed a large number of wafers with low defect density and only a limited number with high defect density, he assumed an exponential defect density distribution as the compounder of his yield model. Seeds himself found that his model overestimated the yield.

### Poisson model

Probably the best-known and most used yield model is the Poisson yield model [2]. The Poisson model uses a delta function as compounder in Murphy's yield integral, which means that it assumes that defects are uniformly distributed and constant across the wafer. Due to the relatively good accuracy and simplicity, [5], the Poisson model is often used for planning purposes and for reporting yield performance trends of manufacturing processes. In an attempt to make this reporting independent of the products that are made in the process, not the yield is reported, but the average defect density. For this purpose the Poisson model is rewritten as:

$$D_0 = \frac{1}{K} \sum_{i=1}^K \frac{-\ln(Y_i)}{A_i} \quad (2.8)$$

Where  $D_0$  is the average density of killing defects and  $K$  is the number of products on which yield is measured. This methodology for defect density reporting can only be adequate if the variance in product sensitivities is small. When the differences between products become too large, more design parameters have to be taken into account.

The Poisson model is known to underestimate the yield of large products. This is a result of the assumption that defects are randomly distributed across the wafer, which is often not the case.

### Negative Binomial model (Gamma defect density distribution)

Defects have the tendency to cluster on the wafer. The negative binomial or Gamma model introduces an extra parameter to account for this effect. The negative binomial model is a very widely used model as well. It uses an extra parameter  $\alpha$  which is equal to the inverse square of the coefficient of variation of the Gamma distribution, but also can be interpreted as a parameter that characterizes the level of defect clustering. Parameter  $\alpha$  increases with decreasing variance in the defect distribution. An important attribute of the Gamma distribution is that  $\alpha$  can be used to emulate other distributions. For instance, when there is little clustering,  $\alpha$  is high and the yield model approaches the Poisson model.

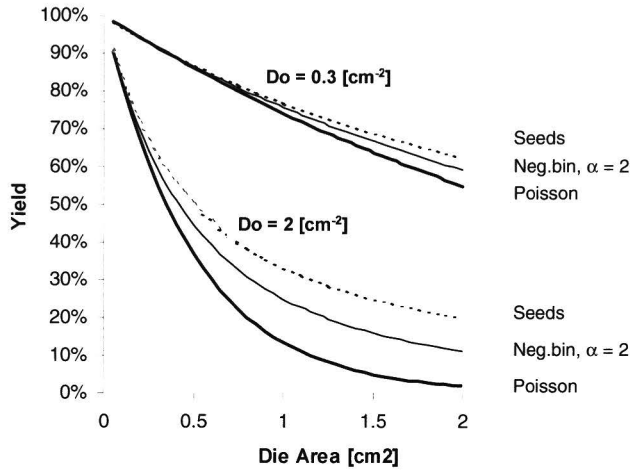
$$Y = \lim_{\alpha \rightarrow \infty} \left(1 + \frac{\lambda}{\alpha}\right)^{-\alpha} = e^{-\lambda} \quad (2.9)$$

Similarly, if the clustering is high,  $\alpha$  will be close to 1, and the yield model approaches the Seeds model:

$$Y = \lim_{\alpha \rightarrow 1} \left(1 + \frac{\lambda}{\alpha}\right)^{-\alpha} = \frac{1}{1 + \lambda} \quad (2.10)$$

$\alpha$  can be derived from the mean and variance of failing devices [17]. The more defect mechanisms exist with varying degrees of clustering, the larger is the overall clustering factor [18,19]. The negative binomial model has proven to be accurate in several cases [20]. The advantage of the model is that the extra clustering parameter gives more fitting capability than in the single parameter yield models such as the Poisson model or Seeds model. However, an extra parameter means also extra parameter extraction costs.

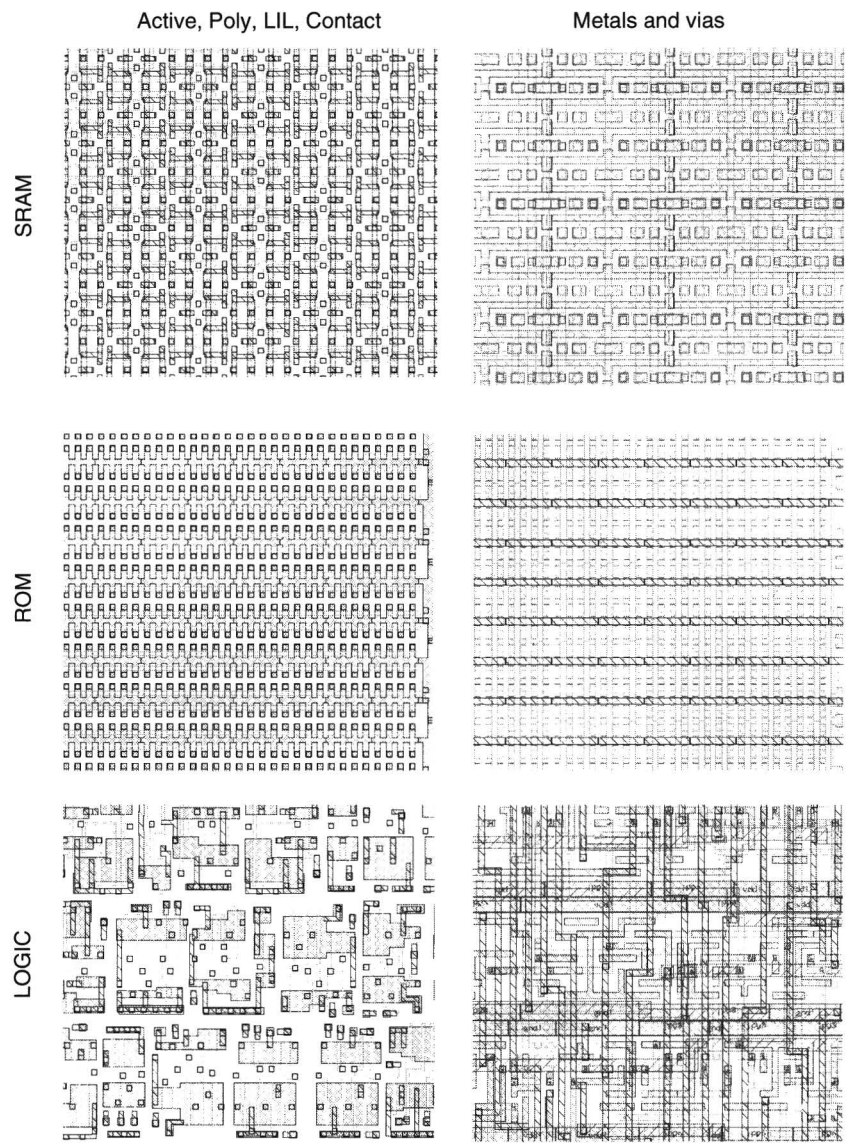
Figure 2.6 compares the yield models described above at typical and high defect densities ( $\sim 0.3 \text{ cm}^{-2}$  and  $2 \text{ cm}^{-2}$ ). At typical defect densities and device areas ( $0.5\text{-}1 \text{ cm}^2$ ) the yield models results do not differentiate very much. For high defect densities the differences in yield model results can be substantial.



**Figure 2.6** Comparison between yield models

### **2.3.2 Yield models based on design attributes other than device area**

Whether or not a defect will affect the functionality of a device depends on the local sensitivity of the product in the area where the defect occurs. The local density of designed structures can for example determine the sensitivity of this area. Under the condition that there is little variance of the sensitivity within the product nor between different products, the product yield can be adequately predicted using the models described in the previous section. This is for example the case in the manufacturing of products that consist of mainly one design style such as memories or completely digital products. In these design styles the structure density is homogeneously distributed across the die, and once the average sensitivity is known, the sensitivity (critical area) of similar products with different sizes can be assumed proportional to the die area and can thus be extrapolated. However, if within the same die, different design styles are used as in applications for mixed signal or digital logic with embedded memory, the spatial sensitivity distribution within a product and also the sensitivity differences between different products can be quite significant [1]. The probability that a defect will kill the die, strongly depends on what location of the IC it occurs. In figure 2.7 an example of different design styles that can exist within one product are shown. SRAM is very densely packed in the front-end layers, but uses only the lower metal layers. Logic area is much less dense than SRAM or ROM, but uses more high metal layers. The particular ROM shown in figure 2.7 uses many more vias than SRAM and logic. It is obvious that the extent of the yield loss for those different design styles will be very different, even though they are manufactured at the same time in the same process. The yield of a product is therefore determined by the sensitivity of each design style itself and by how the die area is divided between the different design styles. This section describes how design sensitivity can be accounted for in the various yield models.

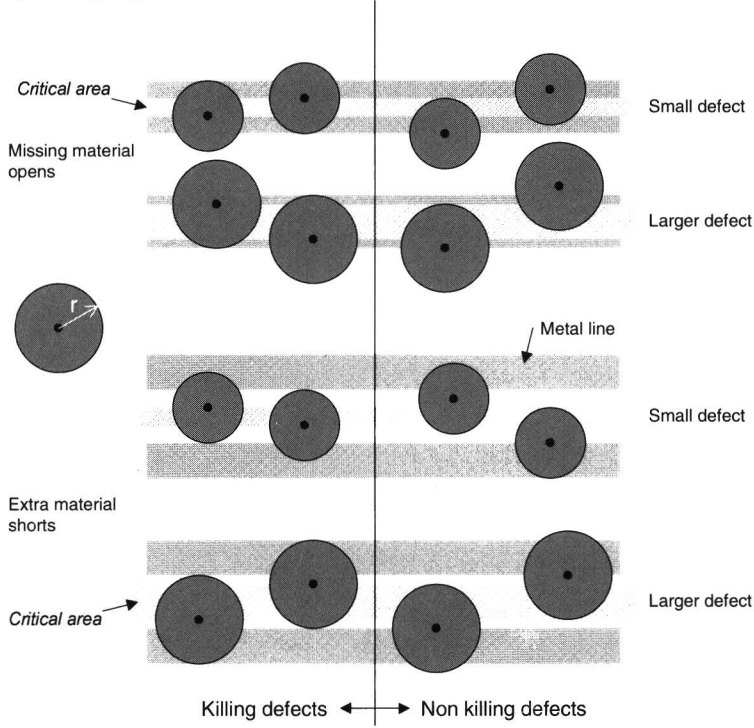


**Figure 2.7** Different design styles with different structure densities and sensitivities

### Critical area yield model

As was indicated before, not the total die area, but only part of it is sensitive to defects. The part of the circuit that is sensitive to yield loss is often expressed in terms of *critical area*, which can be defined as the area in the die in which the center of a defect must be situated to create a fault [21-23]. Because the critical area depends both on the type and size of the defect, it is always reported for a specific defect type and is measured as a function of the size of the defect. The defect model that is used to describe the failure mechanism is crucial for the critical area yield model.

Figure 2.8 shows an example where the killing defects are modeled by disk shaped material of extra or missing material. As can be seen from the figure, the critical area (shaded) depends on the radius of the defect.



**Figure 2.8** Concept of critical area (shaded) for shorts and opens

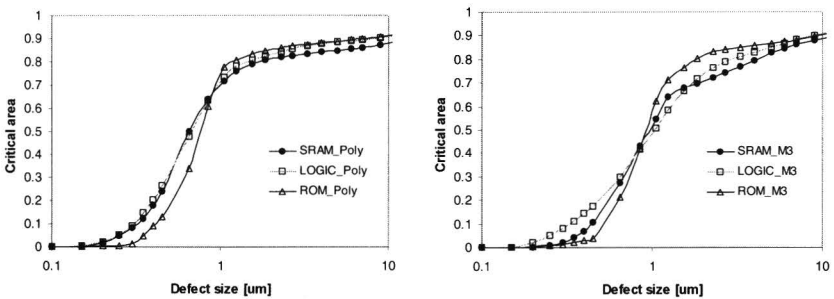
When the critical area  $A_{ci}(r)$  as a function of defect size  $r$  can be calculated for defects of type  $i$ , and the corresponding defect size distribution of the manufacturing process  $D_i(r)$  is known, the fault density can be calculated as:

$$\lambda_i = \int_0^{\infty} A_{Ci}(r) D_i(r) dr \quad (2.11)$$

and if there are  $N$  defect types for which, based on the defect density distribution, a Poisson based yield model can be used, the total yield of a product can be calculated as

$$Y = \prod_{i=1}^N e^{-\lambda_i} \quad (2.12)$$

Figure 2.9 shows typical normalized critical area curves for poly and metal 3 for SRAM, ROM and standard cell logic blocks, as shown in figure 2.7.



**Figure 2.9** Critical area curves for poly and metal 3 shorts extracted from SRAM, ROM and random logic designs

The defect size distribution is often modeled using

$$D_i(r) = \frac{K}{r^p} \quad (2.13)$$

Where  $K$  describes the defect density level and  $p$  describes the rate of increase of defect population with the defect size [24-26]. Values between 2 and 4 are reported in literature for  $p$ .

The value of  $p$  for a manufacturing process is a very important factor in yield calculations. It describes the relationship between the number of small and large defects. It can be shown that the value of  $p$  determines whether it makes sense to shrink ICs [25]. When a die is shrunk with a certain factor, the die becomes smaller and therefore will contain less defects when the defect density remains unchanged. However, at the same time, due to the shrink the die will be more sensitive to the defects. When  $p$  equals 3, these two effects exactly compensate each other, and the yield for the shrunk die and the original die will remain equal,



but more dies can be placed on the wafer and thus the wafer efficiency is increased. When  $p$  is larger than 3, the impact of the smaller defects on the shrunk die will be relatively larger and the yield of the shrunk die decreases.

Critical area yield models have been derived for opens and shorts for conducting materials such as metals, poly and diffusion areas. Yield models based on the extraction of critical area for vias and contacts exist as well [27,28].

### 2.3.3 Critical area model derivatives

In real life situations, critical area yield models have proven to accurately predict spot defect related yield loss in various memory, digital and mixed signal applications. However, the continuous characterization of the manufacturing process in terms of defect size distributions is a difficult task which involves many in-line inspections or test structure measurements on silicon [29-33]. In addition, substantial costs are involved in the computation of critical area for different products and for different types of defects for each manufacturing layer. (See also chapter 3). The costs associated implementing a critical area model are therefore often the reason why simpler models are derived from the critical area model. Such models can then be based on the notion that the critical area of an IC is determined by the density of structures in the circuit which usually correlates well with much simpler design attributes such as the transistor density, the number of nets, the total interconnect length, the number of vias, or mask transmission coefficients. Such attributes do not require extensive product and process characterization because they are much easier to obtain. An additional advantage in the application of such models is that often the necessary design parameters are already available in an early stage of the design phases of a product. Yield predictions and the related design tradeoffs can therefore be made before the complete layout is finished. Such yield models can be classified as critical area yield model derivatives.

#### *One layer critical area model*

Within a category of products there is a high probability that the critical area in one layer correlates well with critical area in other layers. A high transistor density usually means a high number of nets and therefore the critical area in the metal layers will be high as well. Therefore, in [34] for example, a one-layer critical area model is proposed in which metal 1 critical area is taken for product sensitivity characterization. The advantage being that only critical area of only one metal layer needs to be extracted.

#### *Critical area slope model*

To simplify the extraction even more, in [34] it is shown that the initial slope of the critical area curve characterizes critical area very well and can be used adequately to predict the yield loss due to spot defects. This can be explained by the fact that small defects occur much more frequently than large defects. (See equation 2.13)

### *Transistor based yield models*

In some applications metal 1 is mostly used as a local interconnect and therefore the critical area of metal 1 correlates well with the number of transistors in the circuit. In such cases the fault density can be modeled by

$$\lambda = NPoF \quad (2.14)$$

where  $N$  is the number of transistors and  $PoF$  the probability of failure that needs to be fitted to the actual measured product yields. It is obvious that the extraction of the number of transistors from the design takes only a fraction of the time that is required for calculation of the critical area of an IC.

A more advanced transistor based model is described in [35]. This model takes into account the number of minimum feature size squares that are needed to define a single “average” transistor ( $d_d$ ), minimum feature size of the manufacturing process ( $f$ ),

$$\lambda = \frac{N_{tr} d_d D}{f^{p-2}} \quad (2.15)$$

$D$  and  $p$  are defect characterization parameters.

### *Yield model based on the netlist*

In [36] a model is described that takes the netlist of a circuit and defect size distribution parameters as an input to predict the yield. For standard cell logic the model produces adequate results because the number of nodes in the netlist of a product correlates well with the average critical area that is generated after routing. This is especially the case when the different products are designed using the same design tools.

### 2.3.4 Yield model for any design attribute

Design rule manuals are often interpreted in a “digital” manner: structures that are designed according to design rule have a 0% probability of fail, while structures that violate the design rules always have a 100% probability of fail. In practice however this is not the case as is shown in figure 2.10A that shows the probability of fail of a structure as a function of a design attribute. The yield of any designed structure is a function of the applied design attributes such as spacings, widths, areas, overlaps, extensions or combination of those.

If the probability of fail for a structure as a function of the design attributes is known  $PoF(DA)$ , and the number of occurrences in an IC layout of each of those structures as a function of that same attribute is  $N(DA)$ , then the total fault density of the design attribute in the product is:

$$\lambda_{DA} = \int_{DA=-\infty}^{DA=\infty} N(DA) \cdot PoF(DA) dDA \quad (2.16)$$

as is shown in figure 2.10B,C.

If the probability of fail for a design attribute is varying across the wafer (x,y) then the fault density can be expressed as

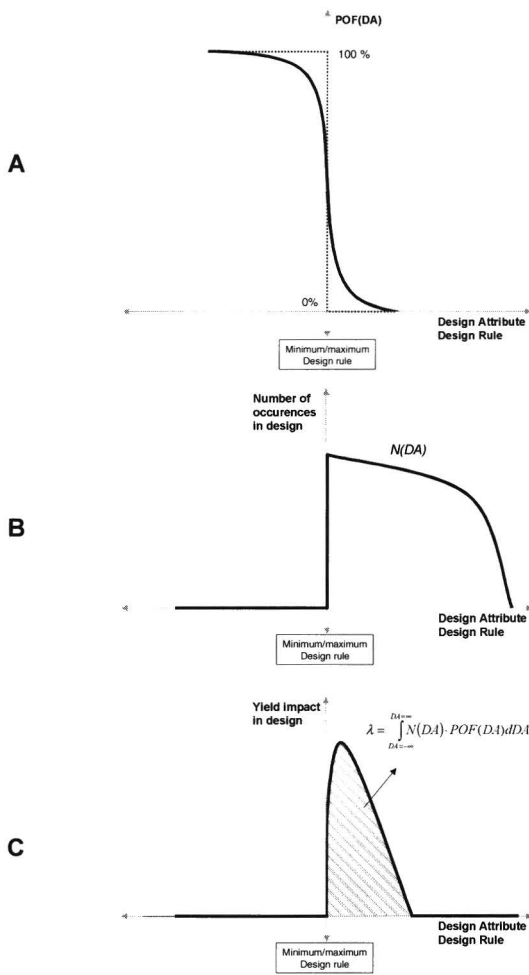
$$\lambda_{DA} = \int_{y=-\infty}^{y=+\infty} \int_{x=-\infty}^{x=+\infty} \int_{DA=-\infty}^{DA=\infty} N(DA, x, y) \cdot PoF(DA, x, y) dDA dx dy \quad (2.17)$$

If the probability of fail for a designed structure is constant over a wafer and there is no dependency on a design attribute (This can for example be the case for vias that are implemented in only one size) then

$$\lambda_{DA} = N_{DA} PoF_{DA} \quad (2.18)$$

Where  $N_{DA}$  is the number of occurrences of the structure in the product (e.g. number of vias) and  $PoF_{DA}$  is the failure rate of that structure, as is measured from for example a test structure.

Critical area yield models for vias exist, however, because of the computational effort calculating critical area of vias as a function of defect size, often model (2.18) is used for calculating yield loss due to vias. It is obvious that the extraction of the number of vias from the design takes only a fraction of the time required for critical area extraction.



**Figure 2.10** Probability of fail, occurrence in a product, and fault density of a structure as a function of a design attribute such as spacing, area, width, or overlap

### 2.3.5 Experimental comparison of yield models

In the experiment described below a comparison was made between different yield models [36]. The purpose of this experiment was to verify whether it is possible to develop a yield model that is as simple as the Poisson model with respect to design parameter extraction and at the same time is as adequate as a critical area model to explain the yield differences between products with different design styles in the same manufacturing process.

#### Models

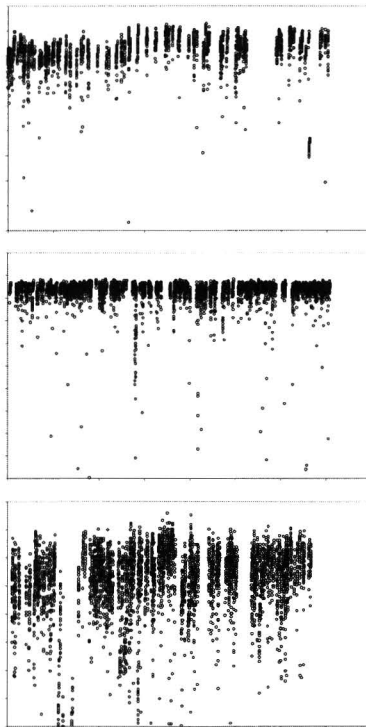
The set of models as listed in table 2.3 were chosen, assuming that the yield equals  $e^{-\lambda}$ . Models 1-4 are described in the previous sections. Models 5-8 are new transistor-based models.

Model #	$\lambda$	Parameters
1	$A_{ch} \times D_0$	$A_{ch}$ : chip area; $D_0$ : defect density;
2	$\int_0^\infty A_{cr}(r) \times D(r) dr$	$A_{cr}(r)$ : critical area for defects of radius $r$ ; $D(r)$ : defect density for defects of radius $r$ ;
3	$Ntr \times Dd \times Fs^p \times D$	$Ntr$ : number of transistors; $Dd$ : design density; $Fs$ : feature size; $D$ : defect density; $p$ : power factor for feature size;
4	$Ntr \times Dd \times Fs^p \times D^{\frac{d^*}{Dd}}$	$Ntr, Dd, Fs, p, D$ : same as above; $d^*$ : power factor for defect density;
5	$P_{of1} \times Ntr$	$Ntr$ : same as in Simple Model 1; $P_{of1}$ : possibility of fail;
6	$P_{of2} \times Ntr \times Dd^r$	$Ntr, Dd$ : same as in Simple Model 1; $r$ : power factor for design density; $P_{of2}$ : possibility of fail;
7	$P_{of3} \times Ntr \times Dd^r \times Fs^p$	$Ntr, Dd, Fs, p$ : same as in Simple Model 1; $r$ : same as in design density model; $P_{of3}$ : possibility of fail;
8	$P_{of4} \times Ntr \times Dd^r \times Fs^p \times Lm^k$	$Ntr, Dd, Fs, p, r$ : same as above; $Lm$ : number of metal layers; $k$ : power factor for number of metal layers; $P_{of4}$ : possibility of fail.

**Table 2.3** Models under investigation

*Manufacturing data*

Manufacturing yield data was taken from products running in one CMOS fab line with different ( $0.5\mu\text{m}$ ,  $0.4\mu\text{m}$ ,  $0.35\mu\text{m}$  and  $0.25\mu\text{m}$ ) feature sizes. In total a subset of 23 products were selected for analysis. Only mature products with no known parametric yield loss components were chosen for the analysis. Immature products and products with substantial analog components were also excluded from the experiment. Special attention was paid to the choice of the period of time in which data was collected. To make sure that the estimated yield model parameters were stable (i.e. were unaffected by the ongoing yield learning process), the periods of observation were limited to relatively short intervals of approximately 100 days, covering a total period of 300 days. Within each time interval it can be assumed that the large variety of products is exposed to the same process conditions. Information about the sample size is indicated in table 2.4. For proprietary reasons only three categories of sizes are mentioned: SMALL – if sample size is more than 100, MEDIUM – if more than 500 and LARGE – if the sample size is more than 1000 wafers. In total the number of wafers used for this experiment was more than 40000. Figure 2.11 shows examples of the wafer level yield trends that were used for the experiment.



**Figure 2.11** Scatter plots of the yield for three different products

Product	Technology		Area *	Num. Of Tr.*	Time Zone 1		Time Zone 2		Time Zone 3	
	F.Size( $\mu$ m)	#M.Layers			Yield	S. Size	Yield	S. Size	Yield	S. Size
1	0.50	3	1.00	0.0017			0.979	Small	0.985	Small
2	0.50	3	1.98	0.0062			0.979	Small	0.978	Small
3	0.40	3	4.33	0.1357	0.890	Medium	0.895	Large	0.900	Small
4	0.40	3	2.16	0.0425	0.958	Medium	0.953	Medium		
5	0.40	3	1.70	0.0590	0.967	Small	0.967	Small		
6	0.40	3	3.03	0.0611	0.943	Large	0.943	Large	0.945	Large
7	0.40	3	3.53	0.0967	0.912	Large	0.926	Large	0.917	Large
8	0.40	3	2.99	0.0613	0.946	Small	0.942	Small	0.941	Small
9	0.40	3	3.22	0.0803	0.931	Large	0.931	Large	0.930	Large
10	0.40	3	1.42	0.0317	0.965	Small	0.964	Small	0.964	Small
11	0.35	5	6.63	0.5852	0.643	Small			0.741	Small
12	0.35	5	3.24	0.2161	0.829	Large	0.822	Large		
13	0.35	5	2.14	0.1597	0.886	Medium	0.877			Small
14	0.35	5	2.38	0.1990	0.875	Medium	0.900	Medium	0.897	Small
15	0.35	5	5.27	0.3552	0.752	Large	0.775	Medium	0.791	Small
16	0.35	5	4.17	0.2086	0.796	Medium	0.829	Large	0.847	Large
17	0.35	5	5.49	0.2782	0.762	Medium	0.785	Medium		
18	0.35	5	9.85	0.4603	0.653	Small	0.652	Small	0.722	Medium
19	0.35	5	12.71	0.5640	0.597	Medium	0.640	Large	0.630	Large
20	0.35	5	2.84	0.1637			0.873	Medium	0.910	Small
21	0.35	5	20.54	1.2391			0.382	Small	0.416	Small
22	0.35	5	10.07	0.5784	0.562	Small			0.632	Small
23	0.25	5	1.40	0.2173			0.804	Small	0.801	Small

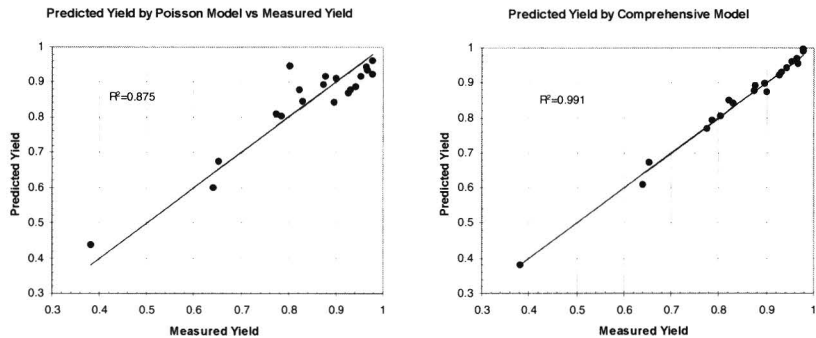
**Table 2.4** Product manufacturing data. \*Note: All data has been normalized for proprietary reasons

### Results

In order to filter out low yielding wafers due to accidents, for each time period the median value of the wafer level product yields was taken to fit the yield models listed in table 2.3. The model parameters were tuned to minimize the average error. Table 2.5 shows the obtained average error, maximum error and correlation coefficient. Figure 2.12 shows the predicted yield versus the measured yield for all the models listed in table 2.3. The obtained results with model-8 are the most accurate.

Model #	Average error for time zone			Maximum error for time zone			R <sup>2</sup> for time zone		
	1	2	3	1	2	3	1	2	3
1	5.20%	5.10%	5.00%	9.80%	14.10%	14.70%	0.863	0.875	0.89
2	2.10%	3.60%	4.10%	4.50%	12.00%	15.50%	0.975	0.963	0.947
3	2.90%	2.20%	1.90%	7.20%	6.40%	5.00%	0.952	0.976	0.983
4	2.40%	2.80%	3.30%	5.10%	9.30%	11.20%	0.977	0.96	0.953
5	2.10%	1.90%	2.90%	4.70%	5.00%	8.20%	0.973	0.983	0.964
6	1.60%	1.80%	2.20%	3.80%	4.50%	7.30%	0.988	0.985	0.978
7	1.40%	1.40%	1.40%	3.50%	3.30%	3.30%	0.989	0.991	0.992
8	1.40%	1.50%	1.30%	3.50%	3.20%	3.20%	0.989	0.991	0.991

**Table 2.5** Model tuning results per time zone. Each time zone represents 100 days of manufacturing



**Figure 2.12** Predicted versus modeled yield for model 1 (Poisson) and model 8

*Conclusions*

Based on the results above it can be concluded that accurate spot defect related yield prediction is feasible with simple models. The more design parameters that correlate with critical area are included in the model the better the model predicts the differences in yield loss between products. The advantage of the models used for the above comparison is that the necessary design attributes can easily be determined before the design process is completed. Therefore yield prediction can be performed at very low cost and still be accurate enough to enable appropriate design-manufacturing tradeoffs.



## 2.4 Conclusions and discussion

With the small windows of opportunity for modern VLSI products on the market, yield and the predictability thereof are important parameters determining economic success or failure. Thus, predictive yield modeling capabilities are crucial in both the design and manufacturing environments of VLSI products.

In this chapter existing yield models are described. In specific situations the models are reported to adequately describe yield loss for many different failure mechanisms. As an example, in a practical situation of a stable manufacturing process of which the yield loss was mainly driven by random defects, a comparison of existing and new yield models was done. Results show that in such a situation accurate spot defect related yield prediction is feasible. Whereas models that only take into account the area of the IC do not predict the yield very accurately, models that take into account design density parameters do. The more design parameters that correlate with critical area are included in the model, the better the differences in yield loss between products is predicted.

The challenge of yield prediction for a user now does not lie in the development of even more new yield models, but in the accurate extraction of parameters to calibrate these models. What models and parameters to use depends on the specific goal and on the ability to accommodate the associated costs of yield model parameter extraction. For some applications parameter extraction may be easy and inexpensive, for others difficult and costly. Therefore the user of a yield model should always clearly determine what are the goals and corresponding costs of the yield prediction. Is absolute yield prediction accuracy for example really important? Or does he only need to explain relative differences between products? The yield prediction and yield model parameter extraction methodology described in the next chapter plays an important role in dealing with this trade-off.

## References Chapter 2

1. Stapper C.H. and R.J. Rosner, Integrated Circuit Yield Management and Yield Analysis: Development and Implementation, IEEE Trans. SEM, 8, 1995, pp. 95-102
2. Cheek G. and G. O'Donoghue, Yield Models in a Design for Manufacturability Environment: A bibliography, IEEE/SEMI int. Semiconductor Manufacturing Science Symposium, -, 1993, pp. 133-135
3. Maly W., Yield Models: A Comparative Study, Int. Workshop on Defect and Fault Tolerance in VLSI Systems, 2, 1990, pp. 15-31
4. Stapper C.H., Fact and fiction in yield modeling, Microelectron. J., 20, 1989, pp. 129-151
5. Ferris-Prabhu A.V., On the Assumptions Contained in Semiconductor Yield Models, IEEE Trans. CAD, 11, 1992, pp. 966-975
6. Li M., Milor L., Computing Parametric Yield Using Adaptive Statistical Piecewise Linear Models, IEDM, 1996, pp. 473-476
7. Milor L., A. Sangiovanni-Vincentelli, Computing Parametric Yield Accurately and Efficiently, IEDM, 1990, pp. 116-119
8. Hofstein, S.R. and Heiman, F.P., The silicon insulated gate field effect transistor, proc. IEEE, vol. 51, no. 9, 1963, pp. 1190-1202
9. Paz, O., Lawson, T.R., Modification of Poisson statistics: modeling defects induced by diffusion, IEEE J. Solid State Circuits, SC-12 no. 5, 1977, pp. 540-546
10. Stapper, C.H., LSI yield modeling and process monitoring, IBM J. Res. Dev., 20, 1976, pp. 228-234
11. Stapper C.H., The Defect-Sensitivity Effect of Memory Chips, IEEE J. SC, 21, 1986, pp. 193-198
12. Murphy B.T., Cost-Size Optima of Monolithic Integrated Circuits, Proc. IEEE, 52, 1964, pp. 1537-1545
13. Seeds R.B., Yield and Cost Analysis of Bipolar LSI, IEEE int. electron Devices Meeting, oct, 1967.
14. Seeds R.B., Yield Economic and Logistic Models for Complex Digital Arrays, "IEEE int. Conv. Rec, Pt. 6", 1967, pp. 60-61
15. Okabe T. et al., Analysis of Yield of Integrated Circuits and a New Expression for the Yield, Electrical Engineering in Japan, 92, 1972, pp. 135-141
16. Stapper C.H., On Murphy's Yield Integral, IEEE Trans. SEM, 4, 1991, pp. 294-297
17. Stapper C.H., The effects of wafer to wafer defect density variations on integrated circuit defect and fault distributions, IBM J. Res. Develop., 29, 1985, pp. 87-97
18. Collica R.S., The Effect of the Number of Defect Mechanisms on Fault Clustering and its Detection Using Yield Model Parameters, IEEE Trans. SEM, 5, 1992, pp. 189-195
19. Ferris-Prabhu A.V., A Cluster-Modified Poisson Model for Estimating Defect Density and Yield, IEEE Trans. SEM, 3, 1990, pp. 54-59
20. Cunningham J.A., The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing, IEEE Trans. SEM, 3, 1990, pp. 60-71.

21. Maly W, Modeling of Point Defect Related Yield losses for CAD of VLSI Circuits, Proc. Of ICCAD-84, Nov, 1984, pp.161-163.
22. Maly W. and J.Deszczka, Yield Estimation Model for VLSI Artwork evaluation, Electronics Letters, 19, 1983, pp.226-227.
23. Maly W., Modeling of Lithography Related Yield Losses for CAD of VLSI Circuits, IEEE Trans. CAD, 4, 1985, pp.166-177.
24. Maly W. et al., "Characterization of Type, Size and Density of Spot Defects in the Metalization Layer", Yield Modeling and Defect Tolerance in VLSI, Adam Hilger Philadelphia", 1988, pp.71-90.
25. Stapper C.H. Modeling of defects in integrated circuit photolithographic patterns, IBM J. Res. Develop., 28, 1984, pp.461-475.
26. Unsoeld A. Stromdichte von Teilchen verschiedener Massen, Der neue Kosmos, 1967, pp.70.
27. Ouyang C. and W. Maly, Efficient Extraction of Critical Area in Large VLSI IC's, Proc. IEEE Int. Symposium on Semiconductor Manufacturing, 1996, pp.301-304
28. Ouyang C., W. Pleskacz and W. Maly, "Extraction of Critical Area for Opens in Large VLSI Circuits", Proc. IEEE In. Workshop on Defect and Fault Tolerance of VLSI Systems, -, 1996, pp.21-29.
29. Hess C, L. H. Weiland, "Novel Methodology to Include all Measured Extension Values per Defect to Improve Defect Size Distributions", IEEE/SEMI Advanced Semiconductor Manufacturing Conference, -, 1998, pp.197-202.
30. Milor L. et al., "Yield Modeling Based on In-Line Scanner Defect Sizing and Circuit's Critical Area", IEEE Trans. SEM, 12, 1999, pp.26-35.
31. Dudivier F., M. Rivier, "Approximation of Critical Area of IC's with Simple Parameters Extracted from the Layout", Int. Workshop on Defect and Fault Tolerance in VLSI Systems, 1995, pp.1-9.
32. Khare J.B., W. Maly, and M.E. Thomas, "Extraction of Defect Size Distributions in an IC Layer Using Test Structure Data", IEEE Trans. SEM, 7, 1994, pp.354-368
33. Maly W. et al , "Double Bridge Test Structures for the Evaluation of Type, Size and Density of Spot Defects", Carnegie Mellon University Research Report CMUCAD-87-2, 1987.
34. Heineken H.T., J. Khare, and W. Maly, "Yield Loss Forecasting in the Early Phases of the VLSI Design Process", IEEE Custom IC Conf., 1996.
35. Maly W., H.T. Heineken, and F. Agricola", A Simple New Yield Model, Semic. Int. July, 1994, pp.148-154.
36. Yanwen Fei, Paul Simon, and Wojciech Maly "New Yield Models for DSM Manufacturing", IEDM Conference 2001

---

## **Chapter 3**

# **Yield Prediction Methodology and Model Parameter Extraction**

## 3

### 3.1 Introduction

Yield modeling is not a goal in itself neither in a design environment nor in a manufacturing environment. However, as discussed in the previous chapter, predictive yield models are essential both for forecasting product cost and for managing yield improvement activities. The usefulness of any yield model depends on the ability of the user to calibrate the necessary parameters. This chapter describes the extraction of the yield model parameters related to the design and the process.

Section 3.2 discusses yield prediction methodology in general. In section 3.3 the extraction of process related yield model parameters is presented. Section 3.4 discusses design related yield model parameter extraction and section 3.5 describes the development and implementation of an industrial manufacturability assessment environment (MAE).

### 3.2 Yield prediction methodology

The probability of a failure of an IC is a function of spatial distribution of process conditions that may cause a defect, and of the fraction of the die area that produces a fault when exposed to such conditions. Within this concept an IC can be perceived as a large collection of different design configurations, each having a different probability of failure that is determined by the combination with the local process conditions. So, if there are  $K$  different design attributes in a product design, and the yield loss due to each of those is  $L_i$ , then the total yield of the product can be expressed as

$$Y = \prod_{i=1}^K (1 - L_i) \quad (3.1)$$

In which  $L_i$  is a function of the occurrence of a design attribute  $i$  ( $DA_i$ ) on location  $(x,y)$  on the wafer, and the spatial and statistical variation of the process conditions that, in combination with the design attribute, cause the fault:

$$L_i = f\langle DA_i(x, y), \mu_i(x, y), \sigma_i(x, y) \rangle \quad (3.2)$$

in which the mean  $\mu(x, y)$  and variance  $\sigma(x, y)$  describe the statistical distribution of the process parameter as a function of the location on the wafer. For example the yield loss due to open vias in a particular area of the wafer can be a function of the etching rate distribution and the number of vias in that region. In the case that for all K design attributes  $DA_i(x, y)$ ,  $\mu_i(x, y)$  and  $\sigma_i(x, y)$  are known, the yield loss pareto for the product can be made.

Theoretically, all design attributes  $DA(x, y)$  can be extracted from a design. However, this is not very practical because of computation limitations. Therefore, in practice, a sensible subset of design extractions need to be done for which, based on experience, yield loss can be expected and for which it is possible to determine  $\mu_i(x, y)$  and  $\sigma_i(x, y)$ . Extraction of design attributes is discussed in section 3.4.

The determination of the spatial distribution of process conditions that cause a fault ( $\mu_i(x, y)$  and  $\sigma_i(x, y)$ ) is far more complicated and is discussed in section 3.3.

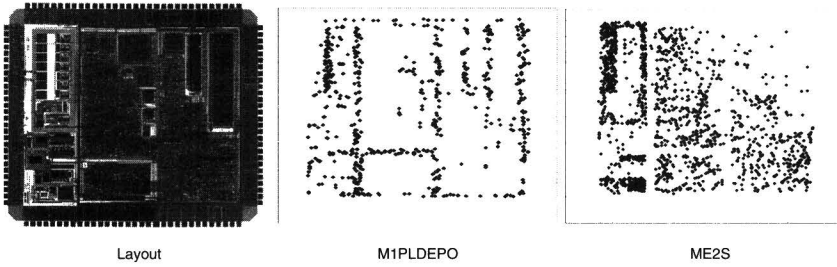
### 3.3 Process parameter extraction

Two ways of approaching the determination of process related yield model parameters can be distinguished: the *modeling approach*, and the *empirical approach*. Both are described below.

#### *Modeling approach*

In the modeling approach an effort is made to fully understand the failure mechanisms that cause the fault. The relationship between the design vulnerability and the process variation are modeled [17]. However, often the physics involved are not well understood, and in addition often the failure mechanisms involve new materials of which the behavior is not well characterized yet. This makes the modeling approach difficult if not impossible. Even if it would be possible to create a model describing the design-process interaction, in most cases it will be difficult or impractical to obtain the data needed for model calibration. Therefore, the cost of the yield model development and the calibration of it are usually very high. For instance in-line defectivity measurements could be used to determine the defect size and spatial distribution of defects in order to calibrate a critical area model. A defect sensitivity map of the product could then be extracted from the layout (see figure 5.21) to predict the yield. However, although in-line measurements can be effective for monitoring purposes, the interpretation and determination of absolute defectivity levels is sometimes difficult.

To study in more detail the above problem, the following experiment was conducted. In-line defectivity measurements were combined to form a stacked die-map as shown in figure 3.1. Clearly the layout of the product can be recognized in the defect patterns. In this case the inspection tool is more likely to trigger on defects that are located in open areas than in dense areas.



**Figure 3.1** *The sensitivity of in-line defectivity measurements is influenced by the layout*

Therefore the defect statistics found on this product may not represent reality. In addition, defect size measurements, classification of the defects, and the determination of the kill ratio per defect is a difficult task that is not well implemented in software yet. Consequently, the calibration of the critical area model by in-line measurements needs extensive human interpretation and is therefore impractical.

The advantage of the modeling approach however, is that once the failure mechanism is well understood and modeled, it is generically applicable to other design attributes and products as well.

#### *Empirical approach*

In the empirical approach, the yield model user is less interested in the process conditions or failure mechanism itself. Only the occurrence of a vulnerable design attribute on the wafer ( $DA_i(x,y)$ ) and the probability of failure of that particular design attribute ( $POF_{DA_i}(x,y)$ ) is of interest:

$$L_i = f\langle DA_i(x,y), POF_{DA_i}(x,y) \rangle \quad [3.3]$$

$POF_{DA_i}(x,y)$  can be characterized by using test structures that mimic the design attribute. Once  $POF_{DA_i}(x,y)$  of a design attribute, and its occurrence rate on the die are known, the yield loss can be calculated. However, in the absence of failure models the test structure results are not generically applicable to the whole design space. Consequently it is only possible to predict the yield loss of the structures in the product that are covered exactly by the corresponding test structures. In most cases this means that due to the restrictions in the number of test structures that can be used, only part of the yield loss can be calculated and the yield loss pareto will be incomplete.

Another disadvantage of a test structure approach [15,16] is that by definition the test structures on test reticles do not exactly replicate the layout conditions of products. The density of structures around, above and beneath the test structure (topography) is different from the product. Therefore, there is a possibility that due to test structure environment extra failure mechanisms are introduced that are

specific to the test structure and do not occur in the product, resulting in a too pessimistic yield prediction.

### 3.3.1 Test structure based yield model parameter extraction

The yield prediction methodology that is described here, uses both the empirical and modeling approaches, and can be used for yield model calibration during yield ramping and for product yield prediction during the design phases of a product. The methodology is based on using different kinds of test structures for yield evaluation called YEMs (Yield Evaluation Monitor).

For a large part the use of these test structures is influenced by the availability and maturity of the failure models. Generally, it can be stated that the better the models describe the interaction between design and manufacturing process, the better specific test structure results can be extrapolated to other design attributes and the less test structure area is needed. For example for determination of defects size distribution in metal layers for use in a critical area yield model, not all possible track pitches need to be covered with the test structures. A limited set of pitches is sufficient to extrapolate the results to other spacings using standard defect size distribution models. In this case the availability of a failure model therefore limits the test structure area that is needed to characterize the yield loss. However, during yield ramping, by definition not all failure mechanisms are known. The available list of known failure mechanisms is incomplete or inaccurate and extrapolation to other design attributes is therefore not possible. To overcome this problem the complete design space could be covered in separate test structures in order to represent all possible layout configurations that occur in the product. The  $POF_{DA_i}(x,y)$  of each test structure can then be translated into a product yield loss for that particular design attribute. However, limitations in the available silicon area will restrict test structure area, and the resulting number of fails per wafer will be too small to be able to extract statistically valid yield model parameters.

From these considerations it can be concluded that, for the development of YEMs, the tradeoff of the cost of test structures versus the cost of obtaining model (parameter)s should be considered. The part of the design space for which failure models exist, the models should be used in order to limit the test structure resources needed to characterize the yield loss. For the part of the design space where no failure models are available, test structures should represent the product design attributes as completely as possible within the available resource limits.

In order to save silicon area, a second type of test structure can be used in which (parts of) a product or IP-block is used to form the test structure (MIMIC test structures). For example the metal1-via1-metal2 design of an SRAM cell can be used to build up a via-string test structure for measuring the probability of failure of vias in SRAM. An advantage of such a test structure is that there exists an exact structural similarity between the IP block and the test structure [10]. The resulting failure mechanisms will therefore be identical since for instance the etching



conditions such as loading effects in the test structure are the same as in the SRAM. The yield of such a test structure can directly be translated to the SRAM via yield by

$$Y_{Via\_SRAM\_block} = Y_{teststructure} \left( \frac{Area_{SRAM\_block}}{Area_{MIMIC\_test\_structure}} \right) \quad [3.4]$$

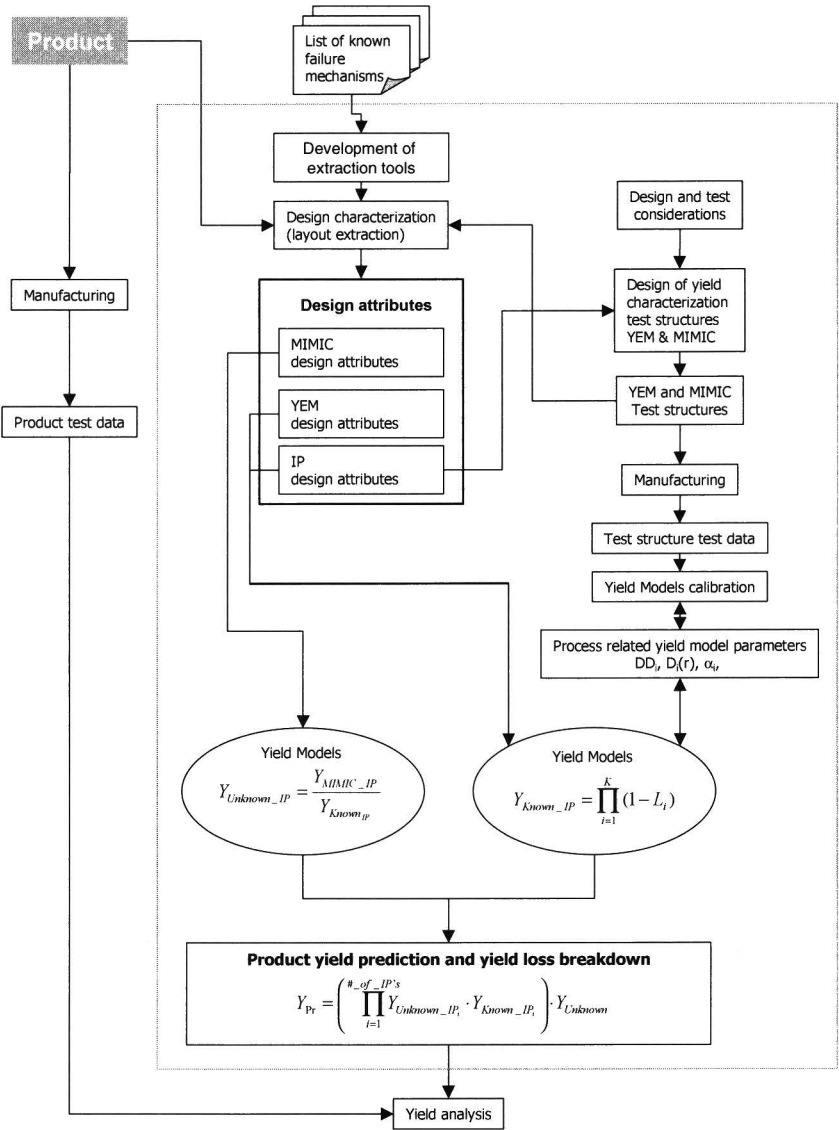
A MIMIC test structure therefore can give the user information on the yield loss on IP-block level. For example the via yield of an SRAM block, or the poly-poly shorts in a standard cell logic block can be evaluated using MIMICs. A drawback of this approach is that because such test structures are built up of several layers, there are a number of possible failure mechanisms that are covered by the test structure that cannot be distinguished.

Table 3.1 summarizes the above considerations.

Failure models	Test structure approach	Example	Drawbacks
<b>Known (anticipated) failure model</b>	Test structures cover a limited set of design attributes. Pof for other design attributes can be extrapolated using the failure model.	A limited amount of spacings between tracks are used in comb meander test structures for determining defect size distribution.	<ul style="list-style-type: none"> <li>• Inaccuracy of the models</li> <li>• Risk of leaving out new kinds of failure mechanisms</li> </ul>
<b>Unknown failure model</b>	All possible design attributes in the product are covered in the test structures.	YEMs All possible design attributes in the product are covered	<ul style="list-style-type: none"> <li>• Silicon area</li> <li>• Test costs</li> <li>• Extrapolation to other design attributes</li> </ul>
	Test structures mimic (a part of) the product	Several SRAM or logic layers are copied into a test structure to determine the POF for the different design styles. (MIMIC test structures)	<ul style="list-style-type: none"> <li>• Test structures cover several failure mechanisms at the same time that cannot be distinguished</li> <li>• Extrapolation to other products</li> </ul>

**Table 3.1** Test structure approach for yield model parameter extraction

As a result of the considerations summarized above a yield prediction methodology has been developed as schematically shown in figure 3.2. The methodology that uses a combination of the test structure approaches discussed above. The key items of the flow are explained on the next pages. Additional test structure design considerations and examples are discussed in the next section.



**Figure 3.2** Test structure based yield prediction methodology

### List of known failure mechanisms

For both the development of test structures and layout characterization tools, a list of failure mechanisms that can be expected to occur, is the starting point. Such a list can be based on the experience of process development and yield ramping of

previous technologies. Table 3.2 shows an example of such a (partial) list of failure mechanisms.

Failure mechanism	Defect	Event / Probable cause	Example of design parameter in yield model
ACT – ACT short	Parasitic Transistors	Trench Depth Implantation Profiles Trench Oxide Quality	Active area Distribution of active area sizes
	Micro scratch	STI CMP	Length of STI edge, Distribution of active-active spacing
	STI Nitride Blocked Etch	Flake / Particle Generation	Distribution of active-active spacing Critical area for active shorts as a function of defect size
	Extra LIL	Litho / patterning Particles	Distribution of active-active spacing
	Junction Leakage	Metallic Contamination	Active area
		Alcaline or Ionic contamination	Active area
		Crystallographic Defects	Active Area, Active width distribution
		Stress / Slip	
		Stress induced leakage current	STI edge
		Too much $\text{TiSi}_2$ was formed: Too much Titan was deposited Too high RTP temperature Too much stress in Active By silicide formation	Silicided Active Area, Device Width
		Poor step coverage of the borderless nitride	
		Too much STI Oxide loss Volcanoes	

Failure mechanism	Defect	Event / Probable cause	Example of design parameter in yield model
ACT – ACT open	Missing Silicide	Particles Bad silicide quality	Active width distribution Active critical area for opens as a function of defect size
	Missing Implant		Active width distribution
	Missing Active	CD variation Necking / notching	Active width distribution
	High Silicide Resistivity	Too long selective etch Too little $\text{TiSi}_2$ was formed: Too little Titan was deposited Too low RTP temperature	Silicided Active Area
POL – ACT short	Top Corner Rounding	Step height < 0 GON uniformity	Poly-active spacing distribution
	Gate oxide defect Threshold voltage shift	Antenna-effect / Plasma Charging	Antenna ratio
	Local GON thinning	Contamination / Defects	Active area
		Poly etch	Poly edge
	Misalignment	SACOX overetch	STI edge
		wafer stepper problems	Overlaps
POL – POL short	Bridging	Too small spacer Too short selective etch Too much $\text{TiSi}_3\text{N}_2\text{O}_2$ was formed: Too much Titan was deposited Too high RTP temperature	Poly Width, Silicided Active Area
	Poly Stringers	Step height > 0: Reticle Lay-out CMP ON-etch SACOX strip	Density of Active, STI edge
	Extra poly	Flake / Particle Generation Blocked Etch	Poly-Poly spacing distribution
	Poly CD-dense	Litho	Poly density distribution
	Poly residues	Poly underetch	Density of Active, STI edge
POL – POL open	Missing silicide	Too long selective etch Too little $\text{TiSi}_2$ was formed: Too little Titan was deposited Too high RTP temperature	Poly area, Poly width distribution,
	N+-P+ transition	Missing silicide Missing implant CD variation (necking)	Poly width distribution
	Poly voids	Too long selective etch Too little $\text{TiSi}_2$ was formed: Too little Titan was deposited Too high RTP temperature too much stress	Poly width distribution, Poly Area, Poly critical area for opens as a function of defect size
POLY variation causing device performance variation	Poly Slope	Non-uniform etch Resist Slope	Poly density, Space distribution, Device length distribution
		ISO-Dense / Geometry Effects	Poly density distribution, Device length distribution
	End of line Shortening	Focus problems	Device Width, Endcap Dimension
	Poly CD variation	Corner CD non-uniformity effects ISO-Dense / Geometry Effects	Poly density distribution, Device length distribution

Failure mechanism	Defect	Event / Probable cause	Example of design parameter in yield model
LIL – POL short	Leakage between LIL and unrelated Poly	Misalignment CD variations	Poly-LIL spacing distribution
LIL – POL open	High Contact Resistivity	Bad planarization due too within die topography	Poly-LIL overlap distribution
		Etched thru Silicide (bad selectivity) ARDE Volcanoes	Poly-LIL overlap distribution
LIL – LIL short	Extra LIL	Patterning	Critical area for LIL shorts as function of defect size
	Extra LIL / Stringer	Dishing Bad planarization due too within die topography	Distribution of STI edges and LIL-LIL spacings
	Extra LIL	Particles CD variations Tungsten residues after CMP Volcanoes	Critical area for LIL shorts as function of defect size
LIL – LIL open	Discontinuity	Particles	Critical area for LIL opens as function of defect size
CNT – LIL open	Resistive contact/ Discontinuity	Polymer formation due too to long etching ARDE Misalignment	CNT-LIL overlaps, Number of contacts
CNT - ME1 open	Resistive contact/ Discontinuity	Planarization Depth of focus particles	Number of contacts, Contact-metal1 overlap distribution
CNT – VIA1 open	Resistive contact/ Discontinuity	Metal 1 CD	Number of Metal1 landing pads
ME n – ME n short	Extra metal	Particles CD variations Tungsten residues after CMP	ME_n critical area for shorts as a function of defect size
ME:n Open	Resistive contact/ Discontinuity	Particles	ME_n critical area for opens as a function of defect size
ME:n – VIA:n Open	Resistive contact/ Discontinuity	Particles, Flakes, Blocked etch	ME_n critical area for opens as a function of defect size
VIA:n - ME:n+1 Open	Resistive contact/ Discontinuity	Particles, Flakes, Blocked etch Via density problems	Number of vias, Number of vias as a function of the surrounding via density
VIA:n – VIA:n+1 open	Resistive contact/ Discontinuity	Metal CD	Number of ME_n landing pads

**Table 3.2** *Example of a (partial) list of anticipated failure mechanisms*

### Product

The critical area of the product design for each individual failure mechanism is characterized so that the critical area for each test structure can be specified. In order to determine the sensitivities of each of IP block in the product, functional blocks are separated from the product first.

During testing of the product usually the yield loss per block is quantified and can be compared to the predicted yield loss.

#### *Design characterization*

For each of the known failure mechanisms a layout extraction tool is developed in order to be able to quantify design parameters of each IP block in the product. The critical area of each test structures is determined using the same layout extraction algorithms.

#### *Test structures*

In order to establish of good correlation between test structure yield and IP-block yield, the test structures represent the IP blocks in terms of both the design attributes and sensitivity [10,14]. The test structure designs therefore are based on the results from the IP design characterization. It makes no sense to explore the manufacturability of design attributes that are not found in the product.

Both YEMs and MIMIC test structures are used. The YEMs are used to quantify the process related yield model parameters for known failure mechanisms. Since (especially during yield ramping) not all failure mechanisms are known, a set of MIMIC test structures are used to capture remaining failure mechanisms that are not covered by the YEMs.

#### *Design attributes*

To be able to compare the sensitivity of the IP cores with the test structures, design characterization of both designs is done with the same extraction tools. Although the goal of a test structure is to show only one failure mechanism, sometimes it is not possible to prevent sensitivity to other failure mechanisms as well. To characterize the sensitivity of the test structure to all known failure mechanisms, all design attributes from each test structure are extracted.

#### *Yield models and calibration*

When both the design related yield model parameters and the process related yield model parameters of the known failure mechanisms are quantified, the corresponding yield models can be calibrated. The yield loss due to these failure mechanisms for the IP blocks can now be calculated using

$$Y_{Known\_IP} = \prod_{i=1}^K (1 - L_i) \quad (3.5)$$

The yield loss in the MIMIC test structures is determined by both the known and unknown failure mechanisms. The unknown part can be determined by

$$Y_{unknown\_MIMIC} = \frac{Y_{MIMIC\_IP}}{Y_{Known\_MIMIC}} \quad (3.6)$$

The IP yield can now be predicted as a product of the known and unknown part.

Due to the test limitations for both the MIMIC and YEM structures, still not all failure mechanisms may be captured. An addition unknown yield loss factor therefore will remain in the product yield prediction:

$$Y_{product} = \left( \prod_{i=1}^{\#\_of\_IPs} Y_{Unknown\_MIMIC\_i} \cdot Y_{Known\_IP\_i} \right) \cdot Y_{not\_captured} \quad (3.7)$$

When the yield of a product is measured,  $Y_{not\_captured}$  can be determined. The yield of a second product (P2) can then be predicted with the same process related yield model parameters and the corresponding design attribute extractions of the product using

$$Y_{P2} = \prod_{i=1}^K (1 - L_i) \cdot Y_{Not\_captured\_P1} \left( \frac{A_1}{A_2} \right) \quad (3.8)$$

where  $A_1$  and  $A_2$  are the product die areas.

### 3.3.2 Other considerations for test structure implementations

The process of choosing different kinds of test structures is complex and depends on many parameters such as the goals the user is trying to achieve, process complexity, and boundary conditions such as available silicon area and test resources. The main test structure design considerations will be discussed in this section.

#### *Full loop versus partial loop test structures*

There are two ways of implementing YEMs and MIMICs. The first one is the integration of the test structures on a so-called “process startup reticle” that is used for process development and yield ramping. Usually one or more products are placed on such a reticle in combination with the test structures. The advantage of such a reticle set is that the failure mechanisms that affect the yield of the product are occurring on the same wafers as the test structures so that the correlation between test structure results and product yield can be made. However, since the reticle size is limited, often a tradeoff has to be made between the number of failure mechanisms that need to be captured with the tests structures and the level of defectivity that can still be detected with reasonable resolution. Another disadvantage of the full loop approach is that it needs to go through the whole manufacturing flow before any data can be used. Depending on the process complexity the feedback time can be significant and yield learning cycles may become too slow.

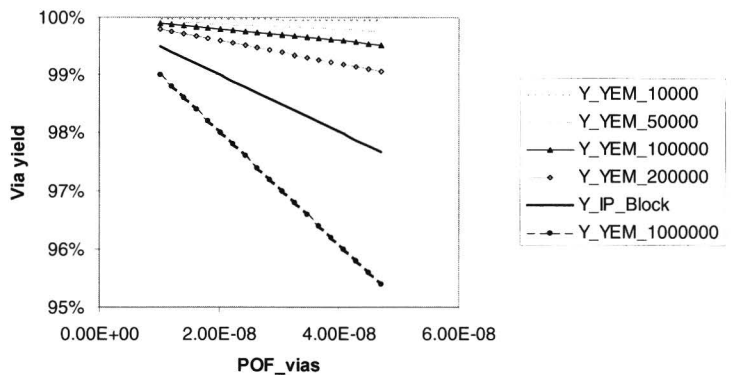
In order to overcome this problem shortloop test structures can be used that cover only part of the process, but enable a much faster yield feedback loop, accelerating yield learning significantly [12]. Shortloop test structures can be used for both (part of) the frontend and backend of the process. In shortloops a subset of the process layers is used. The available area on the reticle has to be divided between a smaller number of test structures, enabling larger test structures and detection of lower defectivity levels than on full loop reticles sets.

#### *Test structure area*

One of the most important considerations to be made during the design is the critical area of the test structure [11]. When the critical area of a test structure is too small, there is a low probability of failure and its yield in most cases will be close to 100%, even though the defectivity level for the failure mechanism may still have a major impact on product yield. If the test structure is too large the opposite may happen. Therefore the size of the test structure depends on the resolution of the yield impact the user wants to identify on the product. In general it can be said that the test structures area should be such that it is able to show a yield loss that is comparable to the yield loss on the product. Therefore the critical area of a test structure should be in the same order of magnitude as the critical area of the product. This is illustrated by the yield predictions for via test structures as a function of the probability of failure for a via as is shown in figure 3.3. The figure shows the calculated yield for via strings with different numbers of vias and for an IP block containing 500000 vias. The IP-block may show several



percent of via yield loss while this may be almost invisible on the test structures with smaller amounts of vias.



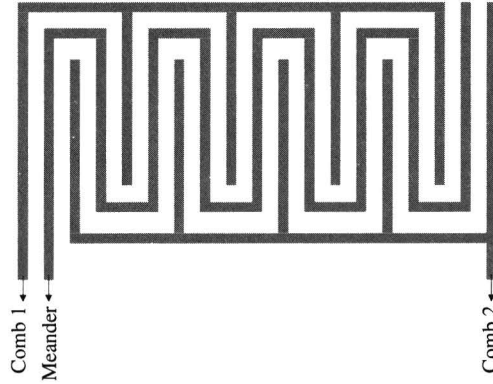
**Figure 3.3** Predicted yield of via test structures as a function of the probability of failure

### 3.3.3 Examples of test structures for yield model parameter extraction

In this section examples are shown of the development of YEMs and MIMIC test structures.

*Example1: Development of YEM for shorts in conducting layers*

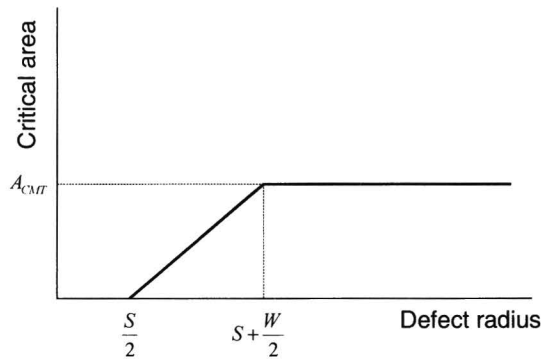
Extra material creating intra layer shorts is one of the most important failure mechanisms. Therefore structures that enable to quantify the yield loss due to this failure mechanism are crucial for yield prediction [6,7,8]. Figure 3.4 shows a schematic example of a simple comb meander test structure that can be used for this purpose. Shorts can be detected by measuring the leakage current between the meander and the combs. Opens are detected by measuring the resistance of the meander. Different spacings between the comb and meander are used to characterize defects size distribution in the process.



**Figure 3.4** Schematic representation of a comb meander test structure

The area to be covered with a comb meander test structure depends on the yield impact the user needs to identify on the product. The critical area of the test structure for a certain defect size is chosen in such a way that the predicted yield impact for the structure is identical to the yield impact in the product or IP core. Typically the comb meanders for smaller spacings can therefore be chosen smaller than the ones for larger spacings since they are more sensitive to yield loss since the smaller defects are more dominant. It is the experience of this author that three spacings are enough to fit the measurement data to the defect distribution curve.

The critical area for shorts in comb meander structures can either be extracted by critical area software or derived as shown in figure 3.5.

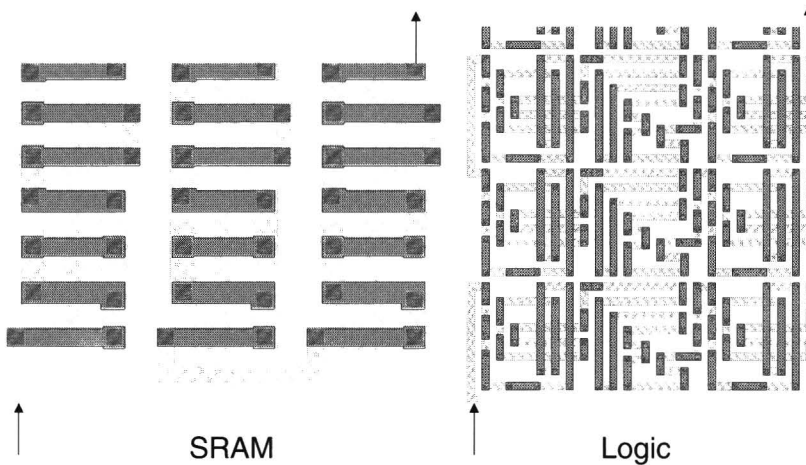


**Figure 3.5** Critical area curve for shorts in a comb meander test structure.  $A_{CMT}$  is the total area of the test structure,  $S$  is the spacing between tracks and  $W$  the width

The yield of the test structure can then be calculated with the critical area model (2.11) and the results can be used to fit the defect size parameters,  $p$  and  $K$ , to the measured yield data (2.13). See also fig 2.17. The fitted values for  $p$  and  $K$  can then be used to predict the yield loss for the IP blocks for which critical area has been extracted. Comb meander structures can be used for extracting yield model parameters for conducting layers such as metals, poly and LIL (local interconnect layer).

*Example 2: Development of YEMs for vias and contacts.*

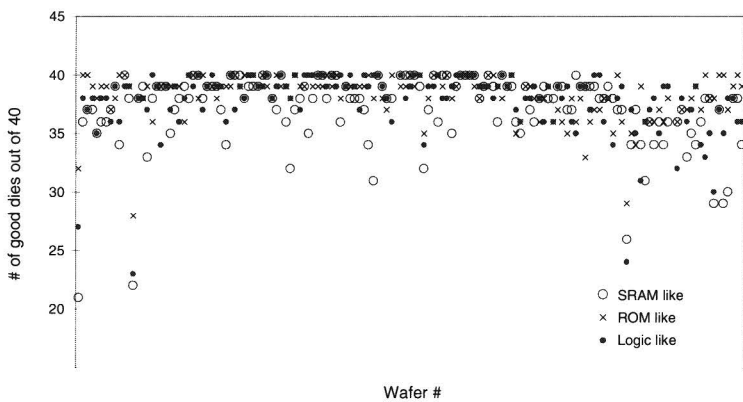
There exist many mechanisms that can cause vias to be defective. For instance, random particles may block the etch or there may be other layout dependent effects such as the etching rates of focus depths that depend on the metal or via density. In order to establish a maximum correlation of the test structure yield with the IP block yield, test structures were designed in such a way that the surroundings of each via resembled the surroundings in the IP block (MIMIC test structures). Figure 3.6 shows examples of via MIMIC test structures for LOGIC and SRAM vias.



**Figure 3.6** *Parts of MIMIC via test structures for SRAM and standard cell logic*

The structures were setup as normal via strings, but the distribution of spacings to neighboring vias and the metal density is similar to the ones in the IP blocks. In order to establish a yield impact on the test structure that is comparable to the via yield impact on the IP blocks, the number of vias that were used, again were in the same order of magnitude as the number of vias in the IP blocks.

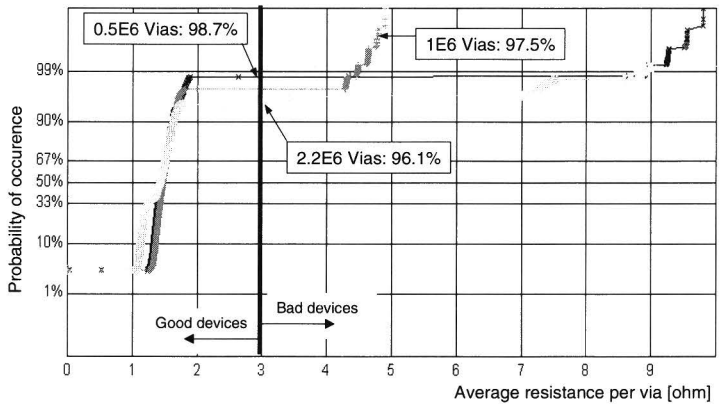
Figure 3.7 shows an example of a yield trend for the two MIMIC structures for SRAM, ROM, and logic via1 structures. In this case there is a clear difference for via yield for the different design styles although the number of vias is the same for all structures.



**Figure 3.7** Wafer level MIMIC yield trends for SRAM, ROM and Logic VIA1 structures

From the above results it can be concluded that for good correlation with IP block yield, the test structures design should resemble the IP design as closely as possible.

In order to determine the via test structure yield as a function of the number of vias, tap-offs at different number of vias are provided on each test structure [13]. The resistance for each tapp-off is measured and subsequently the yield of the is determined determining how many structures fall outside the normal resistance distribution as is shown in figure 3.8



**Figure 3.8** Cumulative distribution of resistance measurements on via strings with different numbers of vias.

In order to translate the above via MIMIC test structure data to the IP block via yield data,

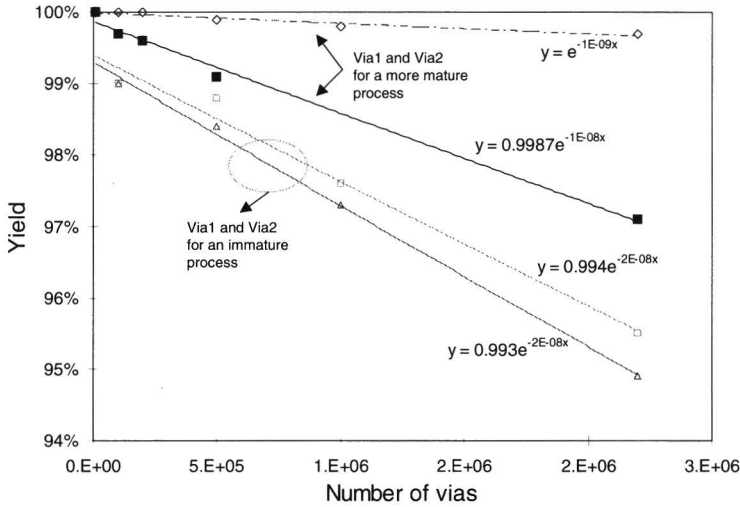
$$Y_{Via\_IP\_block} = Y_{YEM}^{\left( \frac{\#ofvias_{IP\_block}}{\#ofvias_{YEM}} \right)} \quad (3.9)$$

can be used.

Via yield in a product can also be calculated based on test structure data as is shown in figure 3.9 where the probability of failure per via is determined. The via test structure results are shown as a function the number of vias for via1 and via2, both for a process in development and for a more mature process. The yield function

$$Y_{Via\_teststructure} = Y_0 e^{-N \cdot PoF_{Via}} \quad (3.10)$$

where N is the number of vias, is fitted through the measured data to determine  $PoF_{Via}$ .



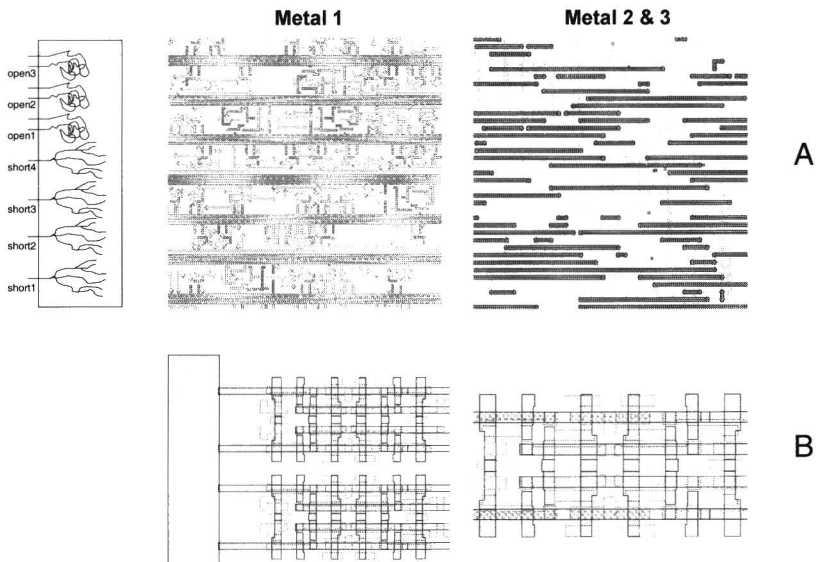
**Figure 3.9** Via string YEM structure results.

Subsequently the via fault density  $\lambda$ , and the via yield in the product can be calculated.

*example 3: Development of a MIMIC test structures*

Figure 3.10-A show an example of a standard-cell-, MIMIC type of test structure that has been developed. The purpose of this test structure is to be able to quantify backend yield for standard cell routing in products. In order to achieve this, a MIMIC layout is generated that resembles standard cell routing as much as possible and at the same time is easily testable for shorts and opens on a standard parametric tester. Standard cells are taken and placement and routing is done with a standard router that is also used for products. A special netlist used as is schematically shown in figure 3.10. The netlist contains three loops to test opens, and four ‘clock tree’ like nets that are intertwined by the router. The complexity of the routing, the number of vias that are used, and the routing density can be varied by changing placement and routing parameters.

As for all MIMIC like test structures, this one also only enables to draw general conclusions on the backend yield loss due to opens and shorts in standard cell IP. The disadvantage of the structure is that it is not possible to de-convolute the yield loss, and attribute it to specific layers in the manufacturing process.



**Figure 3.10** *Examples of standard-cell-backend (A) and SRAM-frontend (B) MIMIC test structures*

Figure 3.10B shows an example of a SRAM frontend MIMIC that is developed to determine SRAM specific yield loss due to contact misalignment and poly-poly shorts or junction leakage. Original SRAM layout is taken and original metal 1 is removed. Metal 1 combs are added that are connected to the contacts to active and poly. Neighboring poly polygons are connected to different comb segments to be able to measure poly-poly shorts. Contacts to active are connected to both combs to be able to measure junction leakage and junction break though voltage.

The advantage of this test structure is that it enables to study the above failure mechanisms in a layout situation specific to SRAM. The structure accurately mimics the SRAM frontend, and the yield of this structure gives a better indication of yield loss in the product due to these failure mechanisms than conventional, repetitive simple test structures.

Again

$$Y_{Frontend\_SRAM\_IP\_block} = Y_{MIMIC\_Teststructure}^{\left( \frac{Area_{SRAM\_IP\_block}}{Area_{MIMIC\_Teststructure}} \right)} \quad (3.11)$$

Can be used to predict the yield loss in other SRAM blocks of different size.

### 3.4 Design parameter extraction: structural layout characterization

The yield of a product is also determined by the susceptibility of the circuit to process conditions that may cause defects. Thus, for evaluation of product sensitivity with respect to a particular kind of yield loss, it is necessary to quantify this sensitivity by extracting the appropriate design attributes from the layout [18-34].

In order to determine the relevant design attributes, an understanding of how physical structures on the wafer interact with particular failure mechanisms is needed. Only based on that knowledge the extraction tools for layout sensitivity analysis can be developed. For example, to predict the yield loss due to open or resistive vias in a product, it makes no sense to extract only critical area for vias as a function of defect size if the main reason for yield loss is a too large within-die variation of via aspect ratios due to CMP imperfections. Ideally, for an ideal via yield loss prediction, extraction tools for all possible via related failure mechanisms need to be implemented. The usual extraction tool development sequence is shown in table 3.3.

	Step	Example
1	Identify failure mechanism	Open vias
2	Determine the layout conditions that are sensitive to the failure mechanism	Due to planarization imperfections there is a thickness variation of inter-metal oxide across the chip. The thickness variation depends on the metal density distribution across the chip. Some vias may therefore have a too large aspect ratio and are not (Completely) etched open.
3	Develop extraction tool	The extraction tool analyses the metal distribution around each via and counts the number of vias for which the metal layout configuration is such that it may result in a "deep via".
4	Extraction from product to determine the product's vulnerability	Sensitive vias in the design are counted and localized

**Table 3.3** *Extraction tool development procedure*

As discussed in the previous chapter, the yield of a product is rarely determined by a single failure mechanism. During the manufacturing of a wafer, many different yield loss mechanisms occur and the yield is therefore a product of a number of factors contributing to the yield loss. During the evolution of the process maturity, the number and types of main yield loss causes change. Consequently, for evaluation of product manufacturability, each failure mechanism needs to be translated into an extraction tool for layout characterization with respect to the individual yield loss terms. However, since not all failure mechanisms are known at any given point in time, it is not possible to



develop a generic or complete extraction tool set. Extraction tool sets will always be based on past yield learning experiences and on new yield loss hypotheses. Furthermore, there exists an almost infinite number of different design attributes of potential interest that can be used for design characterization of a product. For efficiency reasons a choice has to be made with respect to the number and types of attributes that need to be extracted at a certain point in time. Design attributes of interest may be different during for instance process development, yield ramping or process maturity.

In the remainder of this section different layout extraction techniques and the factors that determine the feasibility of these techniques are also discussed.

### 3.4.1 Practical extraction techniques

A product layout is usually described by a collection of polygons. For each polygon the design database contains the contour co-ordinates, the instantiation co-ordinates relative to an origin and the mask layer number. In order to minimize the size of the design database, the repetitive structure of the layout is described in a hierarchical fashion. Cell and block structures therefore need to be described only once and are instantiated when needed.

The extraction of design attributes comes down to counting the number of occurrences of a specific polygon combination. For instance the number of vias or the number of polygons with a certain area are counted. However, most relevant design attributes are not directly available from the layout, but need to be derived by applying special operations on the database. For example to calculate the total amount of gate oxide area in a device, a logical AND needs to be performed between all polygons in the poly layer and all polygons in the active layer. For other design attributes more complex area or edge-based operations are needed such as growing or shrinking of polygons. Because such operations often require large computational effort, much research is done in order to develop more efficient extraction methodologies that minimize the use of computer resources needed while maintaining sufficient accuracy. Important categories of layout characterization techniques are:

#### *1: Design rule checker based polygon operations*

Most layout characterization tools are based on commercially available design rule checking (DRC) software [20,25,33]. The reason for this is that such software is generic and often much experience in the use of the tool is available. The software reads in the GDS file and converts the data into an internal database structure that is optimized for efficient polygon handling and hierarchical searches. Only a rule file in which the polygon operations and the job control parameters are described is necessary to start the extraction. The DRC software will take the GDS file and the rules file, perform the necessary polygon operations and output files from which the results can be parsed. In section 3.4.2 the most common polygon operations that are used for extracting design attributes will be described.

During the conversion from GDS file to the internal data structure, the DRC software generates information on the design hierarchy and cell instantiations that can be used for counting library cells or for calculating embedded memory usage. Often rule files in which the extraction algorithms are implemented are fairly simple to understand and are easily portable from one DRC based tool to another.

### ***2: Stratified Sampling***

Layout extractions based on DRC tools may require large amount of memory and computing time. Depending on the file size, the amount of hierarchy and the type of extractions, large VLSI chips may take days to finish. In an industrial design or manufacturing environment this may lead to unacceptable delays. For the characterization of certain types of design attributes, not all of the data that describes the product layout needs to be analyzed in order to measure its properties. Sample based extraction techniques reduce calculation times significantly [21,22,24,26]. This methodology is based on the notion that IC layouts are usually composed of large blocks of similar layout types. By extracting properties of a sufficiently large number of diverse samples (strata) in which the design parameter shows less variance than the whole population, the characteristics of the whole chip can be estimated. Therefore the method is particularly suited for design parameters such as critical area calculation or via counting for which local variation is less than over the chip as a whole. Because the error bound on estimates based on sampling does not depend on the population size, but on the variance of the population, extraction times do not vary much with size, complexity or hierarchical structure of the design.

A disadvantage of the sampling approach is that it is not possible to localize design attributes.

### ***3: Monte Carlo analysis***

Another method for calculating critical area is the Monte Carlo or “dot throwing” method. In this method a number of defects are introduced in randomly chosen coordinates in the layout. By determining what percentage of these defects created a circuit fault for example by introducing a short or open, the sensitivity of the circuit to a specific failure mechanism can be calculated.

A significant reduction in computing time for critical area can be realized with this methodology. However for other design parameters that require evaluation of more complex polygon operations such as deep vias, this method is not suitable.

## **3.4.2 Extraction toolbox**

As discussed in the previous section DRC based extraction tools such as DRACULA or CALIBRE are widely used in industry. Rule files in which the extraction algorithms are described can easily be understood and implemented in any DRC environment. However, the syntax for rule files for different DRC environments is specific for each tool vendor. In order to describe the extraction algorithms in this thesis in a more generic fashion, definitions of the main polygon

manipulations operations that are frequently used in manufacturability extractions are described shortly in this section. Some examples of extraction algorithms will be discussed in section 3.2.6.

The syntax that will be used in this thesis is as follows:

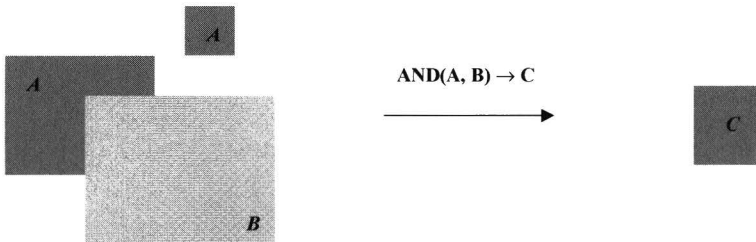
- *Mask layers are indicated within brackets: (A)*
- *Polygon manipulation operations are shown in uppercase*
- *Operation options are shown between [brackets]*
- *The  $\rightarrow$  sign means output of an operation to a new layer*

DRC tools are normally used to find DRC errors in products. Therefore such tools are not intended to calculate and report on the area of certain design attributes as is needed for most extraction tools. In some tools it is therefore necessary the result of an operation is written into a new layer. When the DRC tool generates such a new layer, usually statistics such as total area and the number of polygons are calculated internally and reported into a text file. It is this text file that can then be parsed in order to collect the results of the extractions.

### **Boolean operations**

#### **AND (A, B) $\rightarrow$ (C)**

The AND operation selects all polygons that are common to the two layers A and B and places them into a new layer C. The AND operation is used for example to calculate gate oxide area by calculating the area of AND (poly, active).



**Figure 3.11** AND operation

#### **OR (A, B) $\rightarrow$ (C)**

The OR operation merges all the polygons of the input layers and places them into a new layer C.

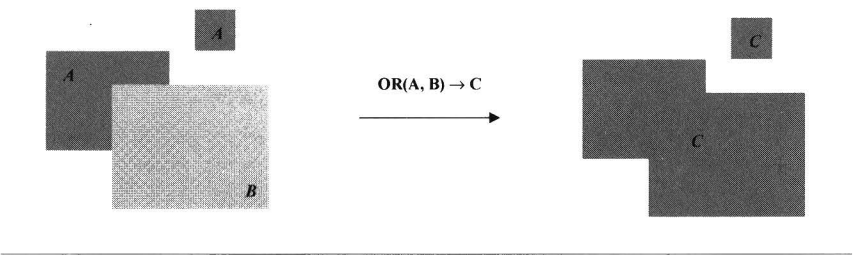


Figure 3.12 OR operation

**XOR (A, B)  $\rightarrow$  (C)**

The XOR operation selects all polygons that are present in only one of the layers. The XOR operation is often used to check whether two designs are the same. C will be empty if A and B are exactly equal.

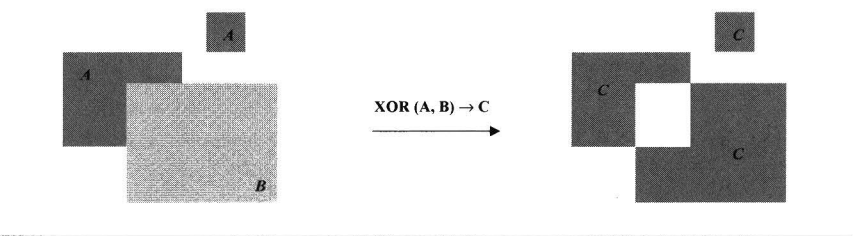


Figure 3.13 XOR operation

**NOT (A, B)  $\rightarrow$  C**

This NOT operation selects all areas in A that are not common to B. The NOT operation is often also referred to as DIFF(A, B) or A MINUS B.

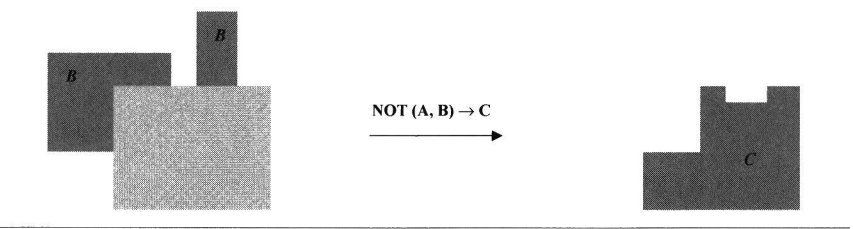


Figure 3.14 NOT operation

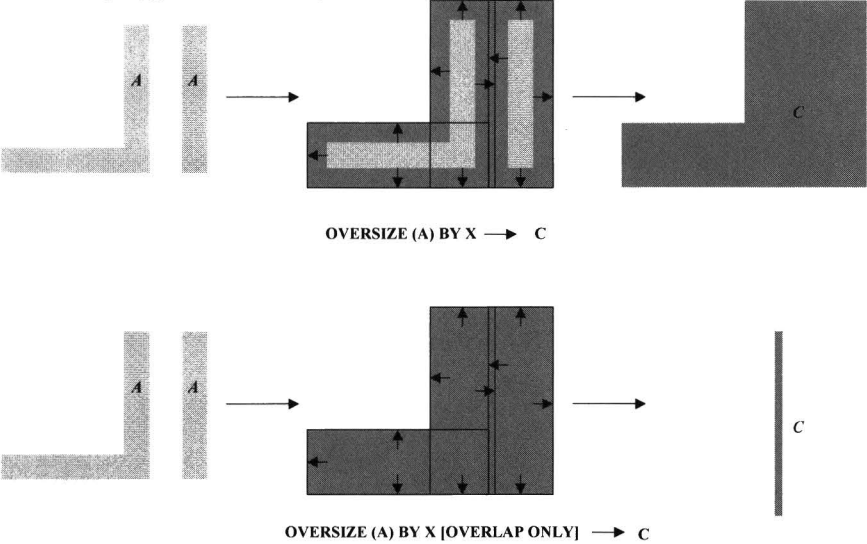
**Sizing operations**

Sizing operations are frequently used in critical area calculations or in algorithms that look for polygons of a certain width. For the purpose of clearly describing

extraction algorithms a distinction is made between oversizing and undersizing. In practical DRC tools, undersizing is done by sizing with a negative value.

**OVERSIZE (A) BY X [OVERLAP ONLY] → (B)**

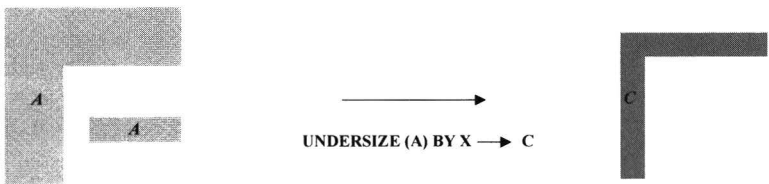
The OVERSIZE operation oversizes all polygons in layer A by X and puts the output in a new layer C. If the OVERLAP ONLY option is switched on, the output will only consist of regions where the oversized polygons overlap (not the oversized polygons themselves).



**Figure 3.15** *Oversize operation*

**UNDERSIZE (A) BY X → (B)**

The UNDERSIZE operation undersizes all polygons in layer A by X and puts the output in a new layer C.



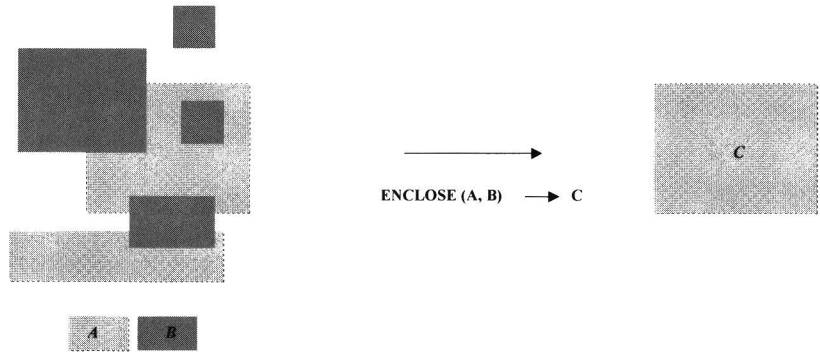
**Figure 3.16** *Undersize operation*

The undersize operation is frequently used in combination with the oversize operation to find polygons with a certain width.

*Other operations*

**ENCLOSE (A, B) → C**

Selects all polygons in layer A that completely enclose polygons in layer B and puts the result in layer C.



---

**Figure 3.17** *Enclose operation*

---

**PRINT\_AREA (A) → file**

Measures the total area of all polygons in layer A and outputs this number into a text file.

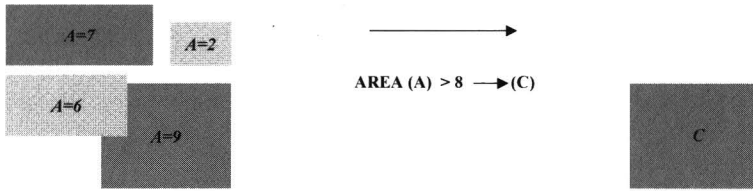
**PRINT\_NR\_OF\_POLYGONS (A) → file**

Measures the total number of polygons in layer A and outputs this number into a text file.

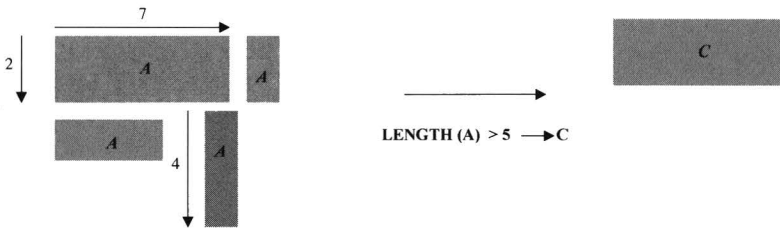
*Measuring operations*

**AREA (A) *constraint* → (C)**

Selects all polygons in layer A that have an area conforming to the constraint and output them to a new layer.

**Figure 3.18** Area operation**LENGTH (A) constraint  $\rightarrow$  (C)**

Selects all polygons in layer A that have a length conforming to the constraint and outputs them to a new layer.

**Figure 3.19** Length operation

### 3.4.3 Examples of design attribute extraction algorithms

In this section some examples of basic extraction algorithms are described.

#### *Extraction transistor related parameters*

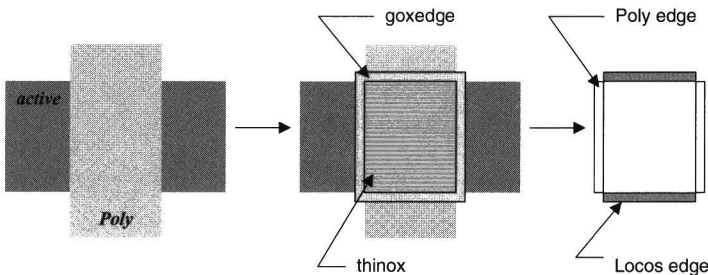
The following algorithm can be used to extract transistor the total gate oxide area, the total length locos edge, and the total length of poly edge in a product. These parameters can then for example be used to estimate the IDDQ current of the circuit.

```

AND (POLY, ACTIVE) → (THINOX)           /* gateoxide area
SIZE (THINOX) BY  $\delta$  → (BLWNOX)         /*  $\delta$  is small : 0.1
 $\mu\text{m}$ 
NOT (BLWNOX, THINOX) → (GOXEDGE)        /* obtain edge
AND (GOXEDGE, POLY) → (LOCOSEDGE)      /* obtain locos
edge
AND (GOXEDGE, ACTIVE) → (POLY EDGE)    /* obtain poly edge
PRINT_AREA (THINOX) → file 1           /* output to file
PRINT_AREA (LOCOSEDGE) → file2         /* output to file
PRINT_AREA (POLYEDGE) → file 3         /* output to file
PRINT_NR-OF POLYGONS (LOCOSEDGE) → file4 /* output to file

```

All polygons in (locosedge) and (polyedge) have a width of  $\delta \mu\text{m}$ . Therefore the total length of poly edges on gate oxide area can be calculated by multiplying the edge area by  $1/\delta$



**Figure 3.20** Transistor related design parameter operation

#### *Extraction of critical area for shorts*

Below an example of a simple extraction algorithm for critical area for shorts for defect sizes of 0.2 to 12  $\mu\text{m}$  is given. Figure 3.21 shows how this is done using the

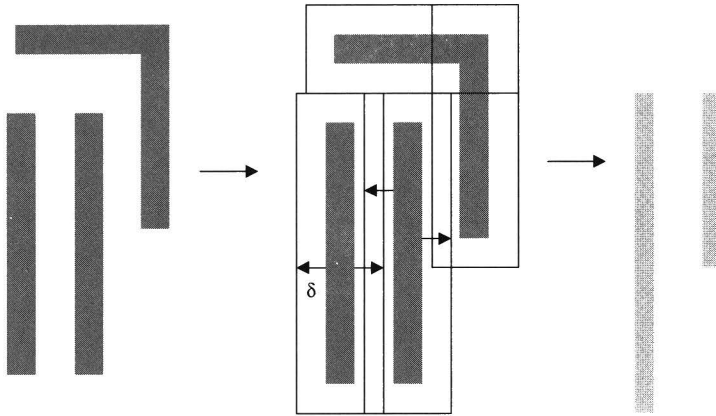


oversize and undersize operations. For large defect sizes computation effort may be large depending on the DRC tool used.

```

For K = 1 to 30 do
  BEGIN
    OVERSIZE (METAL) BY ( $K*0.2$ ) OVERLAP ONLY  $\rightarrow$  (Ck)
    PRINT_AREA (Ck)
  END;

```



**Figure 3.21** *Extracting critical area for shorts*

#### ***Extracting a polygon width pareto***

The extraction algorithm below extracts a pareto of METAL width from 0.5 to 5  $\mu\text{m}$ .

```

For K = 1 to 10 do
  BEGIN
    UNDERSIZE (METAL) BY ( $K*0.5$ )  $\rightarrow$  (MU) /*METAL  $\leq$  K  $\mu\text{m}$  goes
                                           away
    OVERSIZE (MU) BY ( $K*0.5$ )  $\rightarrow$  (MR) /* obtain large metal
    NOT (METAL,MR)  $\rightarrow$  Mk /*obtain METAL with width
                            $\leq$  K
    PRINT_AREA (Mk)
  END;

```

#### ***Extracting number of vias***

The extraction of opens for vias and contacts has extensively been described in literature. In the example below an algorithm for extracting vias used in signal routing is given. First the vias used in signal routing are separated from via banks

that are used in power routing and bondpads. It is assumed that all routing is done in tracks of 1  $\mu\text{m}$  or smaller.

```
UNDERSIZE (METAL ) BY (0.5) → (MGROW) /*METAL <= 1  $\mu\text{m}$  goes
                                         away
OVERSIZE (MGROW) BY (0.5) → (MLARGE) /*only large metal
                                         polygons remain
NOT (METAL, MLARGE) → (SMALLMETAL) /*only tracks < 1  $\mu\text{m}$ 
                                         remain
AND (SMALLMETAL) → (ROUTINGVIA) /* obtain routing vias
NOT (VIA, SMALLVIA) → (VIABBANKS) /* obtain rest of vias
PRINT_NR_OF_POLYGONS(ROUTINGVIA) /*Output
```

These are just a few examples of how extraction scripts can be implemented. Many implementations that give the same results are possible.

In the next section the development of an industrial system that is used for extensive product design characterization is discussed.

### 3.5 Development of a Manufacturability Assessment Environment (MAE)

As discussed earlier, in a multi-product fab-line, many different products that originate from various design sources are manufactured. Often these products are designed with a variety of different requirements in mind, and therefore different design styles and tools are used. This leads to structural differences between products with respect to for instance, metal uniformity, metal coverage or local metal density. Such variations may result in product specific behavior during process steps such as etching, lithography and planarization. Also differences in design methodologies with respect to performance related issues such as decoupling strategy, clock frequency, floor planning, core libraries or embedded memories, may lead to product dependent vulnerability. Consequently, considerable structural differences between different products may exist while they are manufactured in the same process. Prior to manufacturing such structural differences are usually difficult to anticipate on, and in combination with increasingly complex process technologies they often lead to subtle design-process interactions that cause unexpected yield loss. The success of introducing new products in a manufacturing line therefore is uncertain, especially when the market demands steep volume ramp-up. In such cases it is therefore crucial to be able to quickly characterize each new product that is going to be manufactured. The design characteristics can then be compared to other products that are manufactured in the same process, and deviations from “the average” product can be identified. In such a way possible problems can be anticipated on, and the product introduction risk can be minimized.

This chapter describes a manufacturability assessment environment (MAE), called MAPEX-II, that has been developed to address the above needs [5]. The system is based on ideas that were implemented in MAPEX-I which was developed at Carnegie Mellon University [1,2,3].

Since the MAE system has been the basis for much of the data presented in this thesis, additional arguments for the development and the implementation of the system are described in the following sections.

### 3.5.1 Motivation for the development of a MAE

Among the possible applications of a manufacturability assessment environment are:

*Manufacturing environment:*

- Yield forecasting for planning of manufacturing volume
- Yield analysis during process development and yield ramping (priority setting)
- Normalizing defect densities enabling process or fab benchmarking
- Product yield risk assessment
- Fast defect localization for failure analysis
- Process control
- Building a historical database on the design characteristics of products

*Design environment:*

- Assessment of manufacturability implications of decisions in the design of standard cells, memory generators and IP blocks. Comparison of different design styles with respect to yield
- Assessment of economic viability of a product under consideration and yield forecasting for planning of manufacturing volume

Below each of these applications will be explained.

#### 1- *Yield forecasting for planning of manufacturing volume*

As discussed in 2.3.1, often the yield capability of a manufacturing process is expressed in terms of an average defect density,  $Do$ , which is estimated from the yield and die area of different products using for example a Poisson yield model. Often only the die area and yield of a few high volume products are needed to calibrate the model and to obtain a first order indication of the yield capability of a manufacturing process. The yield and manufacturing costs of any new product can be predicted using the calculated  $Do$ . Therefore, manufacturing lines often are required to commit to a certain  $Do$  level or trend for a given manufacturing process over a certain period of time.

However, in a multi-product manufacturing process it is likely that different products are designed with different boundary conditions in mind. Differences with respect to for example time-to-market or circuit performance requirements lead to different design trade-offs and decisions during the development phase of a product. Different design styles can easily cause differences in sensitivities to various yield loss mechanisms. For example product performance requirements may require different choices of library cells or design tools, which may introduce specific sensitivities. Other differences often spring from product history. Some (parts of) products may have been shrunk, compacted or acquired from different IP core vendors. For some products only extremely high transistor density may lead to acceptable profit margins while for other products time to market is crucial for the product's financial success. Such considerations may lead to substantial differences in the design styles with respect to the amount of effort

put into increasing the transistor and routing density on the device. This will eventually lead to differences in defect sensitivity.

For that reason A MAE system that characterizes incoming products is needed for any yield prediction methodology that is meant to plan manufacturing volume.

## 2- *Yield analysis during process development and yield ramping (priority setting)*

Another area of application of a MAE is driven by the need for establishing the correlation between of test structure data and product yield data. During the development of new manufacturing processes or during yield ramping, often only special test structure reticles are used. In order to set priorities in process improvement activities it is important to understand how the yield signals obtained from such test structures relate to product yield.

## 3- *Normalizing defect densities; enabling process or fab benchmarking*

A third reason for exhaustive product characterization is driven by the need for identification of adequate defect density trends of manufacturing processes. Usually Do measurements are based on a high volume product running in a particular process. However, since product lifetimes get increasingly shorter, the product on which Do calculations are based, needs to be changed frequently. Due to the differences in sensitivity between such monitor products, discontinuity in Do trends may arise, although the process yield capability stays unchanged. Therefore there is a need for a normalized Do measurement in which the products sensitivity is taken into account. This can only be achieved if predictive yield models and the related design parameters are available.

## 4- *Product Yield Risk Assessment (PYRA)*

In a multi-product manufacturing environment layout extractions for manufacturing risk assessment are very useful to enable engineers to check whether or not an incoming product is significantly different from the products that are already in production. This may give up-front indication of possible manufacturing problems. For instance the types of library cells that are used in a design can be extracted to check whether the new product contains cells that are not yet yield-wise verified in other products. The same holds for IP-cores and embedded memories. In the case of a yield burst caused by design marginality in a certain block, cell or embedded memory, it is then straightforward to check whether other products that are being manufactured at that moment contain similar cells. Then appropriate actions can be taken so that the yield loss can be constrained.

Another example of checking the similarity between products is the extraction of the pattern density for all layers in a design. If for a new product a substantial deviation from nominal products is detected, it may be useful to give the first lot of the product special attention during the corresponding processing steps to reduce the risk of miss processing. For instance product dependent etching recipes may be applied. Also, local poly and metal densities can be extracted in order to predict within-die variation of inter-metal oxide thickness due to CMP

imperfections. In this way possible problems during the planarization or via formation process can be predicted.

### 5- *Fast defect localization for failure analysis*

Physical failure analysis can be accelerated significantly by the immediate availability of the coordinates of worst case locations for certain failure mechanisms in a product. For instance the coordinates of worst-case vias and contacts with respect to aspect ratio can be extracted so that failure analysts can navigate automatically to the right location with SEM or FIB machines. Without the extraction data readily available, the random search for such worst-case vias is a very time consuming operation that will reduce the speed of yield learning. Worst case coordinates of sensitive layout configurations have effectively been used for automatically guided in-line SEM inspections [4].

### 6- *Process control*

Although for every manufacturing process an attempt is made to develop generic process recipes, some recipes or tools need to be tuned to specific products. Especially processes such as etching or ashing that are sensitive to the density of structures on the wafer sometimes need product dependent recipes. Up-front knowledge on product layout can accelerate recipe development significantly.

### 7- *Building a historical database on the design characteristics of products*

Process architecture and design rules for new processes originate from design density requirements of next generation IC's. A database containing design attributes of all products that are made in a certain technology can be used to verify whether these requirements are met. Also trends with respect to for example embedded memory usage, routing density, transistor density or usage of library cells across process generations can be studied. Without systematic use of a MAE this kind of data is very hard to obtain.

### 8- *Assessment of manufacturability implications of decisions in the design of standard cells, memory generators and IP blocks.*

Often designers go through several iterations before a library cell, memory cell, or IP block is finished. Usually, several options are tried and the best option with respect to performance and area is chosen. Using a MAE enables the designer to bring in yield considerations into his tradeoff as well.

Also yield performance of for example different vendors of IP blocks can be assessed. Examples are given in chapter 6.

### 9- *Assessment of economic viability of a product under consideration*

When a product layout is finished the designer can compare the result to other products with respect to yield and verify whether the yield of the product will not reduce its economic feasibility.

Also, using a MAE the yield prediction will be more accurate resulting in a better tuning of the required manufacturing volume to the market conditions.

### 3.5.2 The MAPEX-II system

As has been already stated, for adequate manufacturability assessment an extensive set of extraction tools is needed that extracts a large number of relevant attributes from the layout. In cases where product characterization is required for a large number of products, in addition to the extraction tools, a certain level of automation is needed with respect to the extraction itself, but also with respect to the storage and analysis of the extracted design data. This section describes an automated system that has been developed to serve in an industrial manufacturing facility that produces many different products in several processes, from 0.5  $\mu\text{m}$  to 0.18 $\mu\text{m}$  technologies, with various process options. In this particular fab line more than a hundred different products are introduced each year and all incoming products are to be characterized before start-up of the first wafers. The system is called the MAPEX-II (Manufacturability Assessment Parameter Extraction) software framework. The system evolved from the MAPEX [3] software in a period of two years and was extensively used to prepare the material presented in this thesis.

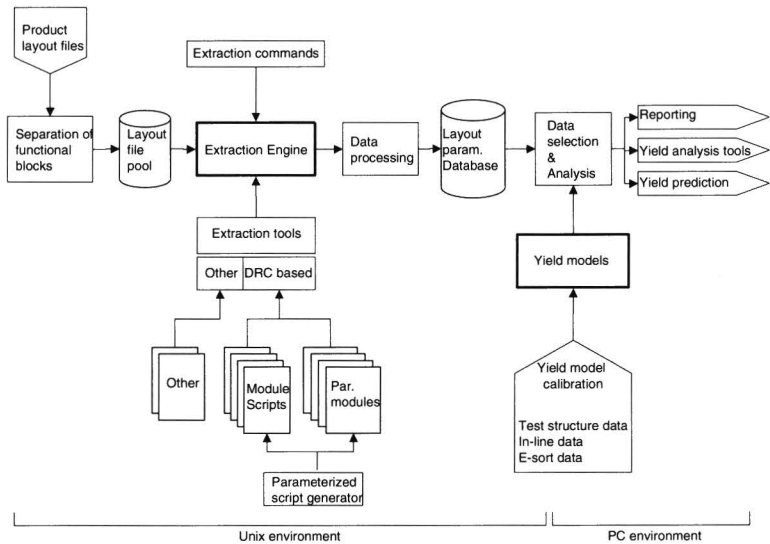
#### *Requirements of the system*

The following requirements have been taken into account for the development of the system:

- For incoming products all relevant extractions should be started up automatically with minimal user interference.
- Output data should be stored in a centrally accessible database without any user intervention. The extracted data of any product should be available at any time to a large variety of people from various disciplines such as product engineers, process development engineers, design engineers and account managers.
- The system should be able to fully characterize at least one average sized product within 24 hours.
- Removal or insertion of additional extraction tools and related yield model (parameters) in the system should be possible for the user with a minimum of effort. This is especially important when new design-process interactions are suspected and the sensitivity of products needs to be assessed rapidly. .
- Default values for the process related yield model parameters such as defect densities and defect size distributions should be available for each manufacturing process and should be changeable by the user.
- Basic manufacturability reports should be generated automatically by the system.
- Output to other yield analysis tools or spreadsheets is possible.

System Implementation

According to the above specification MAPEX-II software environment has been developed. The structure of the system is shown in figure 3.22.



**Figure 3.22** Structure of the MAPEX-II system

Below each of the different blocks in the MAPEX-II structure will be described.

*Separation of functional blocks*

In order to assess sensitivities for different blocks within a product they need to be separated from the original layout. Each block can then be considered as a separate design and stored in the gds file pool. The methodology for separating blocks from a product depends on the hierarchical description of the layout. If blocks are on separate hierarchic levels they are extracted from the hierarchic tree by referring to their structure name. If the different blocks are on the same hierarchic level they can not be extracted by name and they have to be filtered out using a DRC tool that performs an AND operation with an exclusion mask with the appropriate dimensions and output the result to a new layout file.

*Extraction Engine*

The core of the system is the extraction engine. When it is given the command to extract layout attributes from a product, it looks for the layout file in the GDS file pool, it will determine the process the product is manufactured in and it will start up the appropriate extraction tools. When the extractions are finished, the engine



will prepare a file for database uploading. Since the engine starts many different extractions and soft and hardware resources may be limited, a queuing system is used for job handling and priority setting of extraction jobs.

#### *Extraction tools*

DRC based extraction tools in MAPEX-II consist of two parts: a rules module file and a corresponding parameter file. The rules file only describes the operations that need to be done on the layout data. The parameter file then determines what data is to be parsed from the DRC output files. During a DRC based extraction, the engine starts up a new program shell that will take the extraction rule file to generate a product specific job file that is suitable for the DRC tool. The job file is then started and put in the queuing system. When the extraction is finished, the shell will determine from the corresponding parameter file what data needs to be parsed from the DRC output and forwarded to the engine. This mechanism ensures flexibility and modularity and minimizes the amount of programming involved when new extraction scripts need to be developed. Only a new rules module and a new parameter module have to be made. Appropriate job file creation and data parsing from output files is handled automatically.

In case of extraction tools that are not DRC based, (for instance tools that search the hierarchic tree for specific cell names) the engine will simply start-up the tool and put the output data in a file for database uploading.

There exists a large number of different design attributes of potential interest for design characterization. However, for efficiency reasons a choice has to be made with respect to the number and types of attributes that need to be extracted at a certain point in time. Design attributes of interest may be different during for instance process development, yield ramping or process maturity. Table 3.4 lists the extraction tools that have been developed for the MAPEX-II system.

<b>Extracted design attributes</b>	<b>Purpose</b>
<b>General layout information</b>	
Die area	YP, MRA
Number of bondpads and coordinates	Test engineering
<b>Coverage / Density</b>	MRA, PD
Mask coverage / pattern density	Litho and etch recipes, MRA
Size of "white" area (area where nothing is drawn)	PD
<b>Library cells</b>	
Count of all core library cells	MRA, YP
Count of all IO cells	MRA, YP
<b>Embedded memories</b>	
Types: SRAM, ROM, OTP, DRAM	YP, MRA
Nr of blocks	YP, MRA
Number of cells	YP, MRA
Total area of the memory	YP, MRA
<b>Transistors</b>	
Number of N and P transistors	YP, MRA
Total gate oxide area	YP, MRA
Transistor size distribution	YP, MRA
Gate oxide edge length distribution (locos/STI)	YP, MRA, iddq
Gate oxide edge length distribution (active)	YP, MRA, iddq
<b>Contacts and vias</b>	
Nr of vias1-5	YP, MRA
Nr of non-redundant vias	YP, MRA
Number of stacked vias	YP, MRA
Number and coordinates of deep vias and contacts	YP, MRA, FA
Number and coordinates of shalow vias and contacts	YP, MRA, FA
Number and coordinates of lonely vias	YP, MRA, FA
Metal over Via overlap distribution	YP, MRA
<b>Critical area as a function of defect size for shorts and opens</b>	
Shorts in conducting layers (metals, Poly, Active, LIL) With and without connectivity information	YP, MRA
Opens in conducting layers	YP, MRA
<b>Metal</b>	
Metal coverage map	MRA, CMP
Wire (net) length distribution	YP, MRA
Wire width distribution	YP, MRA
<b>Design rules</b>	
For the most important design rules: Number of occurrences on minimum design rule and on minumum design rule+ 1,2 and 3 grids	YP, MRA, PD
<b>Special items</b>	
Charging sensitivity, antenna ratios	YP, MRA
<b>Windowing</b>	
Windowing for all parameters (layout is divided into smaller portions. For each portion the above parameters are extracted	YP, MRA

**Table 3.4** Basic list of layout extraction capabilities YP=Yield Prediction; MRA = Manufacturing Risk Assessment; PD=Process Development; FA=Failure Analysis

### Data retrieval

The front-end of the system is used to extract data from the database and to do a first analysis of the data. It is implemented on a PC platform in order to increase availability of the data to a wide range of people. An example of the data selection window is shown in figure 3.23. The user can select certain products and design parameters corresponding to those products. The front-end generates the necessary database queries to extract the data from the database and inserts the data in a spreadsheet-like tool to do the yield predictions and to produce the standard reporting.

The MAPEX-II system has several yield models such as the Poisson model, the negative binomial model, and the critical area model at its disposal to perform yield calculations for each extracted design parameter. Which yield models apply to which layout parameter can be determined by the user. Also the process dependent parameters such as defect densities and defect size distributions can be set. For each process a default set of values is available that is calibrated through test structure or in-line data. See also figure 3.2.

The screenshot shows a graphical user interface for data extraction. It features several panels and buttons:

- Process:** A list box containing items like C050HM, C050MTP, C075EE, C075FC, C075FM, C0750TP, C1000TC, C1000TP, C1000TC, C1000TL, and C1000TL.
- Customer:** A list box containing items like AM, APIC, BU, CERNUM, C0CAEN, C0CHAM, C0CONUM, C0CSOTON, C0CTAIP, C0CU, and CS DM.
- Products:** A list box containing items like P0000-v01, P0001-v01, P0007-v01, P0005-v01, and P0006-v01.
- DFM Parameter groups:** A list box containing items like cana, cell\_count, coverage, covas, galov, general info, and mem\_count.
- Parameters:** A list box containing items like CA\_ACT\_0.15, CA\_ME1\_0.15, CA\_ME1\_0.2, CA\_ME1\_0.25, CA\_ME1\_0.3, CA\_ME1\_0.35, CA\_ME1\_0.4, CA\_ME1\_0.45, CA\_ME1\_0.5, CA\_ME1\_0.55, CA\_ME1\_0.6, CA\_ME1\_0.8, CA\_ME1\_1, CA\_ME1\_1.2, CA\_ME1\_1.5, and CA\_ME1\_10.
- Experimental Data:** A section with checkboxes for ☒ None, ☐ Include, and ☐ Only.
- Output To:** A section with checkboxes for ☐ Advisor, ☐ ASCII File, and ☒ Table.
- Buttons:** Select Par. Group, Select products, Select Parameters, Exit, View, and Generate Output.
- Timeframe:** A section with From: 21 Jan 96 and To: 21 May 99.

**Figure 3.23** Example of an extraction data retrieval window

### 3.6 Conclusions

Adequate use of yield models plays an important role in the design and manufacturing of ICs. On the manufacturing side, yield ramping can only be achieved if major failure mechanisms are quickly identified, and more importantly, if the impact of these failure mechanisms on product yields can be assessed. Only then correct priorities can be set in possible improvement actions. On the design side, yield models enable designers to quantify the yield impact of design decisions they make.

For adequate use of such yield models it is essential to characterize the relation between each individual failure mechanism that can be identified in the manufacturing process and the corresponding sensitivity of the products to that failure mechanism. The yield prediction methodology described in this chapter has proven to play an important role in achieving this goal. The methodology comprises two major parts: product design characterization and process or defect characterization. For both types of characterization knowledge on known failure mechanisms is the starting point. Product *design characterization* is straightforward in the sense that much software is available on which layout extractions can be based. The difficulty is more in the choice of design attributes that are relevant to extract. Development of extraction tools costs can become non-negligible and a sensible subset of the total set of extractable design attributes needs to be chosen according to the goals the user tries to achieve. For yield ramping purposes for example, the extracted attributes may be different from the extractions needed during the design of a product.

*Process characterization* is far more difficult, and therefore usually the costs for calibrating the yield models in this respect are much higher. The methodology described in this chapter is based on a combination of a modeling and empirical approach to extract process related yield model parameters. By extrapolating test structure results for certain design attributes according to the failure models, resources for silicon, test and analysis can be limited. For example, to extract defect size distribution to be used in a critical area model, not all possible metal spacings need to be used, but the results of three spacings is usually enough to extrapolate. In that way the costs are limited.

Depending on the accuracy of the yield models and the level of detail of the yield loss breakdown that is needed, the extraction of yield model parameters can become very costly and needs to be justified in relation to the goals the user tries to achieve. The considerations and tradeoffs that are needed in this respect have been discussed in this chapter.

A state of the art manufacturability assessment and yield prediction system was developed and implemented in a CMOS manufacturing environment. As is shown throughout this thesis, the system clearly identifies significant layout differences between products that may lead to specific product-process marginalities, enabling up-front anticipation on product dependent yield loss.

Automation of layout extraction and data storage is crucial to the success of the system. Especially in a multi-product environment, the immediate availability of design related information of any product is of great value. Also the flexibility to develop and add new layout extraction tools to the system is important. This is

particularly the case during process development where subtle design related marginalities are a significant part of the yield pareto. Hypotheses for new failure mechanisms can easily be verified by correlating yield trends of functional blocks with their corresponding design sensitivities. Extensive use of the system has shown that the system provides product and yield engineers with a valuable source of information that would not have been available otherwise. The availability of detailed information on design-process interactions enables a substantial acceleration of the yield learning process that is crucial in today's semiconductor market

## References Chapter 3

1. W. Maly, H. T. Heineken, J. Khare, and P. K. Nag, "Design-Manufacturing Interface: Part I - Vision," Proc. of Design, Automation and Test in Europe Conference 1998 (DATE '98), pp. 550 - 556, Paris, France, Feb. 23-26, 1998.
2. W. Maly, H. T. Heineken, J. Khare, P. K. Nag, P. Simon, and C. Ouyang, "Design-Manufacturing Interface: Part II - Applications," Proc. of Design, Automation and Test in Europe Conference 1998 (DATE '98), pp. 557-562, Paris, France, Feb. 23-26, 1998.
3. H.T. Heineken and W. Maly, "Manufacturability Analysis Environment - MAPEX," in Proc. of the 1994 Custom Integrated Circuits Conference, May 1994, pp. 309-312.
4. P. Simon, W. Maly, D. K. de Vries, E. Bruls, "Design Dependency of Yield Loss Due to Tungsten Residues in Spin on Glass Based Planarization Process," in Proc. of ISSM, San Francisco, Ca., Nov. 1997.
5. Paul Simon, et al. "Layout based manufacturability assessment and yield prediction methodology" Conference on In-line Characterization techniques for performance and Yield Enhancement in Microelectronic Manufacturing, Edinburgh, Scotland, May 1999, pp.282-288.
6. J.B.Khare, W.Maly and M.E.Thomas," Extraction of defects size Distribution in an IC Layer Using Test Structure Data", IEEE transactions on Semiconductor Manufacturing, Vol.7, no.3. August 1994, pp.354-368.
7. Christopher Hess and Larg H. Weiland, "Extraction of wafer-level Defect Density Distributions to improve Yield Prediction", IEEE Transactions on Semiconductor Manufacturing, vol.12.No.2, May 1999, pp.175-183.
8. Christopher Hess and Larg H. Weiland, "Harp Test Structure to Electrically Determine Size Distributions of Killer Defects", IEEE Transactions on Semiconductor Manufacturing, vol.11.No.2, May 1998, pp.194-202.
9. Christopher Hess and Larg H. Weiland, "Customized Checkerboard Test Structures to Localize Interconnection Point Defects", Proc. VLIS Multilevel Interconnection Conference, Vol. 14, June 1997.
10. F. Camerik, "Qualification and Quantification of Process-Induced Product-Related Defects", International Test Conference, 1989, pp.643-651.
11. William.J.J. Rey, "What area should the defect monitor be", Internal Report, Philips Research Labs Eindhoven.

12. Crid YU et.al., "Use of Short-loop Electrical Measurements for Yield Improvement", IEEE Transactions on Semiconductor Manufacturing, vol.8.No.2, May 1995, pp.150-159.
13. M.A. Mitchel et.al., "issues with Contact Defct Test Structures", Proce. IEEE 1992 Int. Conference on Microelectronic Test Structures, Vol 5, March 1992, pp.53-56.
14. T. Hamamoto et.al. "Measurement of Contact Resistace Distribution Using a 4K Contact Array", Proc. IEEE 1995 int. Conference on Microelectronic Test Structures, Vol 8, March 1995, pp. 57-95
15. M.M.A. van Rosmalen, K.Baker, E.M.J.G. Bruls, "Parameter Monitoring: Advantages and Pitfalls", Int. Test Conference 1993, pp.115-123.
16. E.Bruls, "Quality and reliability impact of defect data analysis", IEEE Transactions on Semiconductor Manufacturing, vol8, no. 2, pp.121-129, May 1995.
17. Linda S. Milor, "Yield Modleing Based on In-Line Scanner Defect Sizing and Circuit's Critical Area", IEEE Transactions on Semiconductor Manufacturing, vol12, no. 1, pp.26-34, Feb 1999.
18. Charles Kooperberg, "Circuit layout and Yield", IEEE Journal of Solid State Circuits, vol. 23, No. 4, August 1988, pp. 887-892
19. Witold A Pleskacz, Charles H Ouyang and Wojciech Maly, "A DRC-Based Algorithm for Extraction of Critical Area for Opens in Large VLSI Circuits", IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems. Vol. 18, No. 2, February 1999, pp 151-161
20. Charles H Ouyang , Witold A Pleskacz and Wojciech Maly, "Extraction of Critical Area for Opens in Large VLSI Circuits", 1996 IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems.
21. G.A. Allan and A.J. Walton, "Efficient critical area algorithms and their application to yield improvement and test strategies",IEEE workshop on defect and fault Tolerance in VLSI Systems, Montreal, Quebec, Canada, Oct 1994, pp.88-96.
22. G.A.Allan, , J.P. Elliot, and A.J. Walton, "A defect sensitivty measurement tool enabling comparison of multilevel interconnection strategies",VLSI Multilevel Metal interconnection Conferene, Santa Clara CA, June 1995, pp. 655-657.
23. M.J.Lorenzetti, P. Magil, A. Dalal and P.Franzon."McYield: A CAD tool for functional yield projections for VLSI", IEEE workshop on Defect and Fault

- tolerance in VLSI Systems, Institut National Polytechnique de Grenoble, Nov 1999, pp.100-110.
24. G.A. Allan and A.J. Walton "Hierarchical critical area extraction with the EYE tool,"IEEE Workshop on Defect and Fault Tolerance in VLSI Systems, Lafayette, Louisiana, Nov 1995, pp. 28-36
  25. P.K.Nag and W.Maly "Hierarchical extraction of critical area for shorts in very large scale IC's", IEEE workshop on on Defect and Fault Tolerance in VLSI Systems, Lafayette, Louisiana, Nov 1995, pp. 19-27.
  26. F.Dudivier and G.A. Allan, "Application of a survey sampling critical area computation tool in a manufacturing environment", ", IEEE workshop on on Defect and Fault Tolerance in VLSI Systems, Boston, Massachusetts, Nov 1996, pp.48-52.
  27. W. Maly and J. Deszczka, "Yield estimation model for VLSI artwork evaluation,"Electronic letters, vol.19, no.6, pp.226-227, March 1983.
  28. J.Pineda de Gyvez and C.Di,"IC Defect Sensitivity for Footprint-Type Spot Defects",IEEE Transactions on CAD of Integrated Circuits and Systems, Vol.11, No5,pp.638-658, March 1992.
  29. H.Walker and S.W.Director,"VLASIC: A Catastrophic Fault Simulator for Integrated Circuits", IEEE Transactions of Computer-Aided Design", CAD-5(4),pp.541-556, 1986.
  30. D. Schmitt-Landsiedel, D. Keitel-Schulz, J. Khare, S. Griep and W. Maly, "Critical Area Analysis for Design Based Yield Improvements of VLSI Circuits," in Proc. of the 5th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF 94), Glasgow, Oct. 1994.
  31. R. K. Nurani, et. al., "In-line Yield Prediction Methodologies Using Patterned Wafer Inspection Information," Proc. of The 1996 Int. Symp. on Semiconductor Manufacturing, Tokyo, Oct. 1996, pp. 243-250.
  32. C.H. Stapper and R.J.Rossner, "Integrated Circuit Yield Management and Yield analysis: Development and Implementation", IEEE Transactions on Semiconductor Manufacturing, vol.7, no.4, Nov.1994.
  33. Witold A. Pleskacz, et. al., "A DRC Based Algorithm for Extraction of Critical Areas for Opens in Large VLSI Circuits", IEEE Transactions on Semiconductor Manufacturing, vol.18, no.2, Feb.1999.
  34. I.Bubel, et.al.,"AFFCCA: A Tool for Critical Area Analysis with Circular Defects and Lithography Deformed Layout", 1995 International Workshop on Defect and Fault Tolerance in VLSI Systems.



---

## **Chapter 4**

# **Plasma Process Induced Damage Physics and modeling**

# 4

## 4.1 Introduction

The yield modeling techniques and methodologies discussed so far were based on known failure mechanisms. In real life however, new, unknown failure mechanisms occur with progressing technology. In order to understand the yield impact of such new failure mechanisms, they should be studied and characterized. Plasma charging damage is an example of a yield loss mechanism of which the yield impact has not been sufficiently studied. Therefore, in light of the methodologies discussed in previous chapters, the failure mechanism itself and the yield modeling needs to be developed from scratch.

Plasma processes are widely used in semiconductor manufacturing for etching of poly-silicon, oxide, and metal films. High-density plasmas are also used for deposition of oxides. Unfortunately, while a wafer is being processed, the energetic ions and electrons in the plasma can build up charge on the gate oxide in devices. Due to the imbalance of local ion and electron fluxes in the plasma ambient, structures on the wafer that serve as electrodes or 'antennas' accumulate charge. These charges that are built up on the antenna structures may lead to damaging tunneling currents through connected thin gate oxides [1-6]. The transistors that are damaged in this way may either show parametric deviations such as threshold voltage shifts or gate leakage. On product level such defects may affect the performance of the device or cause yield loss due to for instance unacceptably high Iddq values. Plasma induced damage may also degrade the reliability of the product by revealing itself only in a later stage when the gate oxide is stressed during a period of normal operation [4-5].

During the last decade plasma process induced damage has become more of a threat to advanced VLSI devices. With the scaling of technologies, gate oxides have become thinner and the aspect ratios of metal structures have increased. At the same time the number of metal layers and plasma related processing steps have increased as well, making products more susceptible to plasma induced damage. Today, this concern explains a widespread interest in this phenomenon. Much work has been done to understand the damage mechanisms. However, due

to the complexity of the issue, it is still not possible to completely eliminate the plasma damage from any manufacturing process solely by optimizing process parameters. In addition, equipment and process fluctuations in combination with the difficulty of monitoring the parameters relevant to charging make it difficult to guarantee low damage levels at all times.

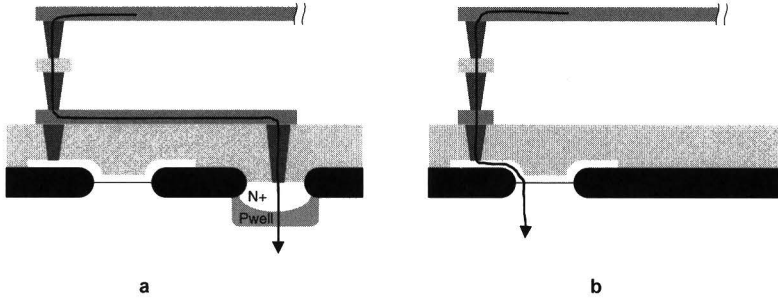
The extent of plasma induced damage, not only depends on process parameters, but also on the sensitivity of the product itself. Products containing high numbers of large antennas are more likely to fail. Therefore, it is important for a designer to be able to realize charging robust products. Design rules or methodologies that are based on the understanding of the relationship between the process conditions that cause charging and the layout configurations that make the product prone to these conditions, are crucial in this respect. Presently, charging robust design is based on simple antenna ratio rules that only take into account the area or perimeter of the structures that are connected to each individual transistor in the product. When a charging sensitive structure is detected, the designer needs to go through a lengthy, often manual operation of removing these structures. In order to prevent antennas from occurring in the first place, designers may also decide to use cell libraries that use a protective reverse biased diode at each gate that is supposed to protect the gate oxide by shunting the charges to the substrate.

This chapter shows that the existing design methodologies for charging robust design are too simplistic and therefore inadequate to address the real problems the designer is facing. Solving the plasma induced damage problem is a typical DfM example for which only an approach that encompasses both the product design and manufacturing conditions, will result in the ability to realize charging robustness.

In this chapter the charging mechanisms and their relationship with yield and reliability loss will be discussed. Then new methodologies for the characterization of the layout dependence of charging and the resulting product sensitivity characterization are described. A new product charging sensitivity index and methodology for charging robust design will be proposed.

## 4.2 Charging failure mechanism

During plasma processing, local or global charge imbalance caused by either plasma conditions itself or by particular layout configurations may cause different charges to be built up on poly or metal structures on the wafer. At the time that each individual layer is created, only part of the connections from all output drains to the input gates is completed. In these cases a leakage path exists from the antenna to the drain area that is connected to it. In most cases the drain area will be large enough to sink the accumulated charge so that no damage is done to the gate oxide. (See figure 4.1a).

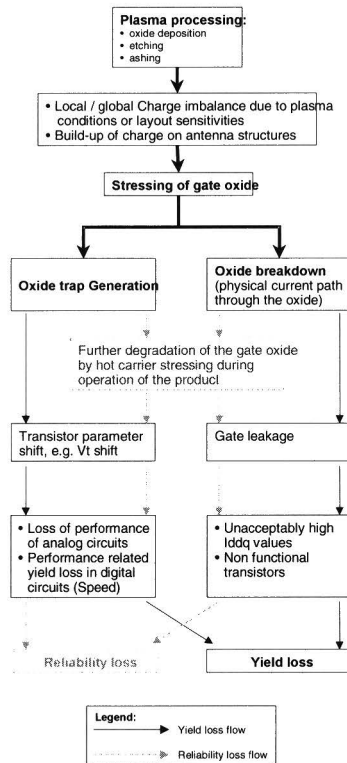


**Figure 4.1** *Metal 3 antenna structure. A: The transistor gate is protected by a drain or diode area. B: No drain area available yet.*

However, until the final metal layer is formed, not all connections from all input gates to output drain areas are formed. For some of the gates only part of the final metal connection will be formed. In such cases there is not yet a leakage path connected to the antenna at the time it is processed, and if the structure is large enough, it may accumulate high levels of charge. (See figure 4.1b). The charge build-up on the antenna will create a steady state voltage on the transistor gate that is connected to it, resulting in electrical stress of the gate oxide. The stress will cause the oxide to breakdown or to degrade by causing new charge trapping in the oxide as well as interface trap generation at the  $\text{SiO}_2\text{-Si}$  interface. The degraded oxide may change transistor characteristics such as sub-threshold voltage and gate leakage currents causing yield loss. Also it is more vulnerable to hot carrier induced degradation and time dependent dielectric breakdown causing reliability failures of the product.

The interface traps generated by plasma damage can be passivated with a subsequent gas annealing step. However, the latent damage is likely to manifest itself in a later stage during circuit operation in the form of degraded hot carrier performance.

The above failure mechanism is summarized in figure 4.2.



**Figure 4.2** Schematic flow of the charging failure mechanisms for reliability and yield loss.

## 4.2.1 Charge imbalance mechanisms

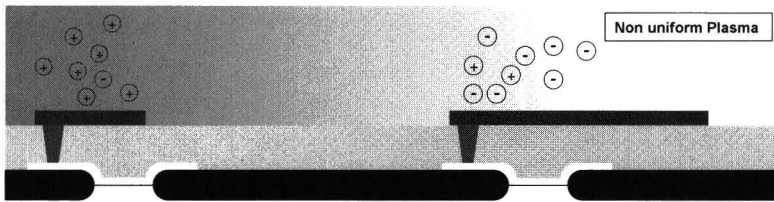
A damaging injection current through a gate oxide is caused by a local or global charge imbalance that can be caused by certain plasma conditions such as plasma uniformity, electron temperature or pressure. The extent of the imbalance also depends on the layout configuration of the antenna in the circuit. For example area, perimeter and surroundings of the antenna play an important role. Often it is difficult to determine which conditions are causing the charge imbalance for a particular process. During the manufacturing, test structures go through different plasma steps such as etching, ashing and deposition, each having different stages with different plasma conditions, and therefore capable of different charging mechanisms. Since a test structure can only be measured at the final stage of the process, it is difficult to distinguish the damage mechanisms it was submitted to for each individual plasma step. For example for a metal antenna test structure it is impossible, without special process changes, to distinguish charging damage

caused by over-etching from charging damage by deposition of the liner or first dielectric oxide layer.

However, several hypotheses for charge imbalance mechanisms have been proposed in literature over the past few years. The most important ones are listed below.

#### 1. *Plasma non-uniformity* [1]

Plasma non-uniformity causes an uneven distribution of charged particles and DC sheath potential across the wafer resulting in damaging currents through the gate oxides. See figure 4.3.

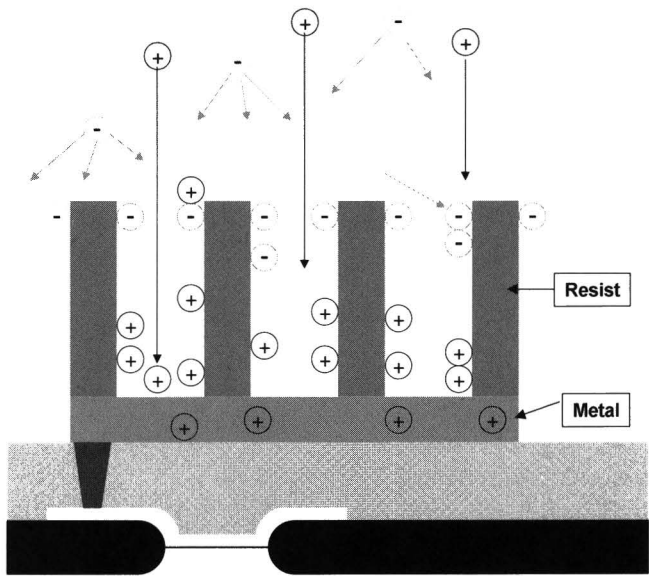


**Figure 4.3** *Plasma non-uniformity leading to currents through the gate oxide.*

#### 2. *Electron shading* [9,11-14]

Charge imbalance due to the electron shading effect is of a more local nature. (See figure 4.4). The effect occurs when the top masking layer (for instance photo resist) charges so that electrons are shaded from the bottom of the trenches being etched. The difference in distributions of angular velocity for electrons and ions cause ions to have a higher probability of entering the tight space between closely spaced lines than electrons [15-16]. As a result, the bottom of the trench is positively charged, which results in excessive tunneling currents through the gate oxide and sometimes also in notching of side walls. Simulation studies have shown the effect of process parameters, aspect ratios and spacing on electron shading [13]. Test structure results show that for the electron shading effect the charging damage increases with narrowing spaces between tracks.

This type of charging is of particular concern since it depends on the design of the circuit and can therefore not be solved by optimizing plasma uniformity. Electron shading can take place during metal etching [7,9,11] as well during oxide deposition [29, 31].

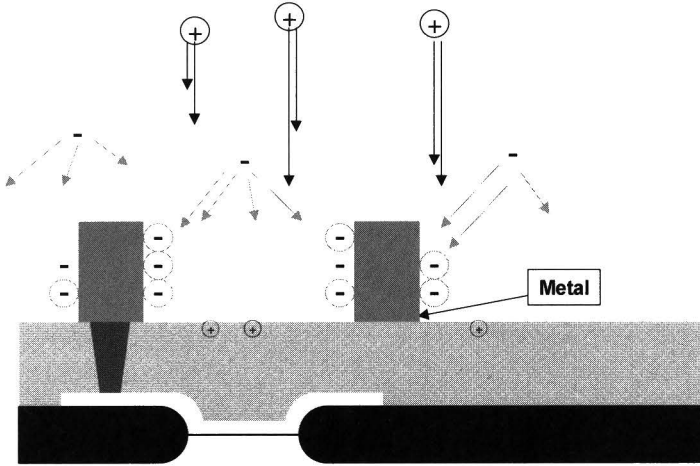


**Figure 4.4** *Electron shading effect.*

3. *Extended/inverse electron shading [22,23,24]*

The extended electron shading is also the result of the difference in angular impact between electrons and ions in the plasma. It occurs during the over-etch of metal or during the deposition of oxide on the metal. See figure 4.5. During these stages of the process the individual metal lines are separated from each other and electrons are impacting the side-walls of the tracks charging them negatively. Contrary to the normal electron shading effect, here the charging effect increases with the line spacing. The larger the opening between the lines, the more electron will charge the side-walls.

Both the electron shading and extended electron shading may be present in the same process, even at the same time. In the later case, some parts of the metal on the wafer will be charged positively due to electron shading and other parts may be charged negatively due to extended electron shading.

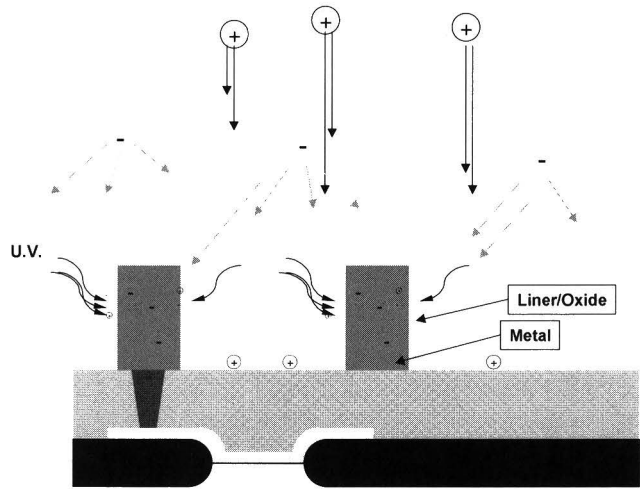


**Figure 4.5** *Extended electron shading effect.*

#### 4. Photoconduction [22,23,29]

During oxide deposition on metal tracks one would expect charging to occur only in the early stages when the metal surface is not covered in oxide yet. Once the metal is covered with the insulator, there is no direct contact of the plasma with the metal and intuitively there is no charging possible. However, in [29, 31] it is shown that after the metal is covered in a blanket of oxide there is still severe charging possible due to the photoconductivity of the oxide due to the UV radiation of the plasma. The severity and polarity of the charging damage depends on the thickness and conformality of the oxide.[22]. (See figure 4.6).

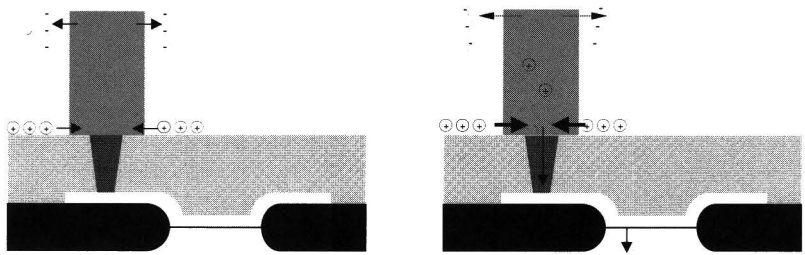




**Figure 4.6** Photo conductivity of the oxide under influence of UV radiation enables the charging of the tracks.

5. Tunneling [30,34]

A second cause for charging to occur during oxide deposition, even after the metal is covered, is the tunneling of charge through the oxide, especially when the oxide deposition occurs in a non-conformal way (the oxide on top and on the edges of the tracks grows faster than on the side-walls). In [35] it is shown through Monte Carlo simulations that metal track charging occurs when the top dielectric is thick enough to prevent tunneling currents, while the side-wall dielectric thickness still allows tunneling current to flow to the metal line. The charging of the side-walls is then caused by electron shading. See fig.4.7.



**Figure 4.7** Tunneling of charge through the deposited oxide.

## 4.2.2 Layout dependency of charging

Depending on how the charge imbalance is built-up, the following layout attributes may be of interest with respect to the extent of the damage:

### 1. *Transistor geometry*

Often the area of the gate oxide area is taken as a measure for the sensitivity of the transistor to be damaged. The idea is that the larger the gate oxide area, the more charge it can endure. However, oxide quality may not be homogenous across the total gate area and local thinning of the oxide may weaken the overall robustness [2]. Also the bird's beak in LOCOS technologies is more vulnerable. Therefore the W/L ratio of the transistor is important [9].

### 2. *Surface area of the conductor*

In the first stages of HDP oxide deposition or metal etching, the surface area of the conductor determines the level of charging in case of plasma non-uniformity. In case of deposition this mechanism lasts until the oxide layer is too thick to allow for charge tunneling through the oxide. At that moment only sidewall effects may play a role.

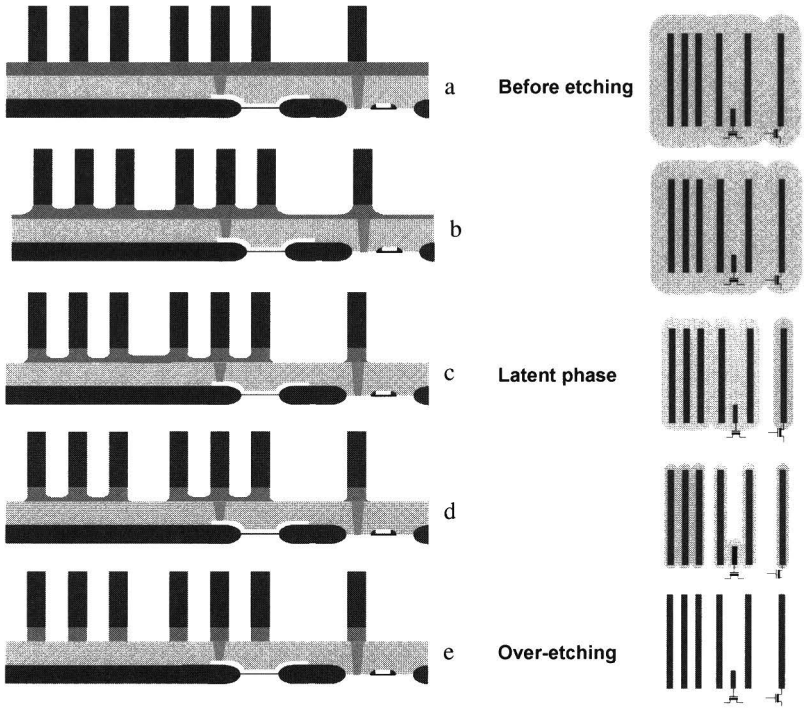
### 3. *Edge length and height of the conductor*

When electron shading is causing the charge build-up, the edge length or perimeter of the conductor determine the charging level. Also the thickness of the tracks in combination with the resist thickness are important parameters.

### 4. *Surroundings of the conductor*

In case of electron shading, the spacing of a track to the neighboring lines is determining the charging level. For electron shading the damage is more severe at small spacings. For the extended electron shading the inverse is true.

Another important layout dependency of etching damage is caused by reactive ion etching (RIE) lag microloading or Aspect Ratio Dependent Etch (ARDE) [15,16]. This effect is explained in fig. 4.8 where the different stages of a metal etch process are shown. During the formation of interconnect metal lines, the wafer is covered with metal. After lithography, the main plasma etch is started (a). At this moment no transistors can be damaged because many drain areas are connected to the metal. As the main etch continues, the etching rate in between metal lines with large spacing is higher than in between closely spaced lines.(b).



**Figure 4.8** Due to the RIE lag effect large ‘islands of connected metal’ are formed during the etching process.

At this moment the charge imbalance that is caused by electron shading is accommodated for by electrons from the substrate coming from a drain of a transistor that is still connected. As the etch continues, (c,d) first larger open areas are cleared and “islands of connected metal” are formed. If such an island is connected to a gate and not to a drain area, latent antennas exist and electron shading may occur. The charge balancing electron current is cut off and the antenna potential will rise. To minimize the charge imbalance, electrons from the substrate will tunnel through the gate oxide and eventually damage the gate oxide. Although it is difficult to distinguish between the two phenomena, the individual charging contributions of RIE lag and electron shading can be studied using dense finger and shaded finger antenna configurations [12]. The latent antennas [11] effect has been studied using transient fuses [9]. The amount of damage is proportional to the total metal area that stays connected during this phase. Once the metal in the trenches is cleared (figure e), all metal lines are disconnected and the over-etching begins. The amount of injected current at that moment is proportional to the length of the sidewalls of the metal that is directly connected to the gate and the stress level will drop accordingly.

### 4.3 Modeling charging induced yield loss

The development of a yield model for charging is not only necessary for predicting the charging yield loss of different products, but also for the development of layout rules for charging robust design. In most cases plasma process induced charging can not completely be removed from a manufacturing process and therefore a limit has to be set for the size and shape of the antennas that are still acceptable from a yield point of view. This section describes how the relationship between the distribution of antennas in a product and the corresponding yield loss can be established.

Although it is generally accepted that plasma processing can inflict damage on gate oxides and thereby cause yield loss [2,3,6,33], in practice it is very difficult to demonstrate which part of the total yield loss of a product is related to charging. There are several reasons for this. Firstly, the product level yield loss symptoms from charging induced defects are very difficult to distinguish from the symptoms of other defects such as intrinsic gate oxide defects or even metal shorts or opens. Also performance related yield loss caused by for example threshold voltage shift may be caused not only by charging, but just as well by (a combination of) many other effects. Furthermore, charging related yield loss may have any spatial pattern on the wafer. In addition, gate oxide defects generated by charging are hard to locate in a product die. Therefore, physical failure analysis is not able to prove that charging has been the reason for yield loss.

Another reason for the difficulty in proving the correlation between product yield and charging is the dissimilarity between scribe line test structures and products. For example, scribe lines are usually much more isolated and are therefore vulnerable to different kinds of charging than the product. Also, charging test structures show threshold voltage shift or gate leakage. Although threshold voltage shift can be interpreted as an indication of whether charging has occurred on the wafer, it may not be the failure mechanism that in the end causes the product to fail.

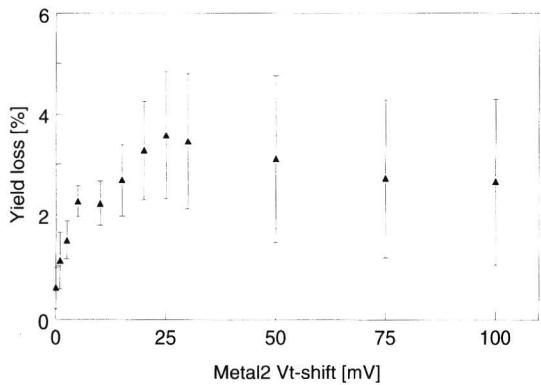
In some cases however, where charging is obvious and a clear signal is seen on scribe line test structures, a correlation with product yield can be made [6]. This was also the case for the experiment described in this section. In this experiment the impact of plasma process induced damage (charging) on the yield of products in 75-120 Å CMOS processes has been analyzed. It is shown that product yield loss is related to the threshold voltage shift of charging sensitive test structures and thus to charging. A yield model is introduced in which the charging related yield loss component of products is expressed by the attributes of antennas in products and the extent of charging measured on test structures. The proposed model is found to predict charging related yield loss using only one process dependent parameter, as is shown for various products.

### 4.3.1 Charging induced yield loss experiment

In this experiment thousands of wafers containing different products in a 120Å, 3 metal layer CMOS process have been analyzed. The wafers were selected from a particular period in time when charging was incurred by a specific set of tools during oxide deposition on the metal tracks.

In order to identify charged wafers, scribe line test structures were used. These antenna test structures consisted of a PMOS transistor with a metal ‘finger’ antenna with different antenna ratios, connected to the gate. (Taking the charging mechanism into account, as a first order approach, for this experiment the antenna ratio was defined as the ratio of metal top area of the antenna and the gate oxide area). Since the charging mechanism in this case appeared to be of uniform nature across the wafer, the wafer level average threshold voltage shift of antenna transistors with respect to reference transistors was taken as a measure of the extent of charging in the analysis. In the experiment the yield has been analyzed in relation with metal-2 charging. This metal layer was the most important charging layer in the processes under study, because in these designs poly-silicon and metal-1 are only used as a local interconnect and therefore large antennas in these layers are not likely to occur. Antennas in the third metal layer were not present at all, because at that level all gates are connected to diffusion areas in any three layer metal process.

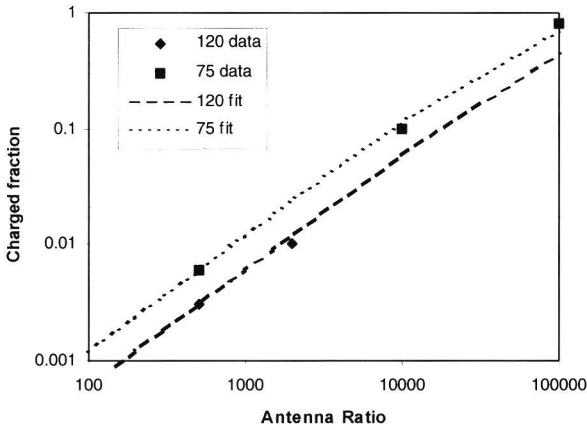
In order to separate the charged wafers from the non-charged wafers for product yield analysis, an optimal wafer separation criterion (in mV threshold voltage shift) was selected by plotting the average yield difference between charged and normal wafers as a function of the selection criterion. See figure 4.9. It seems that for this 120 Å process, the best criterion is a 25 mV shift.



**Figure 4.9** Average yield difference versus Vt shift criterion for charging damage in a 120 Å CMOS process.

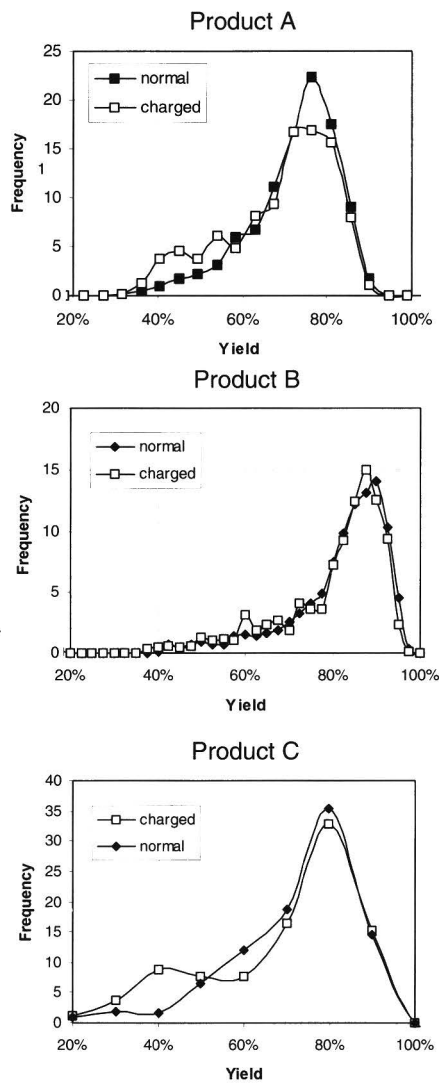
In order to model the contribution of damage of antennas within a design to the product yield, the fail probability of all transistors in the design was determined using the 25 mV threshold voltage shift criterion. Figure 4.10 shows the failed

fraction (defined as having threshold voltage shift  $>25$  mV) of test structures as a function of antenna ratio for the two different technologies. The measured values show a Poisson like behavior with respect to the antenna ratio.



**Figure 4.10** Charged ( $>25$  mV shift) fraction of antenna transistors vs. antenna ratio for 75 and 120 Å processes. The data has been fitted using a Poisson model

In figure 4.11 the yield distribution of 3 different products in the 120 Å process for the normal and ‘charged’ wafers is shown. The difference in average yield for the two classes of wafers was 3.5%, 0.9% and 1.4% respectively.



**Figure 4.11** Yield distributions of three different products for normal and charged wafers. Average yield differences are 3.5%, 0.9% and 1.4% respectively.

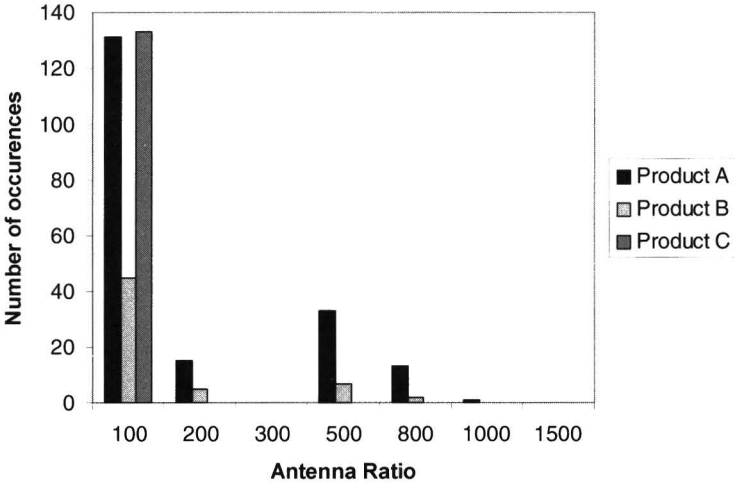
Average yield loss for Product A shows a clear lower average yield for the charged wafers. Product B and C do not exhibit the clear charging related yield distribution difference.

It was excluded that the apparent lower yield of wafers with a  $V_t$  shift on the metal-2 antenna transistor was due to another mechanism than charging. At metal-1 the same charging mechanism took place. Using the same separation

process based on metal-1 antenna Vt-shifts, normal and charged wafers were separated, but now no yield difference was observed.

### 4.3.2 Modeling the charging related yield loss

The differences between the three different products with respect to charging related yield loss can be understood from the results of antenna extraction from the designs of both products; product B has much less and smaller antennas than product A. Product C only contains small antennas. See figure 4.12.



**Figure 4.12** Antenna extraction results for products A,B and C; the number of antennas for different antenna ratios.

The charging related yield loss can be modeled using a Poisson like model that takes into account the antenna ratio distribution of the product in its fault density  $\lambda$ . If for an antenna with antenna ratio  $ar$  the probability of failing is  $POF(ar)$ , then the fault density of all antennas in a product  $X$  can be expressed as a function of the distribution of antenna ratios  $N(ar)_x$  in the product:

$$\lambda_X = \int_{AR=0}^{AR=\infty} N(ar)_x \cdot POF(ar) d(ar) \quad (4.1)$$

in which  $POF(ar)$  is measured from the in-line test structures (figure 4.10) and  $N(ar)_x$  is extracted from the product layout (figure 4.12). The charging related yield for product  $X$  can then be expressed as:



$$Y_{charging} = e^{-c\lambda_N} \quad (4.2)$$

in which  $c$  is a process dependent parameter that models the relationship between the charging related damage and the measured threshold voltage shift. The actual and predicted charging related yield loss is shown in table 4.1. The indicated measured yield loss is the difference in yield between charged and non-charged wafers, separated by the 25 mV criterion as discussed above. The calculated yield loss is the yield loss as predicted by equations 4.1 and 4.2. Good agreement is found, showing that the method used for calculation of the yield loss works quite well.

Product	A	B	C
Number of wafers	1484	527	2629
$c$	0.15	0.15	0.15
$\lambda$	0.264	0.064	0.080
Measured Yield loss	3.5%	0.9%	1.4%
Calculated Yield loss	3.9%	0.9%	1.2%

**Table 4.1** *Measured and predicted yield loss due to charging for different products.*

### 4.3.3 Conclusions and discussion

For this particular time frame of processing a clear relationship is found between product yield loss and plasma process induced damage as measured on antenna transistors. The product dependence in charging related yield loss can be understood using the distribution of antenna ratios present in the design of a product. The charging yield prediction methodology presented here successfully predicts the plasma process induced damage related yield loss, and can be used in practice to determine design rules with respect to charging robust design. It is also shown that for determining the charging-sensitivity of a product the conventional antenna ratio model that only takes into account the size or antenna ratio of individual antennas is inadequate. The total distribution of antenna sizes in the product needs to be evaluated. A large number of relatively small antennas can be more devastating than one large antenna.

In the experiment described in the above section, charging was caused by a high density plasma oxide deposition process step. Apparently, the layout dependence of this particular type of charging mechanism could successfully be modeled by only taking into account the top area of the conductors connected to transistors. However, as has been discussed in previous sections, in other cases the layout dependency can have a much more complex character. In order to develop charging yield models in such cases, the manufacturing process needs to be extensively characterized in that respect so that the antenna extraction algorithms can be developed accordingly. The next section extensively describes plasma process characterization methods.



## 4.4.2 Multiplexed Antenna Monitoring (MAM) test structures

The most effective way of gathering data in a systematic manner can be achieved by using electrical test structures. It has been shown that measuring basic transistor parameters such as gate leakage and threshold voltage shift can be used effectively to determine the extent of charging that has occurred on a wafer [4]. The problem is however, that charging may have only a limited effect on transistor parameters, especially, if antennas with small ratios are investigated. In such cases, large sample sizes are needed to reliably distinguish transistor parameter shifts due to charging from deviations due to normal process variations.

Furthermore, the extent of charging damage randomly varies within one lot. Some wafers may show a strong effect while others are not charged at all. Thus, it is necessary to extensively characterize charging effects for each individual wafer and avoid averaging the data from all wafers in a lot.

Therefore, a good electrical test structure should be able to provide sufficient data for a variety of charging structures on one wafer. Such a test structure should also contain a large variety of antennas. Every antenna should show sensitivity to charging to a specific layer and geometry so that it is possible to relate charging effects to any specific processing step or tool.

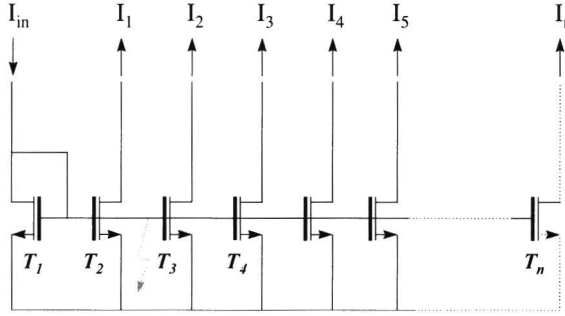
The problem with the above objective is that conventionally, single transistor structures need an extra set of probing pads for every kind of antenna. Consequently, for a large number of different antennas the silicon area overhead consumed by probe pads becomes unacceptable. A large number of pads also has a negative effect on tester time since the wafer prober must reposition the probe for each transistor to be measured. A large portion of total tester time is consumed by re-probing.

To address the above problems in this sections the multiplexed antenna monitoring (MAM, [8,19]) test structure is proposed which:

- enables measurement of both threshold voltage shift and gate leakage of each transistor in the structure;
- requires small area overhead for probing pads;
- facilitates efficient gathering of large numbers of data;

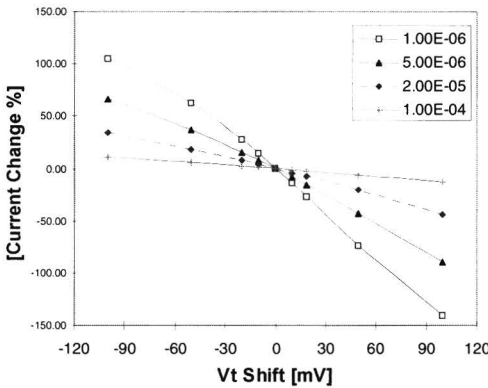
### *Principle of operation*

The MAM test structure essentially is a multiple current mirror consisting of a large number ( $n$ ) of identical transistors as shown in figure 4.14.



**Figure 4.14** MAM current mirror principle.

During testing, the tester feeds a bias current through  $T_1$  and measures the drain currents of remaining transistors ( $T_2$  through  $T_n$ ). In an ideal situation  $I_1$  through  $I_n$  should be identical. In the MAM test structure every odd numbered transistor is connected to an antenna (antenna transistor). Hence every antenna transistor has an identical reference transistor without antenna in its immediate neighborhood. In this arrangement the difference in drain currents between an antenna transistor and its reference transistor will be an indication of the difference in threshold voltages of these two transistors. This difference, in turn, should be seen as being caused by antenna charging, assuming that normal transistor mismatch (due to local process variations) is negligible.

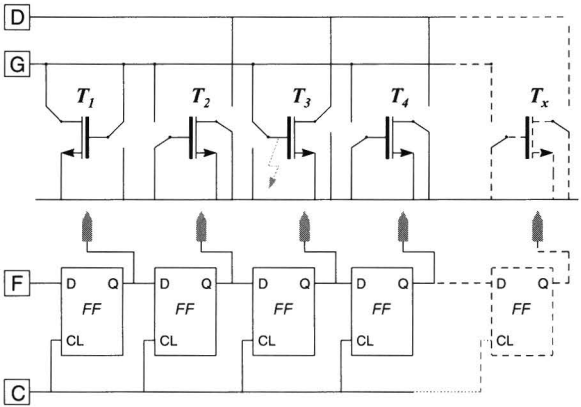


**Figure 4.15**  $\Delta I$  as a function of threshold voltage shift with bias current as a parameter

Figure 4.15 shows the output current differences in terms of threshold voltage shift for different bias currents as obtained via simulation. Note that the sensitivity of the circuit increases with decreasing bias current.

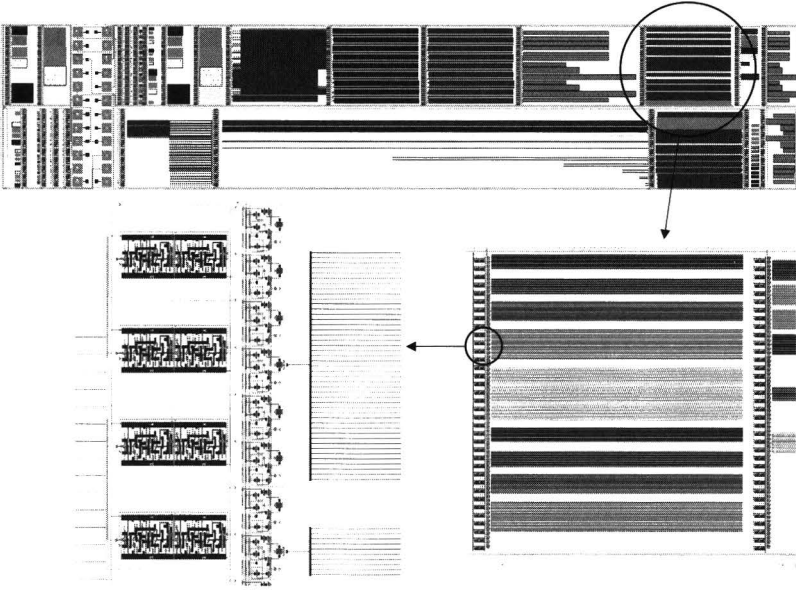
*Test structure design*

Direct implementation of a test structure as shown in figure 4.14 still would require a probe pad for each output current to be measured. Such a solution would require undesirable large area overhead. Therefore, in the MAM test structure, a simple pad-multiplexing strategy was used, as is explained in figure 4.15. To enable selection of each transistor separately, a number of pass transistors controlled by values stored in a shift register was applied.



**Figure 4.15** A shift register is used to drive the switches

The MAM test structure discussed in this paper has 1024 NMOS transistors. All antenna transistors have a  $0.35 \times 0.5 \mu\text{m}$  gate area. The structure has 95 different kinds of antennas with different areas, in different layers, of different shapes. To distinguish between area and perimeter sensitive charging, fork-shaped and plate-shaped antennas of different sizes are included. Figure 4.16 shows the layout of the test structure. Although only 6 pads are necessary to operate the test structure, a standardized  $2 \times 12$  pad layout is chosen in order to be compatible with the probe card of other test structures.



**Figure 4.16** Layout of a total MAM test structure; 1 section showing the antennas; Detail of a section showing the flip-flops and the connected target transistors

#### *Threshold voltage shift measurement in the MAM test structure*

By clocking in the appropriate data into the shift register the gate of the selected transistor ( $T_3$  in figure 4.15) is switched to the gate of the main reference transistor  $T_1$ . The input current is fed through  $T_1$  by the tester using pad [G], resulting in a gate voltage on the selected transistor. The drain of the selected transistor is connected the pad [D], so that the output current can be measured. All other transistors are switched off by connecting their terminals to  $V_{ss}$ . During the subsequent clock cycle the adjacent transistor (located next to  $T_3$ ) is selected and its output current is measured. Finally, the difference in output currents between adjacent transistors  $T_3$  and  $T_4$  is translated into threshold voltage shift.

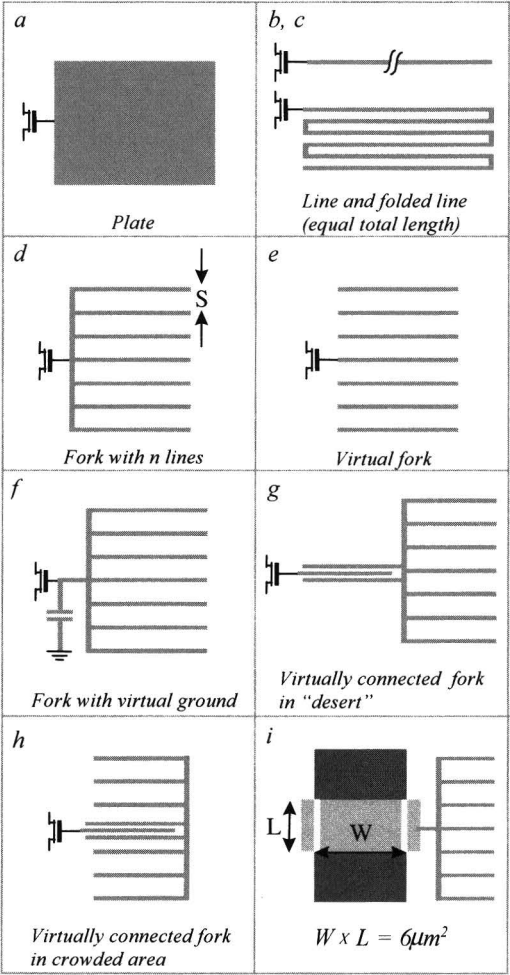
#### *Gate leakage measurement in the MAM test structure*

Gate leakage of each transistor is measured by selecting the transistor via the shift register. A voltage is put onto pad [G] (gate of the transistor) and its source and drain are connected to  $V_{ss}$ . Gate leakage current is measured through the same pad.

#### *Test Structure Detail*

The MAM test structure as described in the previous section was used. Most antennas were drawn in metal layers. Fig. 4.17 shows the key antenna configurations that were used in the experiment. Parameters such as area,

perimeter, spacing and layer were varied. Structures a, b and c (“plate”, “line” and “folded line”) were used to assess relevance of the classical antenna ratio concept. The effect of line spacing has been investigated with antenna structures d, e, and f. The impact of the neighboring metal density on plasma damage has been studied with antenna structures g and h. The effect of transistor geometry has been studied by varying  $L$  and  $W$  of the target transistors while keeping the antenna area constant (structures i).



**Figure 4.17** Antenna configurations used in the experiment

#### 4.4.3 Characterization results of the layout dependence of charging for 0.35 and 0.18 $\mu\text{m}$ processes

This section describes the characterization of the layout dependence of two different processes; a 0.35 $\mu\text{m}$ , 5 metal layer process and a 0.18 $\mu\text{m}$ , 6 layer metal process. The 0.35 $\mu\text{m}$  process was characterized using a MAM test structure. The 0.18 $\mu\text{m}$  process was characterized using conventional structures as described in section 4.4.1.

##### *Experimental results for a 0.35 $\mu\text{m}$ process*

The experiment was conducted to study charging in a standard 0.35  $\mu\text{m}$ , 75  $\square$  gate oxide, 5 metal layer, CMP based technology. The interconnect formation steps as listed in Table 4.1 were the focus of the investigation. Steps flagged with "Y" do pose a possible charging problem.

Process Step	Processing technique	Charging
Via <sub>x</sub> formation		
Metal Deposition	PVD	N
Metal definition	Lithography	N
Metal etch	Plasma etch	Y
Resist strip	Plasma etch + wet etch	Y
First oxide deposition	CVD	N
Second oxide deposition	Plasma	Y
Oxide removal	CMP	N
Via <sub>x+1</sub> formation		

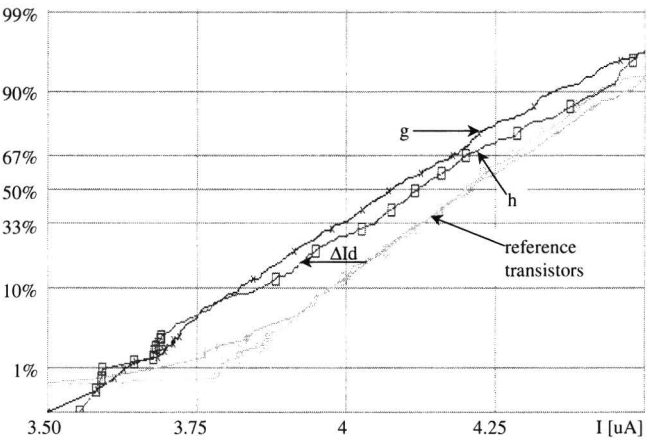
**Table 4.1** Simplified interconnect formation flow

In this experiment two lots of 12 wafers were processed. On each wafer 60 MAM devices were tested. This resulted in  $\sim 1400$  measurement points per antenna configuration. The devices were tested on an industrial product tester. For each antenna-connected transistor the current shift was measured with respect to the reference (antenna free) transistors next to it. To assess the charging impact of the different antennas, the corresponding distributions of drain currents ( $I_d$ ) were compared.

First, to assess the relevance of the classical antenna ratio concept, the cumulative distribution of  $I_d$  currents for the transistors connected to the line antenna and folded line antennas (Fig.4.17 b,c) were compared. Fig.4.18 shows clearly that the output current of the transistors connected to the antennas has shifted. Hence,



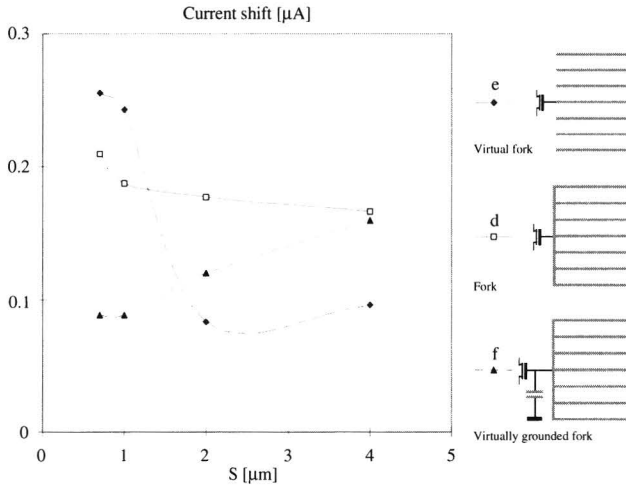
despite having the same antenna ratio, the two antennas produce substantially different levels of gate oxide degradation.



**Figure 4.18** Cumulative current distributions for antennas g and h have shifted with respect to the reference transistors

Figure 4.18 shows the cumulative distribution of  $I_d$  currents for the virtually connected fork (antenna g) and the virtually connected fork in a crowded area (antenna h).

Again, according to the conventional antenna ratio concept both antennas should have little or no impact since the structure that is connected directly to the gate is very small. Figure 4.18 shows that both antennas produce a considerable shift in the  $I_d$  current distribution with respect to the reference transistors. Also there is a difference between g and h, even though the area of both forks are the same. This can be explained in the following manner: the spacing between the center line that is connected to the gate and the fork is minimal. Therefore, the connection of the fork and the transistor will stay intact during almost the total etching time. This way the fork is effectively connected (and is therefore called a virtually connected fork). The difference between g and h can be explained by the microloading effect (section 4.2.2). In the case of the crowded area (h) the etching will be slower and the fork will stay connected even longer and will therefore be able to collect more charge during the etch. Hence, the pattern density of its neighborhood affects the critical “virtual connection” spacing.



**Figure 4.19** Median current shifts for antennas d,e and f as a function of the spacing between lines

To determine the critical distance between patterned shapes at which lines stay virtually connected, antennas d,e, and f were studied. The median of the shift of the current distributions for those antennas is shown in Fig.4.19 as a function of the spacing  $S$  between the lines in the fork. Again the charging effects on the different antenna shapes are very dissimilar. (Note that d and f have the same antenna ratio in the conventional sense). From these results it can be concluded that in this process, charging not only occurs during the metal etch process (electron shading in combination with microloading), but also during oxide deposition (extended electron shading). Both charging mechanisms show a dependence on the spacing between tracks:

The charging effect for virtual-fork antennas of type e decreases with the spacing between the tracks. The closely spaced tracks stay connected during the main etching process, resulting in an effectively large antenna that is charged due to electron shading. At larger spacing the tracks do not stay connected so long during the main etch, and the charging effect drastically decreases at spacings larger than  $1.3 \mu\text{m}$ . Hence, the maximum of the virtual connection distance in this experiment was determined to be a random number varying between  $1.2$  and  $1.7 \mu\text{m}$ .

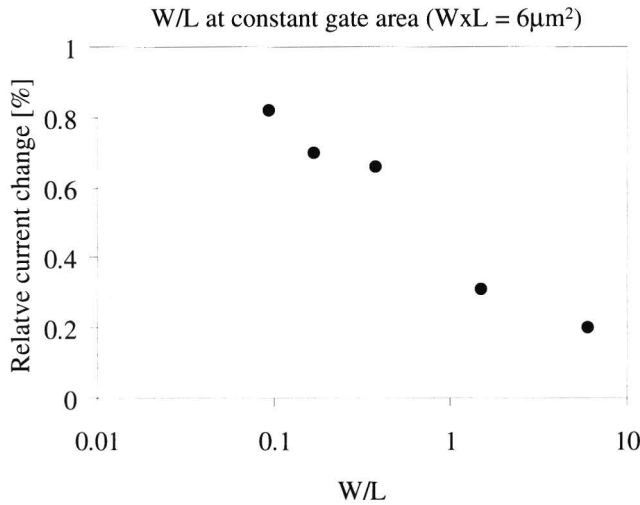
Antennas of type f are virtual grounded during the main etch by the fuse. Therefore the charging due to etching at small spacings is not present. However, the virtually grounded antenna shows increased damage at larger spacings which means that in this process damage occurs due to the extended electron shading effect either during over-etch, resist strip or oxide deposition (see table 4.1).

Antennas of type d shows charging of both types. At small spacings the electron shading effect during etch plays a role. At larger spacings the charging during the

main etch decreases and the charging due to the extended electron shading increases.

It can be concluded from these results that both the electron shading and extended electron shading play a role in this process. Not the area of the conductor connected to that gate itself, but the spacing to neighboring tracks is determining for a large part the extent of charging damage.

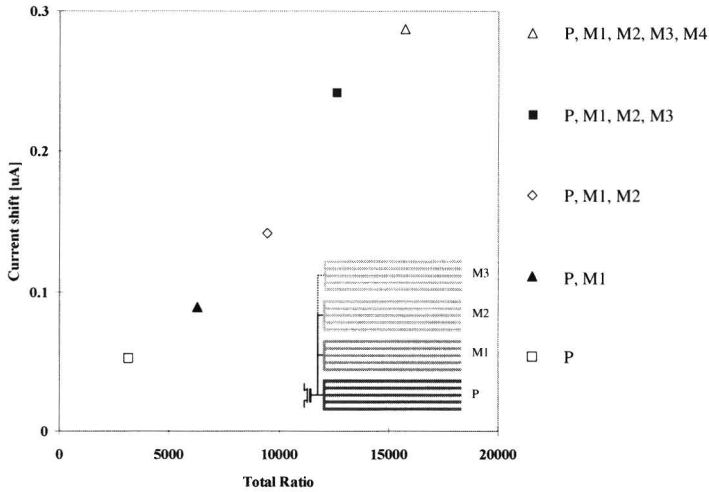
In figure 4.20 the relationship between  $I_d$  current shift and the transistor geometry is shown. In this study, fork antennas were used and again the antenna ratio (both transistor area and antenna area) in the conventional sense was kept constant, but the  $W/L$  ratios were varied.



**Figure 4.20**  $I_d$  change at constant antenna ratio and gate area

From the data it can be concluded again that the area of the gate oxide alone is not enough to model plasma induced damage as is done in the conventional model. Apparently the damage occurs in the neighborhood of the field oxide edge. This can be explained by the thinning of the gate oxide that occurs near the bird's beak of transistors.

Finally, the cumulative nature of the conventional model was investigated. This was done again with fork antennas. One transistor was connected to a relatively small antenna designed in the poly layer. Another transistor was connected to an equal antenna in poly, and an identical antenna in metal 1. A third transistor was connected to identical antennas in poly, metal1 and metal 2. This strategy was repeated up to metal 4. The average  $I_d$  current shifts for these antennas are shown in figure 4.21.



**Figure 4.21** Cumulative nature of damage of different layers

The results show that indeed the damage is accumulated from layer to layer. Also it can be seen that in this experiment charging was the most severe in metal 3.

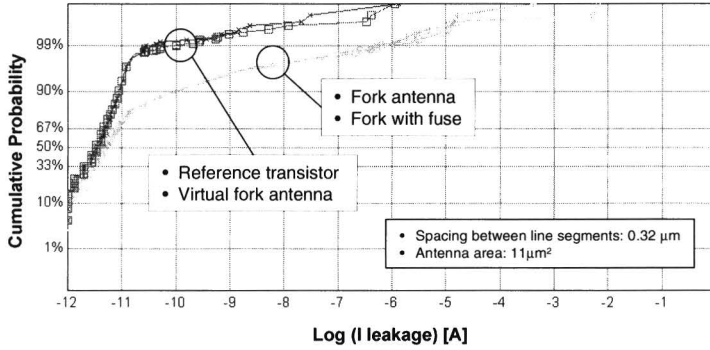
*Results of an investigation of the electron shading effects in a 0.18  $\mu\text{m}$  process.*

A metal-2 charging damage characterization was also done for a 0.18  $\mu\text{m}$  process. The metal-2 formation for this process contains several plasma steps: metal 2 etch, liner deposition and plasma TEOS deposition. Eight different lots were considered for this experiment. 9 measurements of threshold voltage and gate leakage for different antennas were done on each wafer adding up to around 1350 data points per antenna configuration. All wafers went through the complete processing flow, including passivation steps.

Conventional test structures as discussed in section 4.4.1 were used. For this experiment only fork antennas, virtual fork antennas and fork antennas with fuse (see figure 4.17 d, e and f respectively).

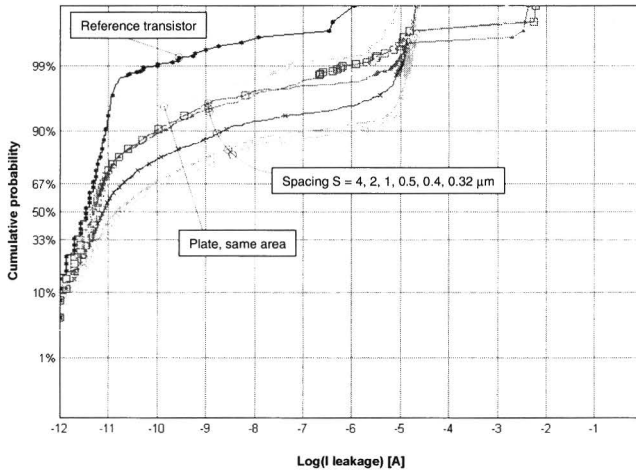
No threshold voltage shift was observed for all antennas. However, gate leakage did occur as is shown in figure 4.22a which shows the cumulative probability plot for the gate leakage measurements for the different antennas. From these results it can be concluded that for this process there is no electron shading present since the gate leakage distribution for the virtual antenna and the reference transistor are equal. The fork antenna and the fork with fuse however show a big charging effect.

In contrast to the results from the 0.35  $\mu\text{m}$  process, the fuse has no effect and therefore the damage occurs either during over-etching (when the fuse is already disconnected), or during the liner or oxide deposition.



**Figure 4.22a** Gate leakage distribution for transistors with different antennas.

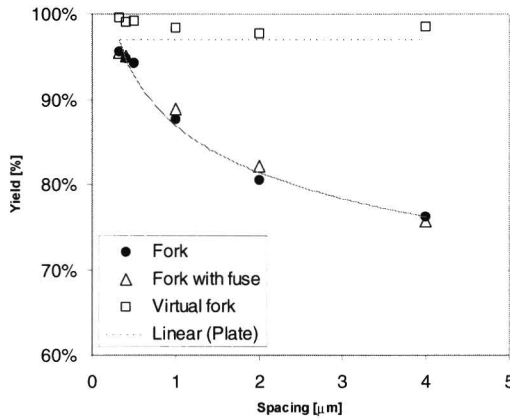
Figure 4.22b shows the cumulative leakage current distribution for fork antennas with different spacings. Also the results for a plate shaped antenna with equal area and the reference transistor are shown.



**Figure 4.22b** Cumulative leakage current distribution for fork antennas with different spacings.

Clearly there is an increase in charging damage on the fork shaped antenna for increasing spacing between the fingers. For this process, tracks that are isolated are clearly more prone to charge build-up closely spaced lines. This can be attributed to an extended electron shading effect during the over-etch or deposition of the liner.

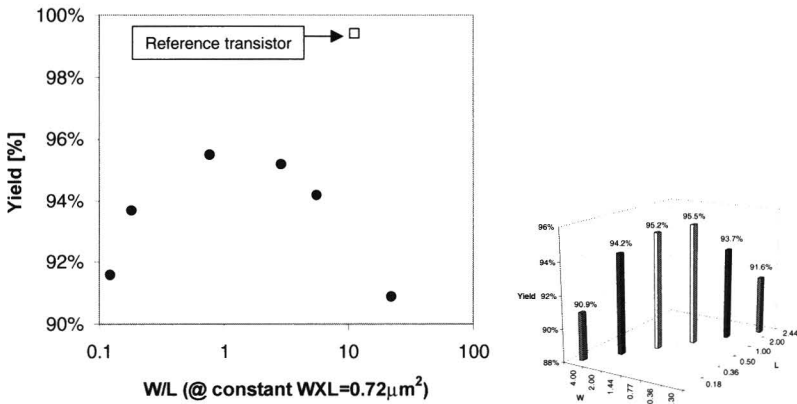
Figure 4.23 shows the yield of the four different antenna configurations as a function of spacing. Transistors with a gate leakage current of more than 1nA were considered to be damaged.



**Figure 4.23** Yield of the different antenna configurations as a function of spacing between the fingers. (leakage current above 1nA is considered defective)

There seems to be a saturation of the damage for larger spacings.

Figure 4.24 shows the yield of different transistors that are connected to equally large antennas. Transistor W and L are varied while keeping gate oxide area stays constant. Note that for the 0.35  $\mu\text{m}$  process a similar measurement was done with different results. Again the assumption that the transistor sensitivity is solely dependent on its gate oxide area is shown to be wrong for this process.



**Figure 4.24** Yield variation for different transistors with equal antenna areas. W and L are varied while gate oxide area (=antenna ratio) is kept constant.

#### 4.4.4 Conclusions with respect to the characterization of the layout dependence of charging damage

A first general conclusion of the experiments described above is that the conventional antenna ratio concept is inadequate to describe the extent of plasma induced damage in real VLSI devices. Also it is shown that different processes can have different charging mechanisms. For example, the  $0.35\mu\text{m}$  process shows electron shading effect while the  $0.18\mu\text{m}$  process is seriously hampered by the extended electron shading effect.

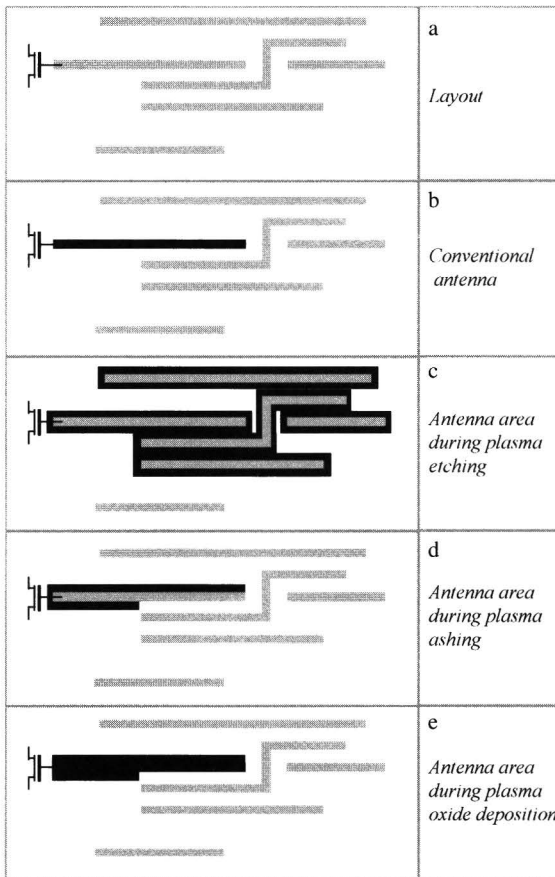
The results show clearly that not only gate oxide area, but also the length of the transistor can determine the sensitivity to charging damage. Also it has been shown that the metal density in the neighborhood of the antenna plays a role in determining the extent of plasma damage.

Another important observation is that the antenna area during etching changes in time. At the beginning of the etching process the wafer is totally covered with metal and at that time many leakage paths to the substrate exist, so there are no antennas. When the etching continues, unnecessary metal is etched away and “islands of metal” will be formed. During continuation of the etch there may arise a situation where a transistor gate is connected to an island that has no connection to substrate anymore. In that case the resulting antenna may become very large. As the etching proceeds, more metal will be removed and the antenna becomes smaller. Finally in the over-etching phase, all metal polygons are separated from each other and at this stage the conventional antenna ratio model may be applied.

Another observation is that for different plasma processing steps there are different antenna sizes. Consider, for instance, the example shown in figure 4.25. In figure 4.25a a segment of a layout is drawn. Figure 4.25b shows what a conventional view on charging would indicate as being a possible antenna (highlighted black). During etching however, islands of connected metal appear that may form substantial antennas as is shown in figure 4.25c. Figure 4.25d shows the part of the layout that makes the connected transistor sensitive to charging during plasma resist removal. As during the etch, only the edges play a role. Figure 4.25e shows the part of the layout that contributes to the damage during plasma deposition of oxide. Since charging damage is cumulative, the total damage done is the sum of figure 4.25c,d and e.

A general conclusion is therefore that an antenna prevention strategy must be based on a new concept of plasma induced damage measure. Such a measure must take into account the cumulative nature of charging induced damage. In addition it should take into account that different processing steps will inflict gate oxide damage via charging of very different charging sensitive areas. This leads to another conclusion that an antenna prevention strategy should involve the following elements: first the processing steps that may contribute to plasma damage should be identified. A MAM-like test structure should be designed, manufactured and tested to reveal individual contributions from each processing step and their layout dependency. From the results of the test structure can be determined what are acceptable levels of charging with respect to reliability, yield and performance. A corresponding antenna ratio model computed as a weighted

sum of the charging sensitive areas for each plasma based manufacturing step can then be developed.



**Figure 4.25** Charging sensitive areas for different plasma processing steps.



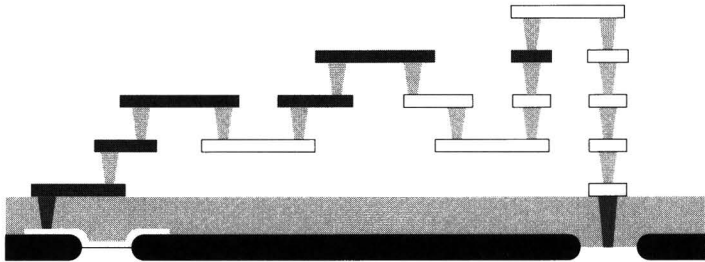
## 4.5 Robust design for charging

Due to the complexity of the charging phenomena discussed in the previous sections, today it is still not possible to completely eliminate the plasma damage from any manufacturing process solely by optimizing process parameters. Moreover, tool and process fluctuations in combination with the difficulty of monitoring the charging relevant parameters make it difficult to systematically guarantee low damage levels. Hence, products have to be designed in such a way that they are as insensitive to plasma damage as possible.

In this section current methods that are available for charging-robust design will be described. It will be shown that these methods are inadequate and that there is a need for different methods that take into account a proper definition of product sensitivity to charging. A new methodology for assessing product sensitivity to charging is proposed.

### 4.5.1 The conventional approach

Since charging damage has a cumulative nature the charging impact for an individual transistor needs to be evaluated by taking into account the whole net that is connected to the gate [21]. Figure 4.26 shows an example of a net that is evaluated for charging damage. The shaded parts of metal add to the damage. The un-shaded parts do not add because they are not connected to the gate yet when they are manufactured.



---

**Figure 4.26** *Metal sections of a layout that have to be taken into account for antenna extraction (shaded)*

---

Once the gates in the product design that exceed a certain antenna ratio are identified, a designer can choose between the insertion of diodes or bridges in order to make a circuit robust to charging. Both methods are described below.

*Diode insertion*

Assume that in the example shown in figure 4.27 a, the metal 3 wire is long and will collect enough charge to damage the gate of the transistor. A reverse biased diode that is connected to the gate of a transistor forms a leakage path for the charge that is built up on the antenna during processing. (Figure 4.27 b). The plasma conditions (high light intensity and temperature) ensure low impedance of the diode during processing so that charge can easily flow to the substrate. During normal operation the diode doesn't affect the functionality of the circuit.

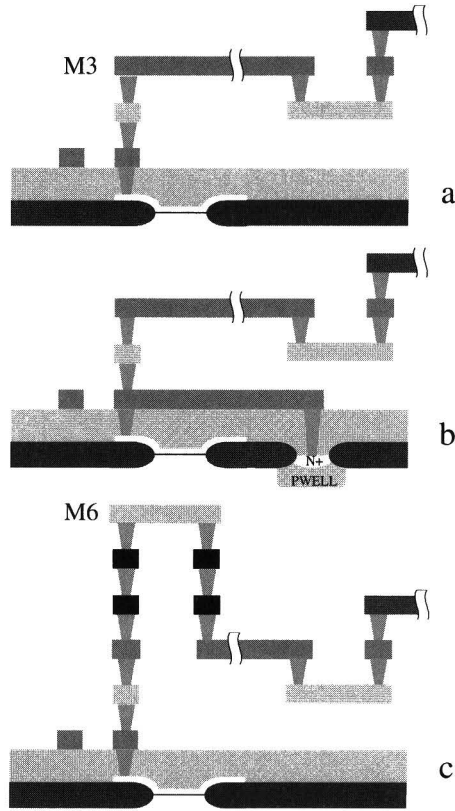
Diodes seem to be a simple solution to the charging problem. Therefore standard cell libraries are sometimes equipped with a diode at every input. Assessment of the charging sensitivity or compliance to charging design rules does not need to be assessed during the design of a product since all gates are protected. For designers this eliminates a lengthy and tedious procedure of detecting and repairing antennas. However, diodes also have some disadvantages with respect to wafer productivity and device performance. Converting a non protected cell library to a library where all inputs are diode protected may have significant implications for the footprints of the cells.

For instance, in the author's case a  $0.35\mu\text{m}$  standard cell library of 350 different cells was converted. Depending on the layout, diode insertion resulted in an area increase of 0 to 10%. The average area increase was 7%. The extent to which diode insertion has an effect on the die area depends on whether the die area is dominated by the metal routing, the cells or the number of bondpads.

In addition, a diode connected to a gate leads to extra input capacitance. Such a capacitance connected to each input degrades the performance of the chip in terms of speed and power consumption. For high performance designers therefore consider diodes not to be a reasonable solution to the plasma charging problem.

*Bridging*

Another design solution to the charging problem is preventing of the charge to build up on the gate oxide by making sure that the gate is not connected to any structure that collects charge during plasma processing. This can be achieved by constructing the routing in such a way that all metal that is being etched at a certain time is not connected to a gate yet. By adding the bridge as is shown in figure 4.27 c, the long metal 3 part of the wire is no longer connected to the gate during metal 3 processing. The accumulated charge can not affect the gate. In order to protect the gate for only metal 3 charging, the bridge does not need to go all the way to metal 6, but can be limited to metal 4. It then does not provide protection for charging on metal 4 and 5 level. For protection on all metal layers, a complete stack of vias to the upper metal layer must be implemented as is shown in figure 4.27 c. The upper metal always completes the connection from gate to the output (drain area) of another transistor. The drain area will provide a leakage path for the accumulated charge.



**Figure 4.27** *Diodes (b) and bridges (c) as charging protection*

Insertion of stacks of vias to each input gate in a cell library also has severe disadvantages. Many stacks of vias lead to many routing obstructions. A router therefore needs more space to be able to make all connections and the result is a larger chip.

In the occasional case where area and speed of the device do not play a role, full insertion of diodes or via stacks may be a fast and easy solution. However, for high performance devices, complete insertion is not acceptable and a more sophisticated solution is needed where diodes or bridges are applied only where it is really necessary. In order to implement such algorithms in routing or post layout processing tools, a definition of what layout configurations are acceptable in terms of charging damage are needed.

Whether or not an accumulated charge inflicts any damage on a device depends on the area and shape of the gate oxide itself, the plasma conditions and the area and shape of the structure that collects the charge (the antenna). To assess the

extent of plasma induced gate oxide damage, often the simple concept of antenna ratio is used. It is defined as the ratio of the cumulative area of poly and metal structures, that is connected to a gate area, but not to a the substrate area or another leakage path, e.g. via a drain area. Hence the antenna ratio AR is:

$$AR = \frac{A_{Poly} + A_{m1} + A_{m2} + A_{m3} + \dots + A_{m(n-1)}}{L \cdot W} \quad (4.3)$$

where:  $A_i$  is the charging sensitive area (top area or the total perimeter of the structure),  $L$  and  $W$  are the length and width of the affected transistor respectively, and  $n$  is the total number of metal layers. The top layer metal  $A_{m(n)}$  is not included in the calculation of the antenna ratio since when this metal is formed, all connections are complete and there always exists a connection to a source or drain area which can sink the accumulated charge safely.

A serious drawback of the use of a charging model described by (4.3) is that it does not take into account the layout dependency of charging that was described in previous sections. It is known that the shape and surroundings of the antennas contributes to the extent of charging which is ignored by this traditional antenna definition. Therefore the algorithms based on this model that are used to remove or prevent antennas are not based on an adequate definition of what layout configurations make a design susceptible to charging damage. Therefore a designer runs the risk of either overestimating charging damage on certain transistors (resulting in unnecessary repair work), or in underestimating charging damage (resulting in un-anticipated yield or reliability loss).

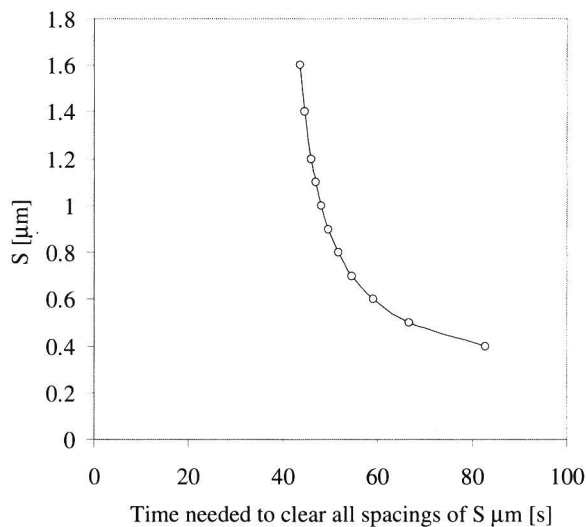
A second drawback of the traditional antenna ratio model is that it only takes into account antenna ratios of individual transistors and a design rules is set so that the maximum allowed antenna ratio may not be exceeded. However, as is shown in section 4.3.2, not the antenna ratio for individual transistors, but the distribution of antenna ratios in the product needs to be evaluated in order to assess its yield loss due to charging. A large number of moderate antennas may cause a product to be more vulnerable than one large antenna.

In order to be able to design products that are less vulnerable to plasma damage, the manufacturing process needs to be evaluated in terms of the layout dependency of charging. Based on the results, a process dependent model of all involved charging mechanisms needs to be built based on which extraction algorithms are developed that are able to detect charging sensitive transistors based on the understanding of the process. In the following section an example for the development of a model for assessing the charging sensitivity for charging during metal etch is given.

### 4.5.2 Sensitivity index model for plasma damage during metal etch

From the considerations in the previous section, it must be concluded that the conventional definition of antenna ratio is inadequate. The (MAM) test structure results show clearly that not only gate oxide area, but also the length of the transistor and the metal density in the neighborhood of the antenna determine the sensitivity to charging damage. By adopting the traditional antenna ratio concept a designer therefore runs the risk of correcting antenna structures that are not really antennas, or worse, structures that will accumulate large amounts of charge are not recognized as being antennas. To solve this problem, in this section a new way of looking at the antenna ratio concept is proposed.

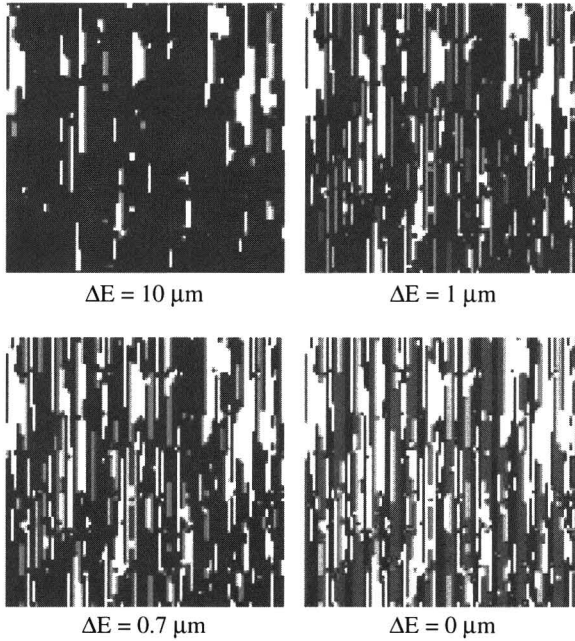
As discussed in section 4.2.2, RIE-lag causes the etching rate to scale with the aspect ratio of metal and resist structures. When the etching rate as a function of aspect ratio is known [e.g. 20], it is straightforward to calculate the etching time that is needed to clear the metal in between metal tracks. See figure 4.28.



**Figure 4.28** Spacing of metal lines as a function of the time needed for clearance based on data from [ 20 ]

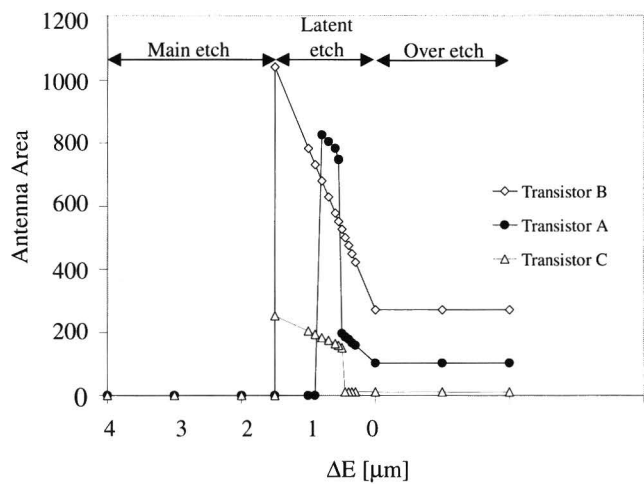
Due to such plasma process characteristic the blanket of deposited metal on the wafer will gradually transform into disconnected "islands" of metal while etching. During this process the areas, shapes and contours of these islands change until they are completely disconnected and form the interconnect pattern that was defined by lithography. To asses the impact of this phenomenon on the antennas that are formed in a product during the etching process, the antenna area can be extracted from a circuit layout as a function of the metal spacings that has been

cleared ( $\Delta E$ ). As an example this has been done for a real product which is produced in a 75 $\square$  gate oxide, 3 metal layer, 0.35 $\mu\text{m}$  CMOS process. The maximum allowed antenna ratio (in the conventional sense) for this product was 300. Extraction of antennas using the conventional antenna ratio model shows that there was no antenna exceeding this limit. Figure 4.29 shows the extracted metal areas for a small part of the product layout, taking into account latent antenna effects at the different stages of etching regimes.  $\Delta E$  is the metal spacing that is cleared. Figure 4.29 shows the resulting extracted antenna areas for a set of transistors in the product at the different stages of etching regimes.

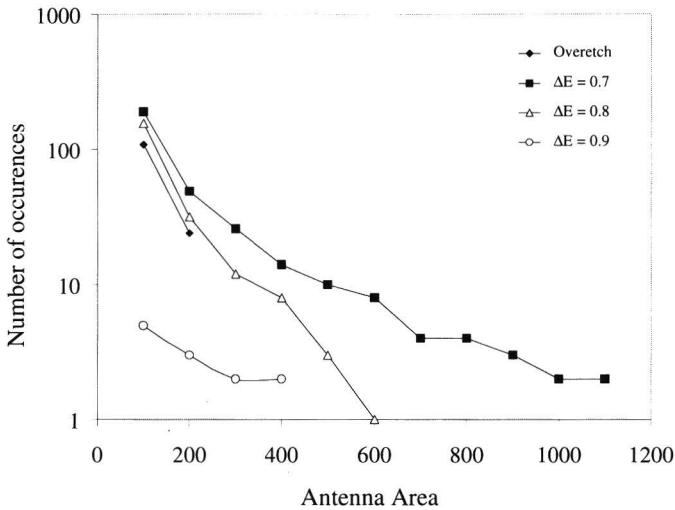


**Figure 4.29** *Simulation of a portion of a product layout during the different phases of metal etching*

Note that transistor A stays connected to a drain area until all open areas larger than 0.9  $\mu\text{m}$  have been cleared. All drain areas are disconnected from the metal island and the resulting antenna area is around 800  $\mu\text{m}^2$ . As etching continues, the side-walls of the connected metal become steeper and the area reduces to 750. Then a large part of the antenna is disconnected and the antenna area falls below 300. For transistors B and C a similar pattern can be observed. In figure 4.31 the number of antennas having an antenna ratio larger than a certain value is given for the different etching stages. As one can see, although in this product the antenna ratio in the conventional sense is not exceeded, the extractions show that there are some transistors that are connected to large antennas areas during a large part of the etching process.



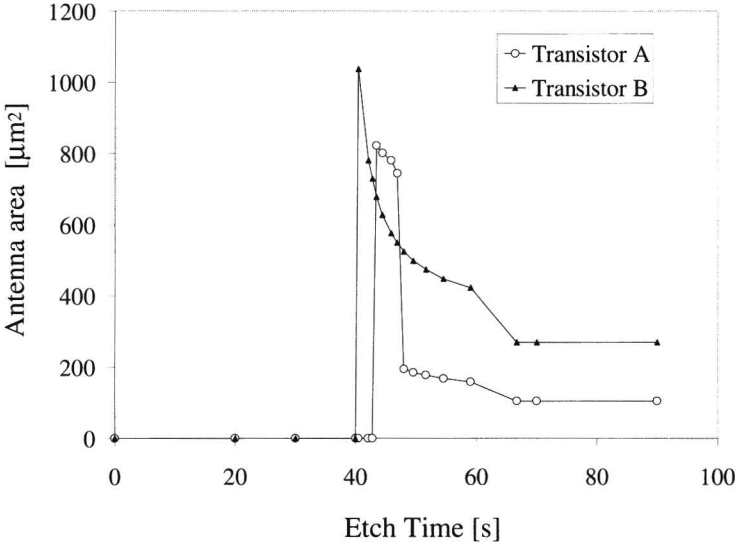
**Figure 4.30** Antenna areas connected to transistors in a product during the etching process.  $\Delta E$  is the spacing cleared. When the metal between all spacings of  $1.5\ \mu m$  and larger is cleared, transistors B and C get disconnected from any substrate contact and large antennas are formed



**Figure 4.31** Distribution of antennas larger than a certain value

The amount of injected charge in those transistors may therefore be much larger than can be anticipated from the conventional antenna ratio model.

Figure 4.30 shows that it is possible to extract from the layout of a circuit how large the antennas are that are connected to the transistors during different stages of the etching process. If the etching process can be characterized in terms of RIE-lag as is shown in figure 4.28, then it is possible to determine the antenna area as a function of time for the different transistors in a device. As an example this was done for transistors A and B. See figure 4.32.



**Figure 4.32** Antenna areas for two transistors in the product as a function of etching time

If now one can assume that the injection current is proportional to the antenna area, then the total accumulated injected charge into the gate oxide of a transistor is

$$Q = \int_{T_0}^{T_{end}} I_{inj} dt = K \int_{T_0}^{T_{end}} A(t) dt \quad (2)$$

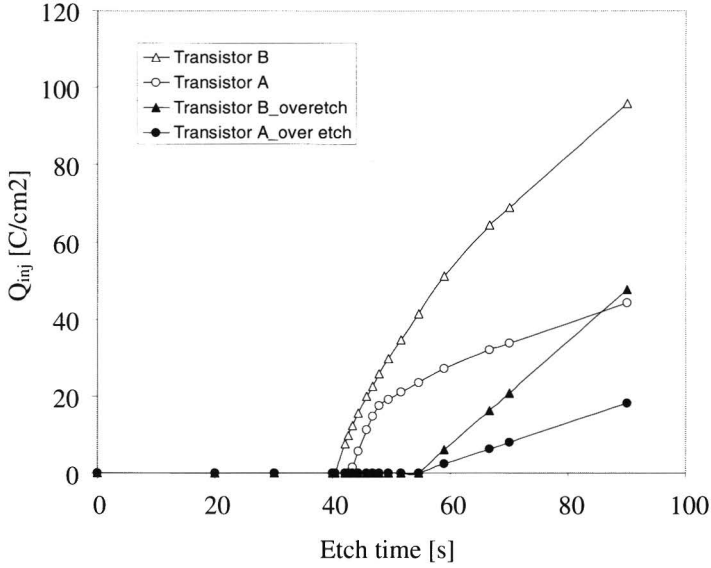
where  $T_{end}$  is the total etching time,  $K$  is a process dependent parameter determined by test structures, and  $A(t)$  is the area of the antenna as a function of time (figure 4.32).

An example of the injected charge for both transistors is shown in figure 4.33. (In this case an arbitrary value for  $K$  is taken). The figure also shows the injected charge for the same transistors, for the case that only over etch is taken into account (as in the conventional antenna ratio model). These calculations show



that the contribution of the latent etching regime to the charging damage to certain transistors in a real product can be significant.

Evaluation of the design with the conventional antenna ratio model may therefore lead to under-estimation of the damage, resulting in the erroneous conclusion that protection is not needed. Hence from the above example as well as from other considerations presented here, it is evident that in order to assess the charging sensitivity of a product, the antenna area as a function of time has to be evaluated for each transistor.



**Figure 4.33** *Injected charge as a function of etching time*

Therefore, a new charging vulnerability index (CVI) for each individual transistor can be defined as:

$$CVI = \frac{Q}{\alpha A_{gate}} \quad (3)$$

where  $Q$  is described by (2),  $A_{gate}$  is the gate area and  $\alpha$  is a parameter describing the transistor geometry. The CVI should be used instead of the conventional antenna ratio (1) in detecting charging sensitive conditions for individual transistors.

### 4.5.3 Quantifying product sensitivity to charging

In the previous section it was shown how product sensitivity for metal etching including RIE-lag can be evaluated. In a manufacturing process there are however more plasma processing steps that may introduce damage such as liner deposition, oxide deposition and resist strip. Therefore it is necessary to develop similar models for the complete manufacturing process.

This leads to the conclusion that a charging robust design strategy should involve the following elements: first the processing steps that may contribute to plasma damage should be identified. A MAM-like test structure should be designed, manufactured and tested to reveal individual contributions from each processing step and their layout dependency. From the results of the test structure it can be determined what types of charging play a role in the manufacturing process and what are acceptable levels of charging with respect to reliability, yield and performance. (See for example section 4.4.3 and 4.3.2.)

Based on these results a corresponding charging vulnerability index (CVI), needs then to be computed as a weighted sum of the charging sensitive areas for each plasma based manufacturing step and for each individual transistor. If  $POF(CVI)$  is the probability of failure of transistors as a function of the CVI, and  $N_x(CVI)$  is the extracted distribution of transistors in product  $x$ , with a charging sensitivity CVI, then in analogy to (4.1), the fault density on the product due to charging is

$$\lambda = \int_{CVI=0}^{CVI=\infty} N_x(CVI) \cdot POF(CVI) dCVI$$

Using this model it is then up to the designer and manufacturer what level of yield impact they consider to be acceptable for a particular type of product. The above considerations are summarized in table 4.2.

STEP	Example	Described in section
Process characterization	(MAM) test structures	4.4.2
↓ Identification of charging steps and layout dependence	Impact of RIE-lag Antenna top area dependence Perimeter dependence Dependence on spacing to neighboring tracks Transistor geometry	4.4.3
↓ Extraction model	Extraction model for each layout dependence	4.5.2
↓ Determination of charging vulnerability index (CVI) for each individual transistor	Computed as a weighted sum of each charging sensitive area for each plasma based manufacturing step	4.5.2
↓ Evaluation of product sensitivity	$\lambda = \int_{CVI=0}^{CVI_{max}} C_s(CVI) \cdot POI(CVI) dCVI$	4.3.2, 4.3.3

**Table 4.2** Flow for quantifying product charging sensitivity

## 4.6 Conclusions

It has been shown experimentally that there is a clear relationship between product yield loss and plasma process induced damage. Based on the results of these experiments a new yield model for charging damage has been proposed that can be used to assess the product dependence of this relationship. In the case discussed in this chapter, the model accurately predicts yield loss as a function of the distribution of antennas that are connected to individual transistors in the products.

Furthermore a new charging monitoring MAM test structure methodology is developed that uses a multiplexing technique in order to save probepad area on silicon so that many different antenna configurations can be measured. This technique proves to be very efficient and has been successfully used to characterize the layout dependence of plasma induced damage for 0.35μm and 0.18μm processes. The results show clearly that there are different damage mechanisms with different layout dependencies. These layout dependencies are not well reflected in conventional methodologies for charging robust design. Therefore designers run the risk of ignoring charging sensitive transistors in the product causing yield and reliability loss.

It is also concluded that the use of diodes or bridges for charging protection purposes has serious drawbacks with respect to circuit area and performance. Therefore there is a need for a methodology that selectively implements diodes or bridges only when needed.

The above considerations have led to a new methodology for the development of charging robust products. In this methodology product layout extraction based on

plasma process characterization with respect to layout dependency are used to assess the product dependent yield loss. It is then up to designers and manufacturers what level of yield loss they consider to be acceptable for certain types of products.

## References

- [1] H.C. Shin and Chenming Hu, "Thin gate oxide damage due to plasma processing" *Sci. Technol.* 1, pp.463-473, 1994.
- [2] S.R. Nariani, et al., "Gate antenna structures for monitoring oxide quality and reliability", *Proc. ICMTS*, 1995, pp. 93-95.
- [3] P. Andrews and A. Blaum, "CMOS-circuit protection against PPID for yield enhancement", *Proc. P2ID symposium*, 1997, pp. 167-170.
- [4] H. Shin and C.Hu, "Thin oxide damage due to plasma processing" *Semicond.sci.Technol.*, Vol.11, pp.463-473, 1996.
- [5] S.R. Nariani and C.T. Gabriel, "A simple wafer-level measurement technique for predicting gate oxide reliability", *IEEE Trans. on Electron Device Letters*, Vol. 16, No. 6, pp. 242-244, 1994.
- [6] Jan-Marc Luchies, Paul Simon, Fred Kuper and Wojciech Maly, "Relation between Product Yield and Plasma Process induced Damage" *Proceedings of the P2ID conference*, pp. 7- 10, June 1998.
- [7] Takayuki Yamada, Koji Eriguchi, Yukiko Kosaka and Kenzo Hatada, "Impacts of Antenna Layout Enhanced Charging Damage on MOSFET Reliability and Performance", *IEDM conference* 1996.
- [8] P. Simon, W. Maly, J.R.M.Luchies and Roland Antheunis, "Multiplexed Antenna Monitoring Test structure," *Proceedings of the P2ID conference*, pp.205-208, June 1998.
- [9] S. Krishnan, K. Brennan, and G. Xing, "A transient Fuse Scheme for Plasma Etch Damage Detection," *Proceedings of the P2ID conference*, pp.201-204, June 1998.
- [10] K.P. Wang, M. Marek-Sadowska and W. Maly, *Proc. 5<sup>th</sup> ACM/SIGDA Physical Design Workshop*, Reston, April 1996, pp.190-197 .
- [11] K. Hashimoto, "New Phenomena of charge damage in plasma etching: Heavy Damage only through dense-line Antenna", *Jap. J. Applied Phys.*, 32, p6109-6113, 1993.
- [12] K. Hashimoto, "Charge damage caused by electron shading effect" *Jap. J. Applied Phys.*, 33, p6013, 1994.
- [13] G.S. Hwang et al., "On the Link Between Electron Shadowing and Charging Damage", *P2ID 1997*, pp. 63-66.
- [14] T. Kinoshita et al., "Analysis of Injection Current Through Thin Gate Oxide During metal Etch", *P2ID 1997*, pp. 45-48.
- [15] J. McVittie, "P2ID 1997", p.433.
- [16] J.C. Arnold et al., "Charging of Pattern features During Plasma Etching", *J. Applied Phys.*, 70[ 10], 1991, p. 5314.
- [17] S. Krishnan et al., "Inductively coupled Plasma (ICP) metal Etch Damage to 35-60A Gate Oxide", *IEDM 1996*, p. 731
- [18] Krishnan et al. "A Transient Fuse Scheme for Plasma Induced Damage Conditions in VLSI Designs", *IEDM 1997* or S. Krishnan et al., *P2ID 1998*, pp. 201-204.
- [19] Paul Simon and Wojciech Maly, "Identification of Plasma Induced Damage Conditions in VLSI Designs", *proc. ICMTS March 1999*, pp1-6.

- 
- [20] Tetsuo Sato, Nobuo Fujiwara and Masahiro Yoneda, "Mechanisms of reactive Ion Etching lag for Aluminum Alloy Etching", *Jpn. J. Appl. Phys.* Vol. 34 1995 pp.1242-2146.
  - [21] Calvin T. Gabriel and Emmanuel de Muizon, "Quantifying a Simple Antenna Design Rule", *Proceedings of the P2ID conference*, May 2000.
  - [22] J-P Carrere, J-C Oberlin, M.Haond, "Topological Dependence of Charging and New Phenomenon During Inductively Coupled Plasma (ICP) CVD Process", *Proceedings of the P2ID conference*, May 2000, pp164-167.
  - [22] J-P Carrere et al., "Electron Shading Characterisation in a HDP Contact Etching Process Using a Patterned Charm Wafer", *Proceedings of the P2ID conference*, May 2000, pp 22-25
  - [23] Martin Creusen et al, "Impact of Reactor- and Transistor Type on Electron Shading Effects", *Proceedings of the P2ID conference*, May 1999, pp. 8-11.
  - [24] Martin Creusen et al, "Impact of Plasma Density and Pattern Aspect Ratio on Plasma Damage in Deep submicron CMOS Technologies", *Proceedings of ESSDERC 1999*, pp.164-167.
  - [25] Srikanth Krishnan and Ajith Amerasekera, "Antenna Protection Strategy for Ultra-Thin Gate MOSFETs", *36<sup>th</sup> Annual International Reliability Physics Symposium*, Reno, Nevada, 1998, pp.302-306.
  - [26] Hiroshi Shirota et al., "A new Router for Reducing "Antenna Effect" in ASIC Design", *IEEE 1998 Custom Integrated Circuits Conference*, pp.601-604.
  - [27] Calvin T. Gabriel and Robert Y. Kim, "Transient Fuse Structures: The Role of Metal Etching vs. Dielectric Deposition", *Proceedings of the P2ID conference*, May 2000, pp 168-171.
  - [28] Reza Rofan and Chenming Hu, "Stress Induced Oxide Leakage", *IEEE Electron Device Letters*, Vol.12, No11, November 1991.
  - [29] Kin P. Chueng, "On the mechanism of Plasma Enhanced Dielectric Deposition Charging Damage", *Proceedings of the P2ID conference*, May 2000, pp 161-163
  - [30] Konstantinos P. Giapis and Gyeong Hwang, "Pattern-dependent Charging and the Role of Electron Tunneling", *Jpn. J. Appl. Phys.* Vol. 37 (1998) pp. 2281-2290 Part1, No 4B, April 1998.
  - [31] Kin P. Chueng and C.S. Pai, "Charging Damage from Plasma Enhanced TEOS Deposition ", *IEEE Electron Device Letters*, Vol.16, No6, June 1995.
  - [32] G.S.Hwang and K.P. Giapis, "On the dependence of Plasma-Induced charging Damage on Antenna Area", *Proceedings of the P2ID conference*, May 2000.
  - [33] P.W. Mason et al, "Relationship Between Yield and Reliability Impact of Plasma Damage to Gate Oxide", *Proceedings of the P2ID conference*, May 2000, pp.2-5.
  - [34] G.S.Hwang and K.P. Giapis, "Modeling of Charging Damage during Interlevel Oxide Deposition in High density Plasmas", *Journal of Applied Physics*, Vol 84, nr 1, July 1998.
  - [35] Gyeong, *Jpn. J. Appl. Phys.* Vol. 84 1998 pp.154-185.
-

---

# Chapter 5

## Design for Manufacturability

**A common language between  
Design, Manufacturing, and Test**

*You know you have achieved perfection in design,  
not when you have nothing more to add,  
but when you have nothing more to take away.*

Antoine de Saint Exupéry

# 5

## 5.1 Introduction

The rate at which new generations of manufacturing processes and IP designs are developed is driven by market needs. The new markets require development of more advanced technologies in combination with construction of ultra modern manufacturing facilities at very high costs (~\$2-3B). The combination of the need for rapid return on investment and extremely narrow market windows for leading edge products put enormous pressure on designers and manufacturers to immediately produce high volume and high yielding first silicon. For a ULSI semiconductor company, the ability to realize fast yield ramp for new technologies and products is therefore becoming a decisive factor to stay in business.

In the early phases of new development projects, design rules are defined and device characteristics are targeted as quickly as possible so that the process and design architectures can be developed simultaneously at a fast rate. However, once design rules and device performance characteristics are fixed, the information flow between design and manufacturing communities often diminishes until production of the first products starts. IP design and process development are often carried out independently.

The question is now whether design rules and simulation parameter files are an adequate common language between the design and process communities to convey information on how design should take place in order to obtain products that optimally fit the manufacturing process. Historically, the same group of people did process development and design of basic cells and consequently process knowledge was transferred to the design. However, because of the increase of complexity of both disciplines, IP development and process development drifted further apart over time. Nowadays, design and process development have a completely different nature and therefore require completely differently skilled people. Often design and manufacturing are carried out in separate organizations that may even be situated in distant locations. Communication is therefore formalized and governed by a simplistic “design rule approach”. This chapter shows that this “throw it over the wall” approach is inadequate to address the rapid yield ramping needs for advanced technologies. It



will be shown that for better manufacturability it is necessary and possible to better integrate design, process and test development. This activity is generally known as Design for Manufacturability or DfM [1].

In other industries, such as the automobile industry, DfM is not uncommon and is accepted as a methodology that is systematically embedded in the company's organization and management structure [2]. Although in the semiconductor industry some DfM activities may take place, it is still far from being accepted as a standard way of working.

The next section discusses the need for DfM in the VLSI semiconductor industry. Section 5.3 proposes a new DfM methodology that joins together IP design, process development and test development in order to achieve high yielding products and fast yield ramp. The subjects that are discussed in previous chapters play an important role in this approach. The consequences of this methodology for design, manufacturing and test are discussed and several examples of DfM techniques are shown in section 5.4, 5.5 and 5.6 respectively. Section 5.7 describes the current and future needs to further implement the DfM methodology. Finally in section 5.8 conclusions are drawn.

## 5.2 A common language between design, manufacturing and test

As stated earlier, the complexity of product design, process development and test development has drastically increased over the past few years. As a consequence, these lines of work have been rapidly drifting apart and are now being executed by different people that are completely differently skilled and are often situated at distant locations. This trend has led to the need for an information exchange format that is simple and can easily be captured in software algorithms. Design rule manuals are an example of this strategy. However, over-simplification inhibits optimal quality and efficiency in the manufacturing of semiconductor products, and therefore a new, more sophisticated approach is needed. In this section this point will be illustrated separately for IP design, manufacturing and test.

### Need for DfM during IP design

Designs for which manufacturing issues are taken into account get to the market sooner, with higher yield, because they fit right into the existing manufacturing process and do not require special procedures. Less engineering resources are drained for costly “fire fighting” of product introduction problems. If a design satisfies the right DfM constraints it will not have to be redesigned for manufacturing. However, in general designers no longer have direct access to the relevant process information that is needed to generate robust designs. Therefore rigid design rules are chosen to convey information about what a layout should comply to, in order to fit the manufacturing process. However, design rules in most cases are very simplistic and do not adequately describe the sensitivities and marginalities that are present in any manufacturing process. Nonetheless, design methodologies and tools are based on those rules. In addition, for design tools the

most important objective is to produce functionality using as less as possible silicon, and therefore the design algorithms make use of the most aggressive design rules by default, *even if it is not necessary to do so*.

Design tools that are used to generate IP-blocks interpret design rules in a rigid way. It is assumed that the yield for design attributes has a digital nature: in case of design rule compliance the yield should be 100%, while in the case of a design rule violation, yield is 0%. On silicon this is not the case. Random defects and design-process marginalities cause the yield to be lower, even at or above the design rule value. Also structures that are designed below design rule are not certain to show 0% yield. Some of the structures will be functional. See also section 2.3.4, figure 2.10.

While a designer strives to minimize the area of a layout, he is constantly making tradeoffs between different design rules. A conventional design rule approach however, does not describe the relative sensitivities of the different design rules in terms of yield loss. The processing window for different design rules can be very different, and the *rigid design rules* that were fixed in the beginning of a development project *do not reflect the design-process marginalities nor the changes in marginality in time*. Therefore, most decisions on design and process architectures are made without assessment of the severity of all possible failure mechanisms or their impact on product yield. In order to design IP that maximally fits the process, information about process marginalities with respect to design attributes must be taken into account. By not doing so, a designer runs the risk of ending up with a non-optimal product-process fit, resulting in yield loss or worse, missing of the market window.

In a DfM approach yield characteristics for all possible design attributes are available so that a designer (or his design tool) is able to weigh the yield impact of the different design options he has. Thereby he has the possibility of making tradeoffs and taking yield into account in his decisions. For example he would be able to assess at what common run length of two metal tracks it is better for critical area reduction to bury of lift a track and add two vias. With the conventional design rule approach this would not be possible. All design rules have the same weight.

In other words: in order to generate robust designs, information on each design attribute and its corresponding relative yield impact is needed. A *ranked list of failure mechanisms* can then be made for a layout, and the designer (or the design tool) can set priorities as to which design attributes need to be changed first to optimize for yield. In a design tool this could be automated for example by minimizing the cost function that comprises the weighted product of yield impacts of all design attributes present in the design.

Not only design rules in the conventional sense are needed, but also information on the yield impact that each of those design attributes will have on the yield of a design. Such a list enables a designer to make tradeoffs between different design rules

Practical examples of techniques that can be used to optimize a design for DfM are given in section 5.4

### **Need for DfM during process development and manufacturing**

Once the design rules are set, process development is done so that manufacturing can accommodate design attributes according to design rules. However, design rules usually only specify basic layout configurations and therefore do not describe the wide variety of layout configuration that are generated by automated design tools. For process development it is however essential to know what can be expected in the product designs with respect to the variety of design attributes and the frequency of occurrence of those attributes. Only then adequate targets can be set with respect to the yields that have to be accomplished on test structures in order to achieve an acceptable yield loss on the product.

For example design rules for vias may specify the size of the via and the minimal metal overlaps. In a product design however, a wide range of via configurations with different via densities and metal overlaps may exist. Design rules usually do not specify the yield impact of those different configurations. For example vias that are situated in very sparse areas may have a higher probability of fail than vias that are situated in an area with nominal density. In order to be able to anticipate on a wide variety of products in a process, it is therefore crucial to characterize the via density distribution in product designs so that adequate test structures can be designed and yield targets can be set.

This is also true for the manufacturing of for example metal structures. By setting minimum spacing design rules for metal tracks it is known what capability the process should have in terms of patterning of the structures. However, it is not known what defect size distributions are needed in order to achieve acceptable yield loss on products due to extra metal particles. This kind of goal can only be set if the critical area generated by the routers is anticipated.

In general it can be stated that in order to be able to develop a process that is able to accommodate a wide variety of products, again (like for design for yield) a *ranked list of failure mechanisms* is needed that indicates what defectivity levels should be achieved for all layout configurations that are present in the products.

Practical examples of DfM techniques in the manufacturing environment are given in section 5.4

### **Need for DfM during test development**

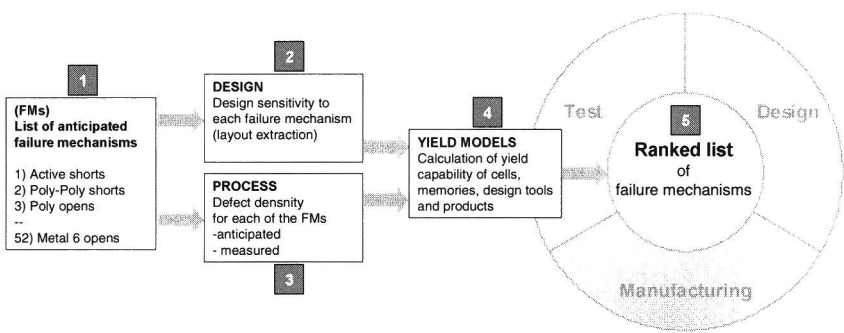
Based on the assumed failure mechanisms automatic test pattern generators produce test vectors. Often only a “stuck-at” fault model is used. In practice such test patterns have good test coverage and as a *by-product* also other failure mechanisms are detected. However, some faults remain undetectable when a simple stuck-at model is used for test pattern generation. The requirements with respect to PPM levels and the increase in the variety of different failure mechanisms in modern manufacturing processes, require a more elaborate assessment of the quality of test pattern generation. Again a *ranked list of failure mechanisms* is of crucial importance for effective test development.

In conclusion one can state that for optimizing the fit of a product design on a manufacturing process, a common language in the form of a ranked list of failure mechanisms is needed. Such a list characterizes the yield impact of all possible design attributes and states their relative importance.

### 5.3 DfM Methodology

The yield of IC manufacturing is determined by the number of ‘killing’ events, the rate at which they occur and the sensitivity of the design to such events. Therefore it seems obvious that the basis for a better common language between design, manufacturing and test should be a characterization of the failure mechanisms and their relative importance. This section describes how such a common language can be achieved.

Figure 5.1 shows the steps that are needed to obtain such a ranked list of failure mechanisms.



**Figure 5.1** Generation of a ranked list of failure mechanisms that can be used in design, manufacturing and test development.

*Step 1: Generation of a list of possible defect causes and related failure mechanisms.*

The process flow is divided into modules and each module is studied to generate an extensive list of all possible failure mechanisms. The list of failure mechanisms is derived by both extrapolation of experiences from older technologies and anticipation on new types of failure mechanisms. Table 5.1 shows an example of (part of) such a list. See also table 3.2.

*Step 2: Design characterization with respect to all failure mechanisms*

For each failure mechanisms in the list its layout dependency of the failure mechanisms is defined so that the corresponding extraction algorithm can be developed. See also table 5.1. Now the critical area for each of the failure mechanisms can be extracted from all types of circuit building blocks. (see chapter 3)

Process module	event	Failure mechanism	Layout dependency	Corresponding yield model	Defect density range	
					Lower	upper
STI	Micro scratch	Active-active short	Active-active critical area	Critical area for shorts	$K=0.01, P=2.8$	$K=0.2, P=4$
	Planarization	poly shorts	Poly-Poly critical area	Critical area for shorts	$K=0.01, P=3$	$K=0.5, P=4$
Poly	Blocked etch	Poly shorts	Poly-Poly critical area	Critical area for shorts	$K=0.02, P=2.8$	$K=0.5, P=4$
Silicide	Missing silicide	Resistive poly / poly opens	Poly width distribution	Critical area for opens	$K=0.01, P=2$	$K=0.1, P=4$
Contact	Blocked etch	Resistive or open contacts	Number of contacts	Poisson (POF)	$POF = 3 \cdot 10^{-6}$	$POF = 3 \cdot 10^{-7}$
Metal1	Blocked etch	Metal shorts	Metal critical area	Critical area for shorts	$K=0.01, P=2.5$	$K=0.4, P=4$
	Particles	Metal opens	Metal width distribution	Critical area for opens	$K=0.01, P=2$	$K=0.1, P=4$
	Planarization	Metal opens	Metal width distribution above densely routed metal areas	Critical area for opens	$K=0.01, P=2$	$K=0.1, P=4$
Via	Blocked etch	Resistive or open vias	Number of vias	Poisson (POF)	$POF = 3 \cdot 10^{-6}$	$POF = 4 \cdot 10^{-7}$
	Patterning	Resistive or open vias	Number of vias above densely routed metal areas	Poisson (POF)	$POF = 3 \cdot 10^{-6}$	$POF = 4 \cdot 10^{-7}$
...	...	...	...	...	...	...

**Table 5.1** *Partial list of possible failure mechanisms with corresponding layout dependencies, yield models and process yield model parameters*

*Step 3: Determination of defect densities for each of the failure mechanisms*

In order to calculate the yield impact of each failure mechanism, its process defect density distribution needs to be characterized. Often this is a difficult task since for many failure mechanisms data is only scarcely available. Defect density data for each failure mechanism can be extracted either from in-line measurements or from test structure data (see also chapter 3). Although for some defects such data is not available, especially in the development phase, upper and lower limits can often be estimated based on engineering experience.

The units in which defect density needs to be formulated depends on the corresponding yield models that apply to the failure mechanism. For example for critical area yield models the range of defect size distributions and defect level ( $K$  and  $p$ ) are needed. For open vias the probability of failure may be sufficient. The more failure mechanisms are listed the more detailed the yield loss pareto can be generated.

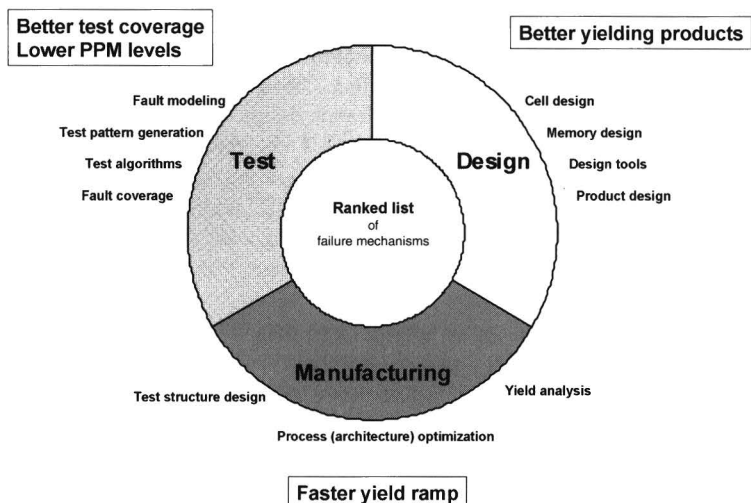
*Step 4: Calculation of the yield impact of each of the failure mechanisms*

Using the layout extraction results, the defect density ranges and the corresponding yield models, now the yield impact range of all individual failure mechanisms can be calculated for any building blocks such as standard cell, memory block , IP block or product. (see also chapter 2)

*Step 5: Ranking of the yield impacts per failure mechanism*

The initial list of failure mechanisms can now be ranked in order of yield impact.

For what purpose the above methodology of ranking of failure mechanisms can be used in design, manufacturing and test development is shown in figure 5.2. In the next section the consequences for the three domains will be discussed in detail.



**Figure 5.2** Use of a ranked list of failure mechanisms in design , test and manufacturing.

### 5.4 DfM in IP design

There usually is little incentive for a designer to optimize his design for yield since he feels yield is the sole responsibility of the manufacturer. Therefore, for most designers functionality and compliance to performance specifications are the only interest.

Design Phases	Design related yield model parameters
Architectural level	<ul style="list-style-type: none"> <li>○ Area of the product</li> <li>○ Sensitivity index per functional block such as micro processors, memories, standard cell blocks, datapaths etc.</li> </ul>
Libraries / memories	<ul style="list-style-type: none"> <li>○ Critical area per cell per layer</li> </ul>
Behavioural description (RTL, VHDL)	<ul style="list-style-type: none"> <li>○ Area and sensitivity index per functional block , e.g. multipliers, registers, adders</li> </ul>
Netlist	<ul style="list-style-type: none"> <li>○ Number and types of standard cells: (e.g NANDs, NORs, Inverters)</li> <li>○ Number of transistors or nets</li> </ul>
Floorplan, place and route	<ul style="list-style-type: none"> <li>○ Area of blocks,</li> <li>○ Number of blocks</li> <li>○ Area of the die</li> <li>○ Critical area for shorts</li> <li>○ Number of vias</li> </ul>
Layout	<ul style="list-style-type: none"> <li>○ Structural description,</li> <li>○ Number of polygons</li> <li>○ Number of layers</li> <li>○ Die area</li> </ul>

**Table 5.2** Different design stages of a product with the corresponding possible yield model inputs.

Many tradeoffs are made without taking into account yield constraints. However, in the design process, at all levels of abstraction, decisions and tradeoffs are made that influence the yield of the final product. Several examples of this fact are illustrated in table 5.2 that shows the different design stages with corresponding design parameters that influence the yield of the final product. Taking into account yield, during library and IP block development and also during the early stages of the product design process can have major beneficial effect on the yield. Especially in markets where profit margins are small, effective use of yield models during the design phases of a product may even determine the economic feasibility of the product.

Yield models can be used in an iterative design process to quantify and visualize yield loss in a design, or as input for constraint driven design. However, as discussed in chapter 2, the practicality of using yield models as indicated in table 5.2 depends very much on the availability and ability to calibrate the corresponding yield model parameters.

The remainder of this section describes several experiments that have been conducted to illustrate how DfM can be used to optimize for yield in various phases of IP design.

### 5.4.1 Standard cells libraries

The effectiveness of optimizing basic building blocks such as standard cell libraries or memory cells is obvious. The resources to optimize the cell designs need to be used only once, while in a later stage the cells are used repetitively throughout the design of many products. Another advantage is that whether or not the library is optimized for yield is transparent to the user and it will therefore not influence the product design time. In practice, cell optimization for yield can be carried out both automatically and manually. Both options are discussed below.

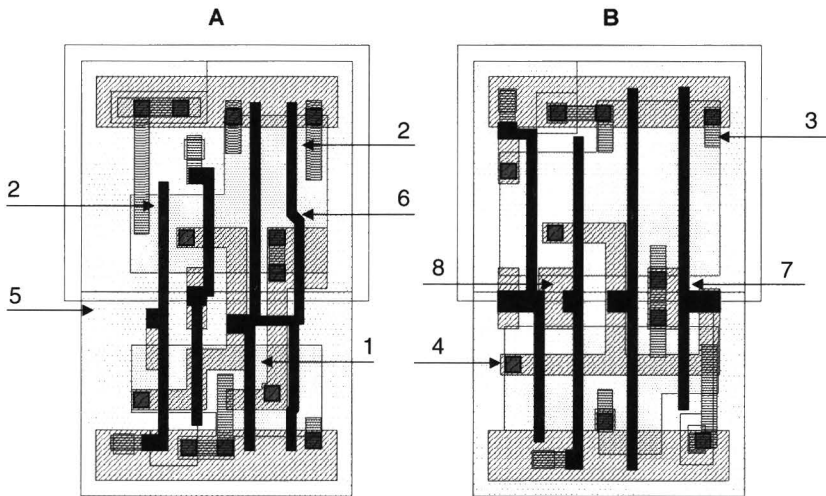
#### *Automatic cell design optimization*

For the design of standard cell libraries or memory cells, usually compaction algorithms are used to migrate an existing design to a new technology. The user defines several boundary conditions such as cell aspect ratio, individual transistor geometries, and the new design rules. The compaction tool then manipulates the layout of the cell until it fits the user defined constraints. However, because yield is usually not taken into account during the compaction, in one of the first steps of the algorithm, many of the dimensions present in the cell are set to minimum values by default in an effort to minimize the new cell size. In many cases this approach leads to cells in which excessive use is made of minimum dimensions, giving rise to unnecessary yield loss. A better compaction approach is to assign costs to the use of certain design attributes and to include the yield model in a cost minimization function. In that way the compaction tool will be able to make the tradeoff between the use of different design rules in terms of yield. A ranked

list of failure mechanisms, as discussed in the previous sections, can be used for that purpose.

#### *Manual cell design optimization*

Yield analysis capability of a layout can also be used in an iterative improvement design process of IP blocks. During the layout phase, the designer (or his design tool) may have many degrees of freedom within the available design space that is defined by the design rules. Therefore he may consider many different layout configurations that all lead to the same functionality. In order to choose the best possible of these options he may consider not only the area of the cell, but also other constraints such as aspect ratio, power or speed. However even within the constraints for area, and performance there are still many layout configurations possible. An example is shown in figure 5.3 that shows two different implementations of the same functionality in a digital core cell. Both cells occupy exactly the same area, but have different layouts. Version A is the result of a “default” cell compaction from a previous technology. Cell B is the result of a manual rework of the cell with “DfM in mind”.



**Figure 5.3** *Digital core cells with same functionality and area. Cell A generated by default compaction. Cell B optimized by hand for DfM.*

Table 5.3 shows the DfM improvements that are made to cell B. It is obvious that such a redesign as is shown above is costly in terms of man-hours, especially if one considers that a complete standard cell library contains many hundreds of different cells. Design optimization of such a library may therefore imply a major

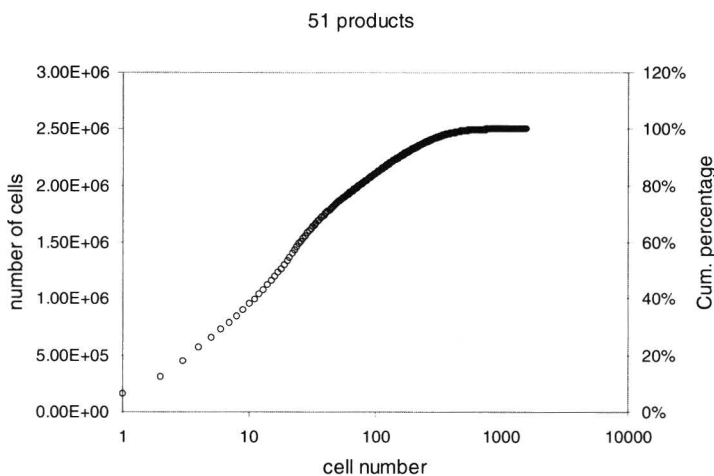


design effort and it may be necessary to prioritize the work and concentrate the yield optimization efforts on the cells that are going to be used the most.

	DfM improvement
1	Less critical area for shorts in metal 1
2	Less critical area for Poly-LIL shorts
3	Extended LIL over active overlaps where possible
4	Extra (0.04μm) metal1 overlap over all contacts
5	More homogeneous usage of cell area
6	45 degree poly angles over active area removed to reduce stress problems
7	Where possible: somewhat wider poly to enhance silicidation
8	Better placement of connection pins to improve routability

**Table 5.3** DfM improvements shown in figure 5.3.

For that purpose the following experiment was conducted. In order to determine what cells are the most frequently used, the library cells of 51 products running in a 0.35 μm technology were extracted using MAPEX-II. All 51 products were designed using the same standard cell library that contained 1500 different cells. Figure 5.4 shows the usage distribution of the 1500 available library cells these products. For 80 % of all the cells that were used in the 51 products, only 75 different cells of 1500 were used (=5%). Clearly it makes sense to concentrate on those 5% of cells to optimize for yield.



**Figure 5.4** Number of used standard library cells in 51 different products

The disadvantage of implementing DfM optimizations by hand, as is done for the example shown in figure 5.3, is that the methodology is not formalized. The tradeoffs the designer is making are based on experience and engineering judgement, and therefore the quality is not consistent. Furthermore the information that is used to optimize for yield depends on the designer's experience. A more structured way of implementing DfM in cell design can only be achieved if the tradeoff procedure is automated in a software tool.

In order to choose the best cell, a yield comparison can then be made based on a ranked list of failure mechanisms. The weak design attributes of each option will become apparent and can be improved in a next version of the design. In order to use this iterative improvement methodology, an adequate yield simulation tool (as discussed in chapter 3) is needed.

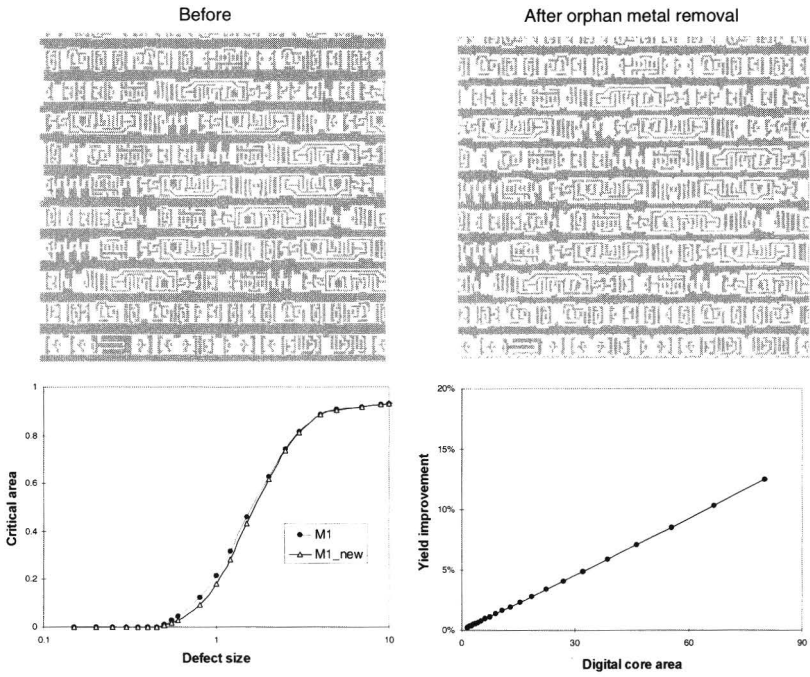
An example of how this methodology can be used to compare different SRAM cells is shown in table 5.4. For this example an arbitrary ranking is chosen.

Design attribute	Weighing factor	SRAM A	Score	SRAM B	Score
Active-Active critical area	0.5	7	3.5	8	4
Poly Critical area	0.7	4	2.8	3	2.1
N+/P+ spacing	0.7	6	4.2	8	5.6
Silicide (poly width)	0.4	2	0.8	5	2
Number of contacts	0.6	6	3.6	8	4.8
Metal overlaps over contact	0.5	6	3	2	1
Metal 1-4 line extensions	0.7	6	4.2	4	2.8
Metal 1-4 critical area	0.8	4	3.2	2	1.6
Number of vias	0.8	10	8	6	4.8
Metal overlap over vias	0.7	8	5.6	3	2.1
Butted contacts	0.3	1	0.3	2	0.6
<b>Total</b>			<b>39.2</b>		<b>31.4</b>

**Table 5.4** *Assessment of the yield capability of two different SRAM cell layouts*

#### *Interconnect design in standard cells*

Depending on the quality of the routing tool that creates the connections between the library cells, a more efficient routing can be achieved when the router has more freedom for “dropping” vias to the standard cells. Therefore some standard cell libraries may equip their pins with extra metal-1 via landing area. A critical area model can be used to analyze the trade-off between the critical area reduction in the higher metal layers due to routing efficiency and the increase in critical area due to the extra metal-1 (sometimes referred to as orphan metal). If it is decided to use a library with excessive metal-1, a post processing design tool can be used to get rid of the orphan metal after routing. In order to assess the yield impact of such a tool, an experiment was done of which results are shown in figure 5.5. The extracted critical area curves for both the original and improved designs, and the resulting yield impact are shown. A critical area yield model is used to assess the justification for such an effort.



**Figure 5.5** Effect of orphan metal removal on critical area characteristic and effect on yield as a function of the digital block area.

In this case critical area reduction by orphan metal removal results in a clear yield improvement, especially for large standard-cell blocks.

#### *Interconnect design in memories*

SRAM cells usually do not require all available metal layers in the process. If, however the manufacturing process offers more metal layers, it might be useful to make use of those by applying wire lifting to the design in an effort to reduce the critical area for shorts. In order to assess this approach an SRAM layout was modified the critical area was compared to the original cell. See figure 5.6. As can be concluded from this figure the critical area for shorts in the SRAM cell is reduced, however extra via connections have to be made. This can make this tradeoff using critical area models for metal and vias.

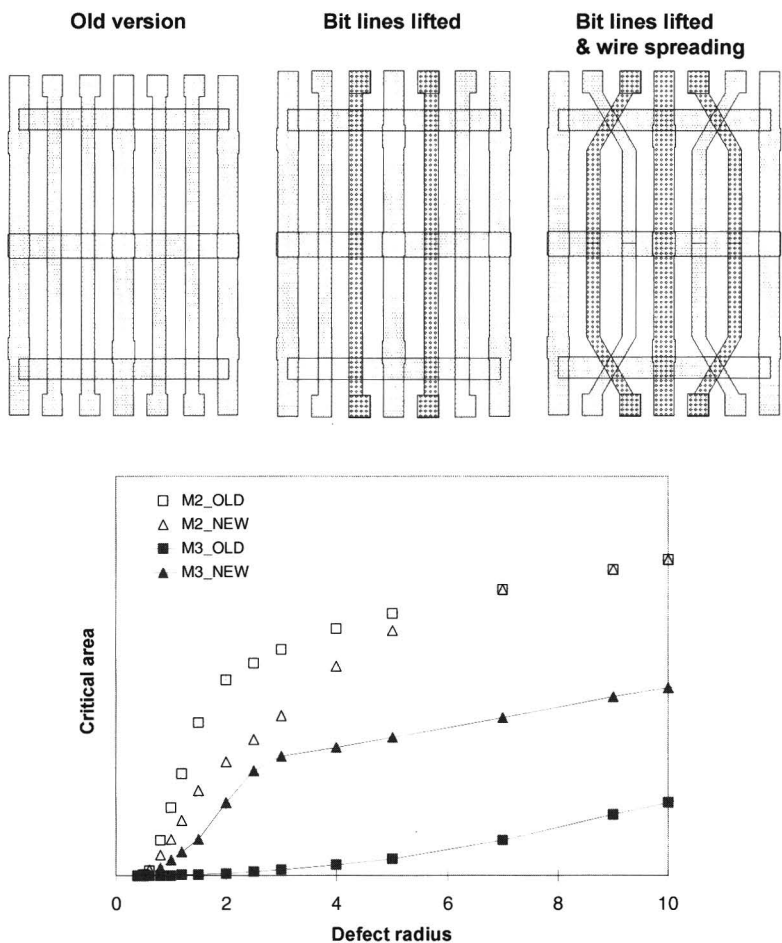


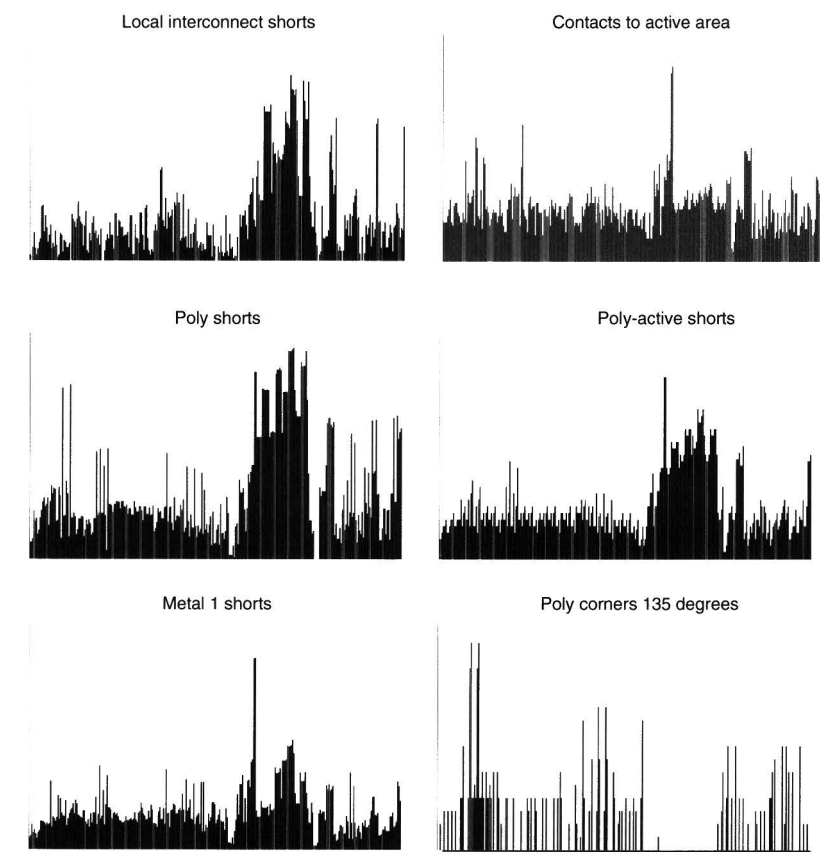
Figure 5.6 Metal layers of an SRAM cell and the effect of wire lifting.

## 5.4.2 Synthesis for high yield

Various behavioral languages may be used to describe functional blocks. These descriptions are not necessarily related to the final physical realization of the circuit. Once the correctness of the behavioral description is verified by simulation, a new description of the circuit on a lower level of abstraction is synthesized. For instance for the realization of a certain Boolean function, several alternatives are possible and the tool that performs the synthesis to the netlist level will make a choice between different available cell combinations. The choice

for certain cell combination is achieved by minimizing cost functions that are governed by user defined constraints such as silicon area, power consumption and speed. Although the area of a cell has a direct relationship with its yield potential, it is not the only parameter that will determine the yield. If the synthesis tool is provided with a figure of merit for yield for each cell and this figure is part of the cost function, it is possible to optimize the synthesis for yield. Such a figure of merit can easily be generated using the ranked list of failure mechanisms. For this purpose a critical area analysis was done for a complete standard cell library of approximately 800 cells. Figure 5.7 shows part of the results of this analysis. Clearly there is a substantial difference in sensitivity to specific failure mechanisms for different cells. Based on these results it is possible to assign a yield index per library cell that can be used during synthesis of IP blocks.

If designs in a certain technology are limited in area by the routing, one might consider using different size library cells for identical functions. Such a library would then contain both high-density cells and high yield cells using somewhat larger area.

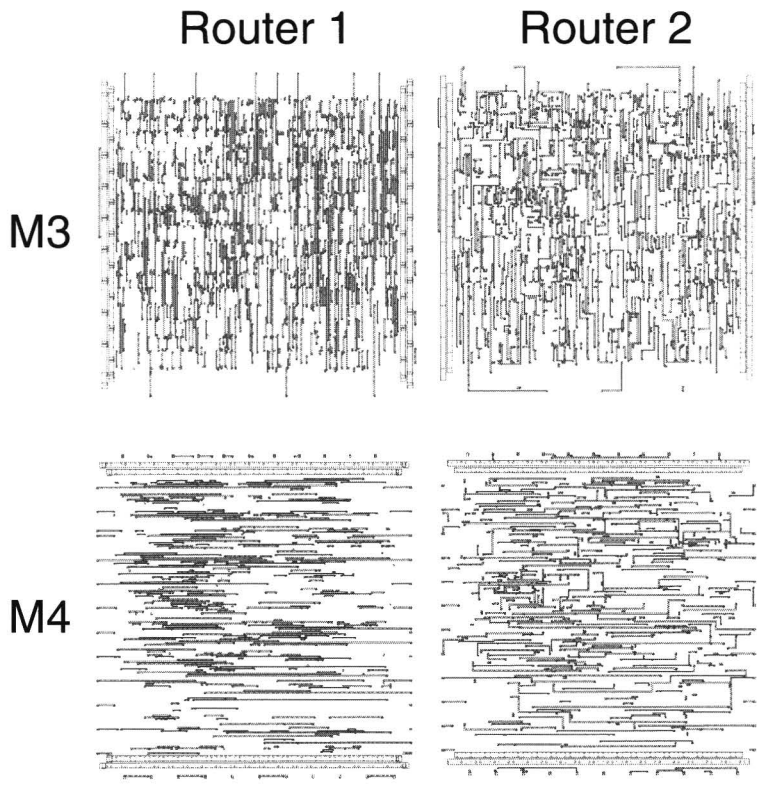


**Figure 5.7** Design characterization for a complete library. Per design attribute critical areas versus library cell number is shown.

### 5.4.3 Routing

The yield of the metal and via layers in products that are manufactured in today's processes is for a large part determined by the critical area for shorts between routed tracks and the number of vias. There are several options possible to optimize the routing in a design for yield. Some of them will be discussed below.

During the routing design tools have the tendency to route at minimum spacing by default. This leads to non-uniform distribution of wires with respect to critical area. Routing at larger spacing obviously would decrease the critical area but would increase the total chip area or would require more metal layers in the manufacturing process. Unfortunately, at this moment no routers are available that effectively use routing constraints that minimize critical area. However, there are differences between routers with respect to the critical area they create. Figure 5.8 and table 5.6 shows the results of an experiment in which the yield capabilities of two different routers were assessed.



**Figure 5.8** Similar standard cell blocks routed with different routing tools resulting in different levels of yield loss. See table 5.7.

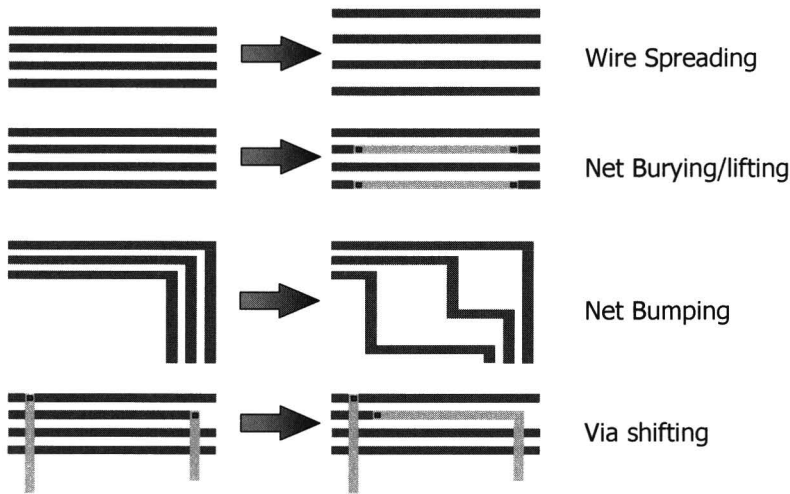
It is clearly shown by the yield prediction based on critical area analysis that different routers can result in a substantial difference in defect sensitivity. Again critical area models are shown to be useful for designers to make the tradeoff between two design tools.

	Router 1	Router 2
Metal1	91.0%	90.8%
Metal2	93.9%	93.5%
Metal3	95.1%	97.4%
Metal4	97.0%	98.4%
Metal5	99.8%	99.2%
Total	78.7%	80.8%

**Table 5.6** Predicted yield of a digital block routed with different routing tools, based on a critical area yield model for shorts.

*Reducing critical area for shorts*

The best way to optimize for yield with respect to shorts in the backend layers is to use minimum spacing as scarcely as possible. However, using a large spacing by default during routing will increase the chip size and reduce the number of dies on a wafer resulting in an overall decrease in efficiency. Therefore routers in general use minimum spacing by default which in some cases results in unnecessary use of minimum spacing. For yield optimization it is therefore beneficial to do a post processing step to decrease the critical area where possible without increasing the die area. Figure 5.9 shows possible solutions to achieve this goal.



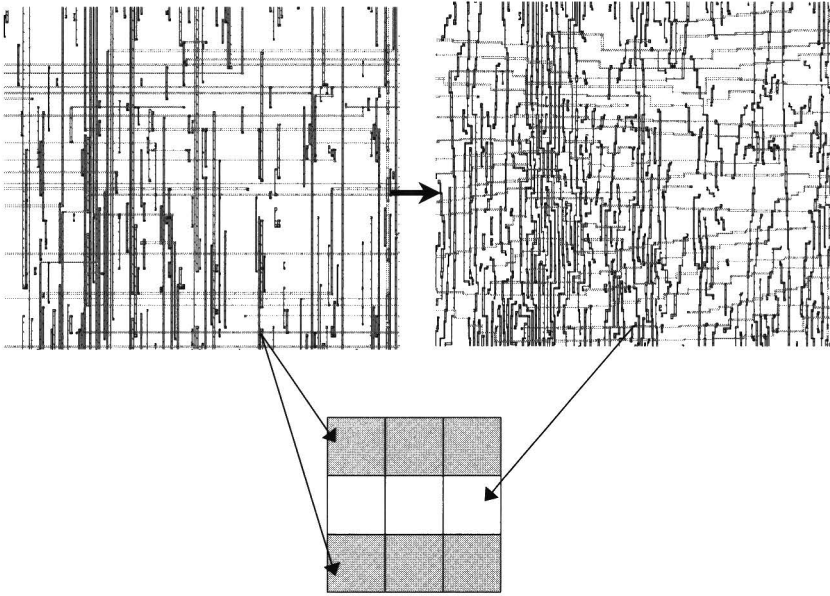
**Figure 5.9** Reduction of critical area in metal routing.

A change in wiring with respect to wire length, number of vias, and spacing, will affect the timing of the device. Therefore it is essential to make any alterations in the routing during the routing itself or just after routing, before timing analysis so that the changes will be taken into account.

Wire lifting or burying only makes sense if the reduction in critical area for shorts justifies the extra yield loss due to the extra vias. This tradeoff can easily be made using the yield models described earlier.

#### *Wire spreading experiment*

To verify the effectiveness of the wire spreading an experiment was done. The wire spreading was done on a real product layout. Both the non wire-spreaded version and the wire-spreaded version of the product were put on a common reticle set so that the yield difference could be measured on wafer level. Figure 5.10 shows part of the layout before and after wire spreading and the reticle layout for this experiment.

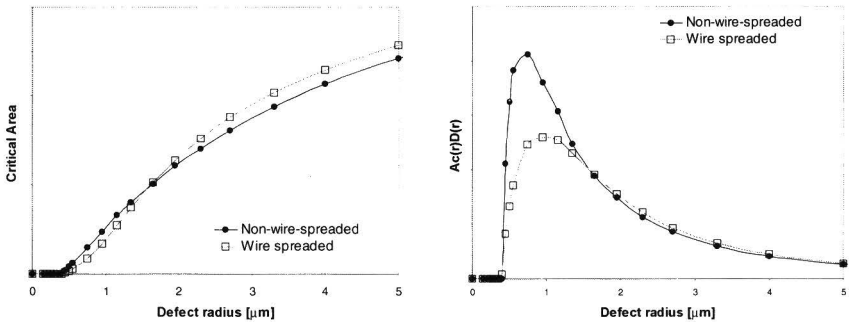


**Figure 5.10** Part of the wire spreaded and original layout of the product. Reticle setup for the experiment.

Figure 5.11 shows the critical area curves and the  $Ac(r)D(r)$  curves for the wire-spreaded and non-wire-spreaded version of the product. (see critical area yield model). The sensitivity of the product for small defects ( $< 1.6 \mu\text{m}$ ) has decreased



while the sensitivity to large defects has increased. Due to the nature of the defect size distribution the predicted overall yield of the wire-spreaded design is higher.



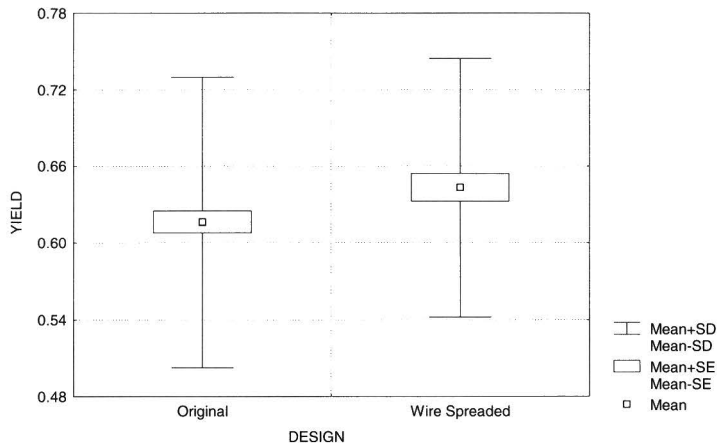
**Figure 5.11** Critical area and  $Ac(r)D(r)$  curves of metal 4 for the wire-spreaded and non-wire spreaded version of the product.

Table 5.2 shows the calculated effective reduction in fault density for each metal layer that was wire-spreaded.

	M2	M3	M4	M5	Average
Original $\lambda$	0.143	0.156	0.110	0.029	0.472
Wire-spreaded $\lambda$	0.140	0.138	0.097	0.026	0.434
Reduction factor	0.981	0.885	0.876	0.886	0.920

**Table 5.7** Reduction of predicted fault density due to wire-spreading for the different layers of the product.

Two lots were processed using the experimental reticle set. Yield of both devices were measured. Figure 5.12 shows the measured yield distributions. A 4% yield difference was measured between the different layouts.

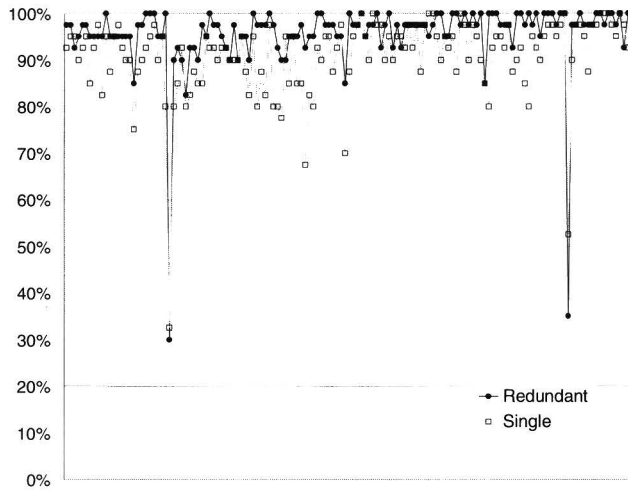


**Figure 5.12** Measured yield difference between the original and wire spreaded product that were manufactured on the same wafers.

The measured increase in yield was in accordance with the predicted yield improvement that was based on a critical area model and a measured defect size distribution on test structures on metal shortloop wafers.

#### *Via doubling experiment*

Depending of the failure mechanism that is causing vias to be (partly) open, placing redundant vias may help reduce the via related yield loss. In the case that large defects are causing the via yield loss, the probability of increasing yield by placing an extra via very close to the original one is very small. However, if a marginality such as metal CD control or another (etch related) problem is causing vias to be open or resistive, doubling the vias may have a beneficial effect on the vias yield as is shown in figure 5.13 which shows a yield trend over several wafers on two similar via test structures. Structure A uses single via chain, structure B uses the same chain, but with redundant vias.



**Figure 5.13** Measured wafer level yield trend for single and doubled via strings.

#### *DfM versus design freedom and design time*

Often designers resist to implement DfM because they consider the extra constraints to restrict their design freedom. However, usually when only conventional design rules are used, there is an overwhelming array of possible solutions to construct the intended functionality, and the designer ends up making arbitrary decisions, possibly ending up with a design that is not optimized for yield. In fact adding extra constraints will limit the number of possibilities for the designer (or his tool) and he will be directed automatically in the right direction, limiting the amount of necessary iterations. Once a layout is finished it will go through a set of subsequent design tools and it will be difficult to incorporate DfM considerations later. The further the cell moves downstream in its development flow, the harder it will be to satisfy additional DfM constraints. The cost of DfM changes rises drastically as the cell progresses toward production. Therefore it makes sense to implement DfM in the design flow as early as possible.

Another reason for designers to oppose DfM actions is that the consideration of more constraints will delay the completion of their design. However, by not taking DfM into account at all, a designer runs the risk of ending up with a non-manufacturable design, delaying the end product even more. Nonetheless, if possible, DfM constraints should be formulated in such a way that they can be captured in software algorithms in an automated design flow. For example a compaction tools used to develop new cell libraries should take into account DfM constraints in its cost functions. Routers should not route at minimum distance by default, but try to spread the wires as homogeneously as possible while limiting the amount of single vias and antennas. DfM actions will then be implemented automatically and remain transparent to the end user.

## 5.5 DfM in the manufacturing environment

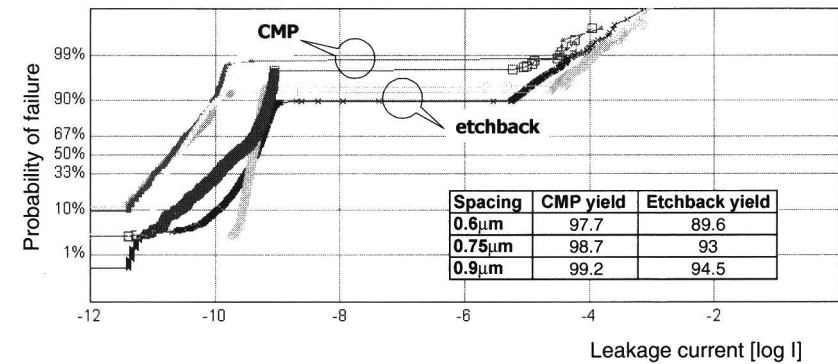
This section describes the implementation of DfM in the manufacturing environment. Process development, product characterization and yield prediction are discussed. Results of several experiments in these domains are shown.

### 5.5.1 Process development

For process development and yield ramping the DfM methodology can play an important role to:

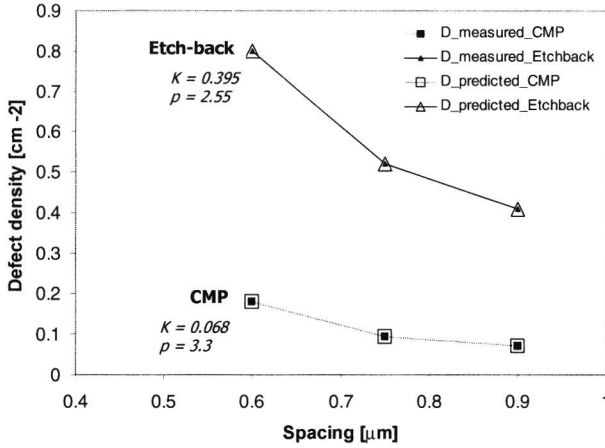
- 1) *Generate a ranked list of failure mechanisms* for each product that is going to be manufactured. A clear picture of the product specific sensitivities is obtained (See chapter 3).
- 2) *Design adequate test structures* that reflect the failure mechanisms that occur in the products. The test structures used need to reflect the layout sensitivities of the products that are manufactured. The DfM methodology enables to assess the relationship between the test structure sensitivity and product sensitivity to specific failure mechanisms. (See chapter 3)
- 3) *Setting clear targets with respect to the defect levels* that need to be realized for each failure mechanism to achieve target product yields. Product target yield can be broken down to target yields for each individual layer or failure mechanism. Using the DfM methodology those targets can be translated into yield targets that need to be obtained on test structures.
- 4) *Characterize products* in order to be able to adjust process parameters for product specific processing. For some processing steps it is difficult to achieve a process window that fits the complete range of products. Better performance can be achieved if the process parameters can be adjusted to the products. For example a metal etching recipe can be sensitive to the amount of metal on the wafer that needs to be etched. If the amount of metal that needs to stay on the wafer is too small, etching problems may occur that result in yield loss. If the metal coverage of the product is known the recipe can easily be adapted to that. Layout extraction can also determine sensitive locations in the die for certain failure mechanisms. Such locations can then be used to do in-line measurements for process control.
- 5) *Assessing and predicting the yield impact of process changes or options*. During continuous yield improvement activities split experiments may be done to study the yield impact of certain process options on products or test structures. Although a significant yield impact may be shown on a test structure, it remains a difficult task to extrapolate the results to product yield. In order to be able to predict product yield impact based on the experimental

test structure results, yield models that take into account the sensitivity of both are indispensable. As an example, an investigation of the yield impact difference of via Tungsten etchback versus Tungsten CMP was conducted. Figure 5.14 shows the cumulative distribution of leakage currents measured on large metal comb-meander test structures (see for example figure 3.4) on a lot on which a split has been performed for both process options.



**Figure 5.14** Comb-meander leakage current distributions (spacings: 0.6 μm, 0.75 μm and 0.9 μm, see fig .3.4) for Tungsten CMP and etch-back.

Test results for the comb-meander structures with different spacings are shown. The leakage currents between comb and meanders are measured. In the current leakage distributions a yield criterion is set at  $10^{-7}$  A, resulting in the yield table in figure 5.14. Clearly the etch-back process has higher defect levels than the CMP process, resulting in better yields on the test structures for the CMP process. However, in order to justify a process change from Tungsten etch-back to Tungsten CMP, the test structure results need to be extrapolated to a yield improvement prediction on real products. In order to do this, the defect size distribution parameters  $p$  and  $K$  for both process options are determined from the test structure results by fitting the critical area based yield prediction of the test structures to the test results as is shown in figure 5.15.



**Figure 5.15** Defect density measurements and predictions as a function of spacing of the comb-meander test structures.

The defect size distribution parameters  $K$  and  $p$  are taken as the fitting parameters. Once  $K$  and  $p$  are determined for both processes, they can be used for the critical area based yield predictions of a set of products as is shown in table 5.5.

Defect density Prediction	Product A	Product B	Product C	Product D	Product E
D_frontend	0.18	0.20	0.16	0.16	0.18
D_Via	0.07	0.08	0.04	0.05	0.06
D_Metal_etchback	0.55	0.60	0.44	0.39	0.50
D_Metal_CMP	0.16	0.18	0.13	0.13	0.15
Dp_etchback	0.80	0.89	0.65	0.60	0.74
Dp_CMP	0.41	0.47	0.34	0.33	0.39
Dp Improvement	0.39	0.42	0.31	0.27	0.35

**Table 5.8** Yield and defect density ( $D$ ) predictions for products A-E based on defect size distribution parameter extraction.

The above experiment shows that with DfM techniques it becomes possible to evaluate the yield impact (and thus the economic justification) of a possible process change on the complete manufacturing volume in a fab. Without proper product characterization and ability to translate test structure results to product yield this would not be possible.

## 5.5.2 Understanding product variability

When introducing new products into a fab, it is important to realize that one can distinguish two situations:

### 1) *Multiple sources of design styles*

In case of a manufacturing line that is accepting designs coming from many different sources, the range of products may be very wide. Many different memories, libraries, and design tools can be used to design the products. This has two disadvantages. First, since there are many sources of design there is no control over the designs other than a design rule manual. Therefore not much can be done to optimize the designs for yield. In addition, the resulting large range of products requires the manufacturing process to have very wide processing windows or to be flexible in the sense that the manufacturing process can be adapted to each individual product. In order to be able to assess in advance what process parameters need to be changed for a particular product, the layout needs to be characterized. This type of flexibility in changing the process requires in depth process knowledge to be able to translate layout parameters to process recipes. In addition the fab logistics and control systems need to be able to handle this flexibility. In other words: being able to handle a wide range of products in one manufacturing process requires a substantial effort and adds to the manufacturing cost of the products. Therefore manufacturing lines that are in this situation invest in minimizing the possible number of design styles by issuing libraries and memories that have been developed in conjunction with the process itself. DfM methodologies as described in the previous section play an important role to realize such high yielding building blocks. In addition extensive characterization of incoming products is indispensable.

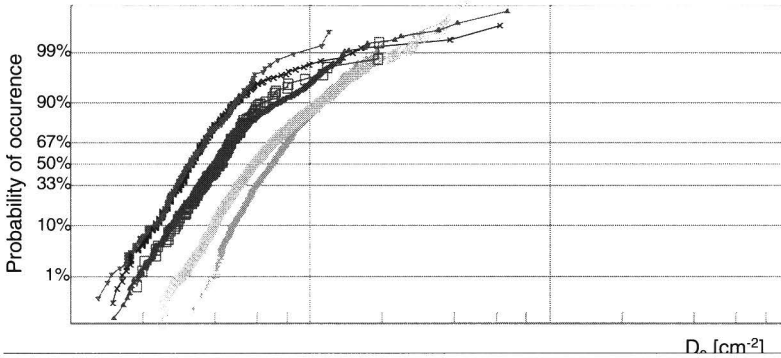
### 2) *Single source of design styles*

A manufacturing line that only manufactures products that are designed with building blocks originating from one common source have a clear strategic advantage with respect to the yield capabilities. However, this advantage can only be capitalized on if a DfM methodology is implemented adequately. For example, extensive effort can be put into the development of high yielding building blocks and design tools. The benefits will then show on all products that are manufactured in the fab. The range of products in the fab will therefore be less wide, and product specific processing may not be necessary (to the extent it is necessary for fabs that run product with multiple source designs). Therefore manufacturing costs can be lower.

In both situations described above it is still essential to be able to assess products that are coming into the fab. As described before, there may exist a significant difference between the different yields of different products running in the same process that cannot be explained by the yield models that only take into account average process defect density and area of the product. In any manufacturing environment there exists a need to understand those differences in order to enable the engineers to prioritize on what product to focus their attention (the product that yields below the expected yield). To illustrate this, the cumulative

distributions of wafer level  $D_0$ 's for different products manufactured during a period of three months in a  $0.35\ \mu\text{m}$  process was studied. See figure 5.16.

Even though from the normal distribution of the wafer level  $D_0$ 's it can be concluded that the products and the manufacturing process are in mature state, a large variety of median  $D_0$  per product can be observed. The figure also shows



**Figure 5.16** Cumulative distribution of  $D_P$  values on wafer level for different products manufactured during a period of three months.

that the concept of using a calculated average process- $D_0$  to predict the yield of all products can lead to significant errors for individual products. More sophisticated yield models and product characterization methods that take into account design density or critical area must be used to understand the yield differences.

The remainder of this section describes first results that have been obtained with the MAPEX-II MAE system (chapter 3) that has been used to characterize product differences in a manufacturing environment. Examples are shown of how products that are manufactured in the same process can show significant structural differences.

#### *Example 1: Choice of Library.*

Figure 5.17 shows a portion of a logic core of three different products running in a  $0.35\ \mu\text{m}$  process. The different products are designed with different library cells resulting in a clear difference in metal 1 density.

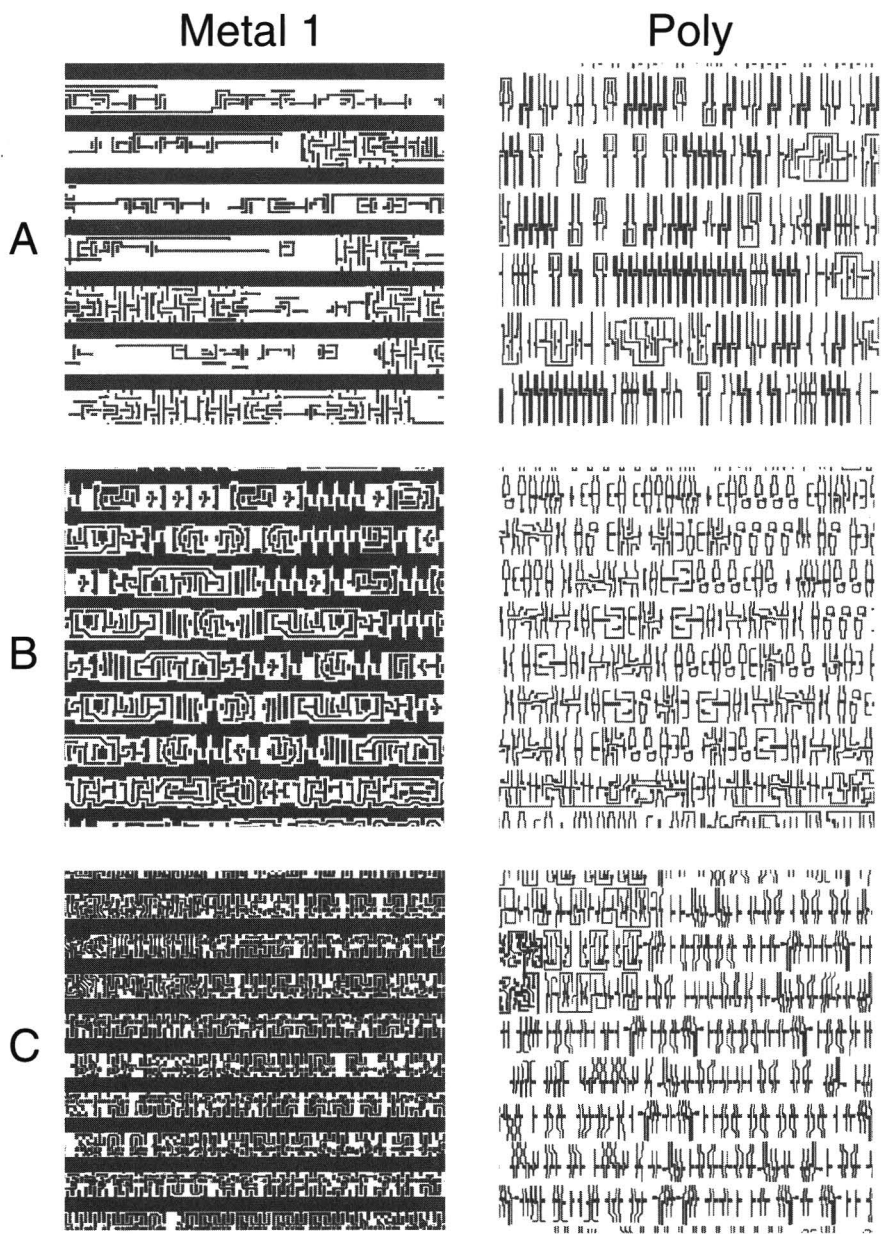
Because of the obvious difference in metal-1 density, the sensitivity to yield loss due to metal-1 bridging is different for these products. Product A is the least sensitive to this failure mechanism. It is a shrunk version of a  $0.5\ \mu\text{m}$  process design. It also uses a large amount of decoupling cells that do not use much metal-1. The logic core of product B has the same netlist, but has been re-synthesized using another library resulting in an overall smaller block with increased sensitivity to defects on metal-1 level. The metal density is further increased by the different layout of the decoupling cells and by redundant metal-1 that is added to the input/output pins of the cells to give the router more landing flexibility for vias. Product C uses a third library of which the pitch in Y direction is similar to the library used in product B. Nevertheless, the increased pitch in X



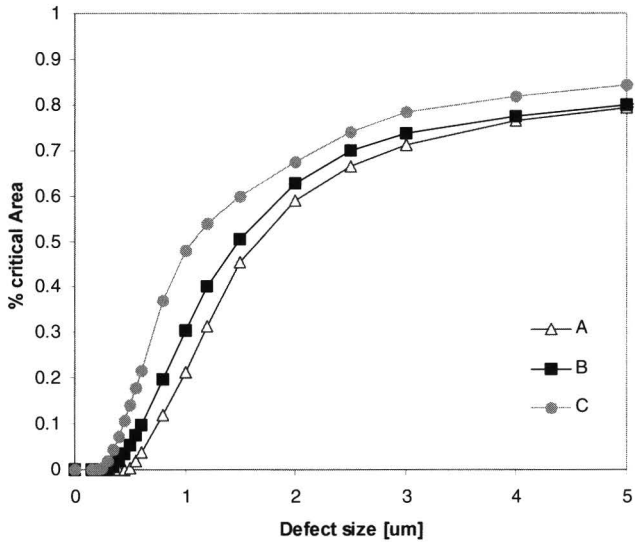
*Chapter 5*

---

direction results not only in a higher transistor density, but also a higher metal density. The extremely high number of contacts to active in this library causes an even higher metal density. Figure 5.18 shows the extracted percentage of critical areas for the three products as a function of defect size. The sensitivity to shorts for small defects (that are much more likely to occur than larger defects) is dominated by the choice of library.



**Figure 5.17** Part of the layouts of different products in which different libraries are used.



**Figure 5.18** Critical area as a function of defect size for metal-1 are determined by the library choice.

*Example 2: Differences in routing strategy*

Figure 5.19 shows the extracted percentage of critical area on the total die area as a function of defect size for different products in a three metal layer 0.35  $\mu\text{m}$  process.

The poly layer is only used for transistors and local interconnect within library cells and memories. A large spread in sensitivity for large defects in poly can be observed. For example for defects with a radius of 10  $\mu\text{m}$ , product A is twice as sensitive as product F.

The sensitivity to shorts in metal-1 is determined by the library cells and embedded memories. Only limited amount of signal routing is done in metal-1. The critical area curve for metal-1 shows less spread in sensitivity than for poly. Product B is the most sensitive.

Sensitivity to shorts in metal-2 is dominated by signal routing density. In this respect product B is also the most sensitive. For this product a relatively large design effort is put into increasing the transistor density in the logical core, causing many areas where routing congestion is high. Metal-3 is mostly used for power routing and a limited amount of signal routing. Therefore the overall sensitivity of all products in metal-3 is relatively low, even for large defects. Product B is the most sensitive for small defects while product A is more sensitive to large defects in metal-3.

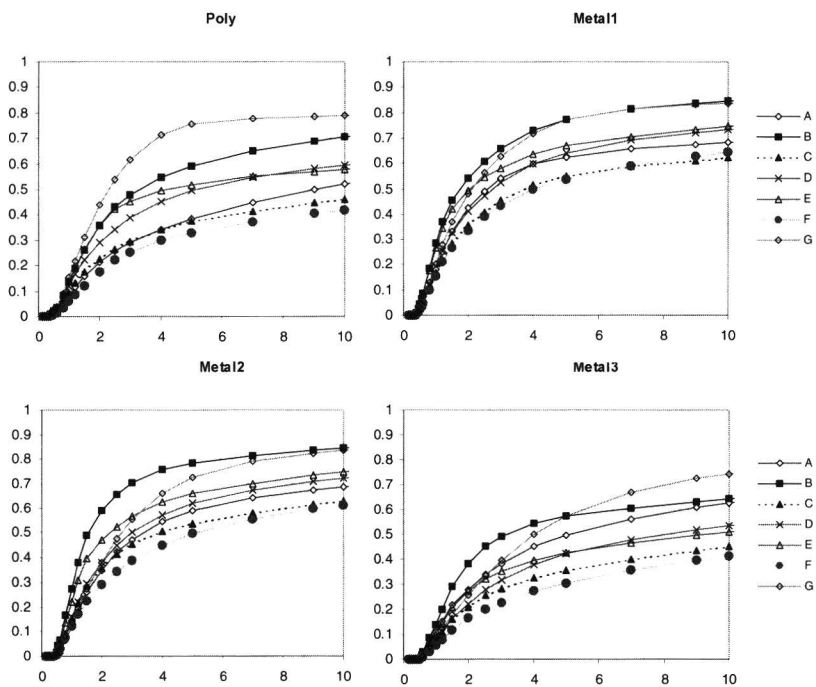


Figure 5.19 Percentage of critical area for different products.

Figure 5.20 shows the calculated impact on Do of the different metal layers for the different products. An arbitrary defect size distribution of  $0.5/r^3$  is assumed for all layers.

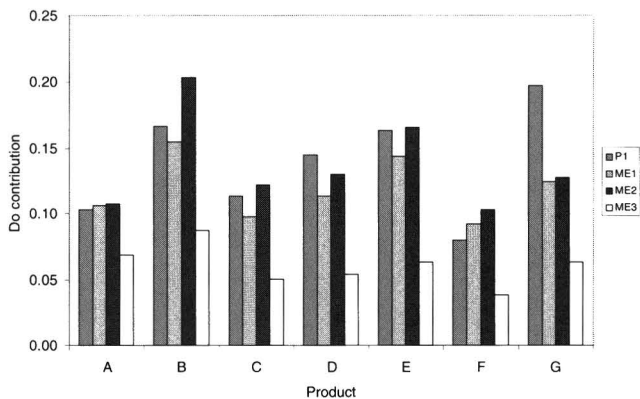
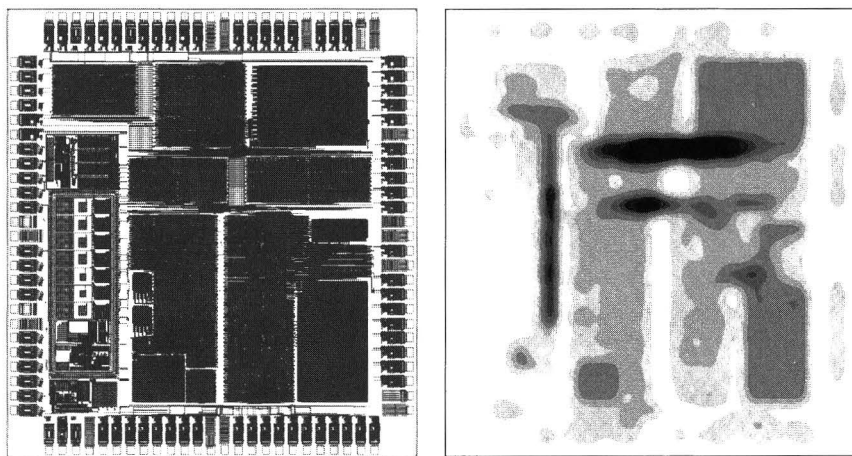


Figure 5.20 Do contribution for different layers of different products.

From the above it can be concluded that the sensitivity to yield loss by metal bridging defects is dominated by the routing strategy that is being used.

Figure 5.21 shows a metal 1 layout of a product and the corresponding sensitivity map for bridging defects with a diameter of  $1\mu\text{m}$ . The gray scales in the sensitivity map indicate the defect sensitivity. The figure shows clearly that in this case the bus systems in between functional blocks contribute to a large part of the overall sensitivity of the circuit. The bus routing strategy determines the spatial sensitivity distribution within a product.



**Figure 5.21** Layout (left) and extracted sensitivity map (right) for metal 1 bridging defects. Darker areas indicate higher sensitivity.

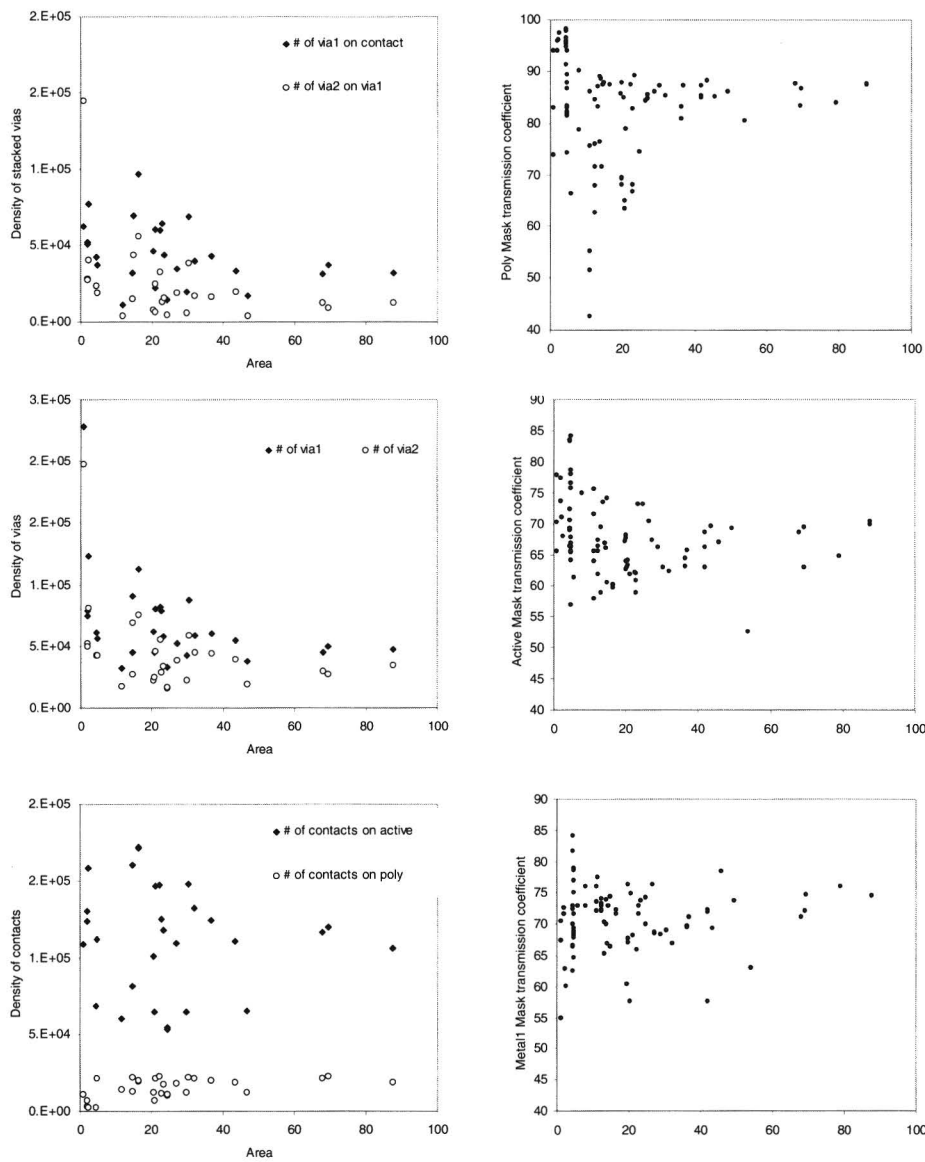
### *Example 3: Vias, contacts*

Figures 5.22 a,b,c show the differences in contact, via and stacked via density for different products. Again the difference in routing tools or strategies cause a substantial difference between products in this respect. The level of sensitivity to yield loss because of these parameters is different for these products and can not be predicted by the die size. Depending on the probability of failure for vias in the process this will result in a difference in yield loss contribution for different products.

### *Example 4: Coverage*

Figures 5.22 d,e,f show the mask transmission coefficients for the poly, active and metal-1 level as a function of die area. Although one could expect similar characteristics from similar products in the same process, there is a significant spread in coverage of these masks leading to a large difference in amount of metal, poly, active or resist material that needs to be removed during etching or

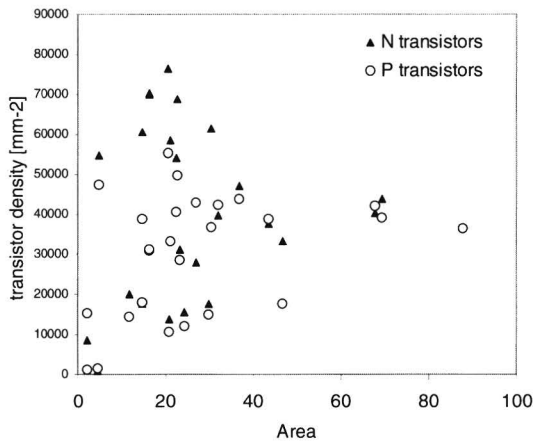
ashing. When a process marginality with respect to one of these parameter arises, the level of yield loss may be product dependent.



**Figure 5.22** Density of contacts, vias and stacked vias for different products and mask transmission coefficients for different products.

*Example 5: Transistor density*

The transistor density of a circuit is determined by many factors such as the amount of embedded memory, the library choice, the amount of decoupling cells and the placement and routing strategy. Figure 5.23 shows the transistor density as a function of die area for different products.

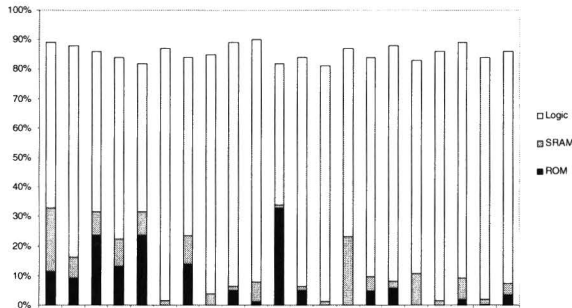


**Figure 5.23** Transistor density of different products as a function of die area.

Again it is shown that similar products in the same process with similar die area may have very different transistor densities.

*Example 6: Embedded memory use*

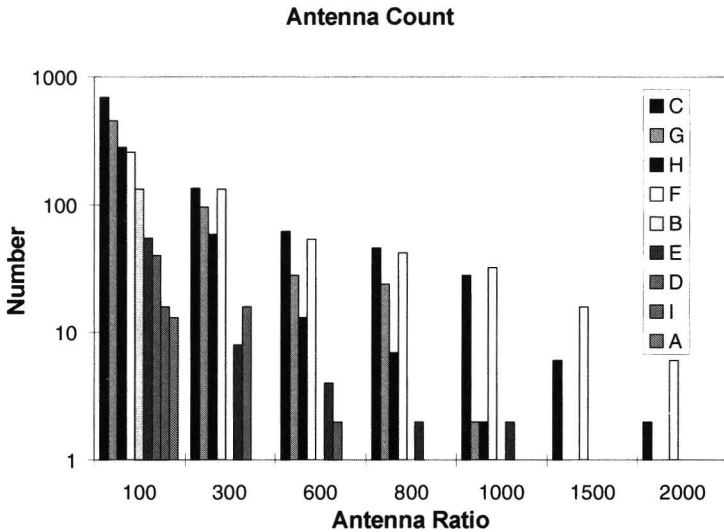
Figure 5.24 shows for a range of products within the same process the amount of the die area that is used for SRAM, ROM, and logic core. A large spread can be observed.



**Figure 5.24** Embedded memory usage in different products

*Example 7: Sensitivity to charging*

Figure 5.25 shows the extracted number of antennas in different products as a function of the antenna ratio. Again a large spread in charging sensitivity can be observed. For instance product A has only few small antennas,, while product C and F have a large number of small antennas but also and some very large ones.



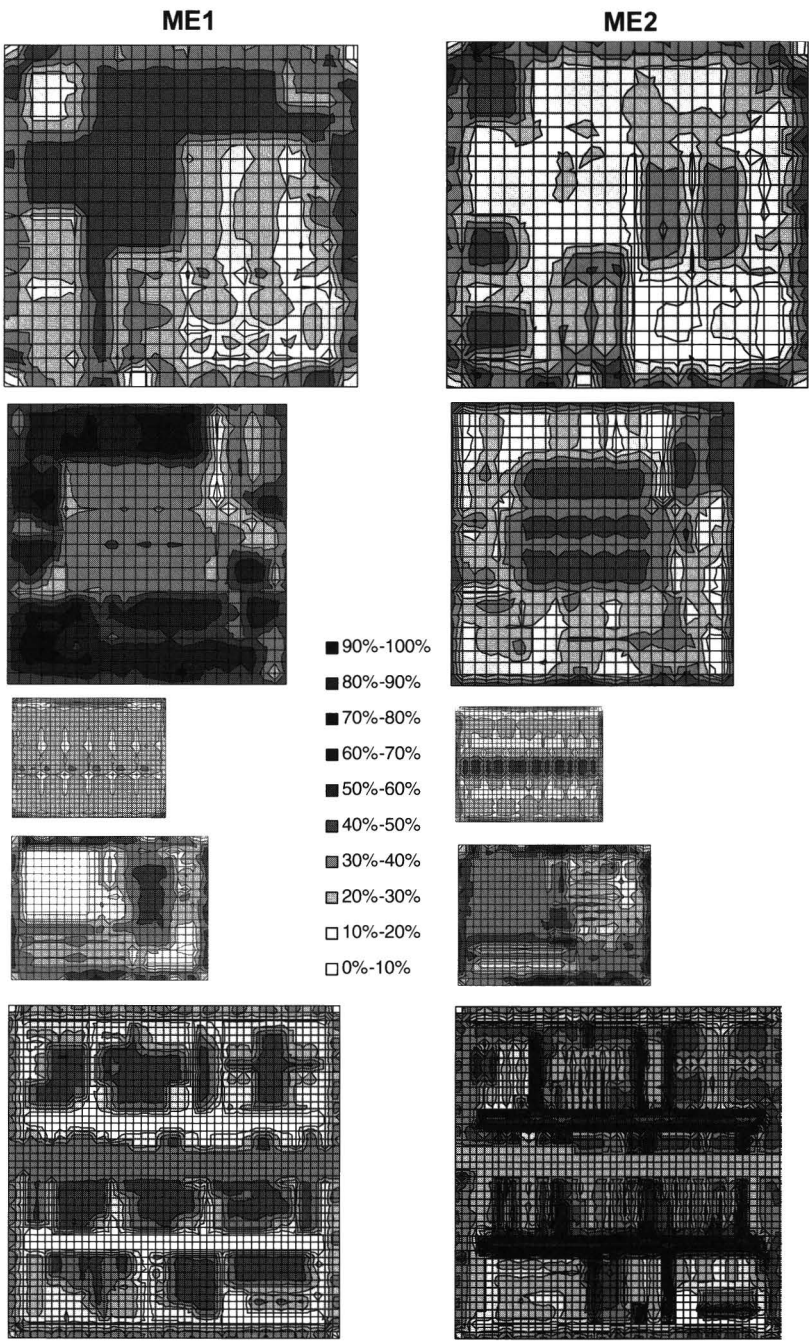
**Figure 5.25** Antenna pareto for different products

*Example 8: Metal coverage distribution*

Many manufacturing processes such as litho, etching and CMP steps are influenced by local density of patterns on the wafer. Figure 5.26 and 5.27 show metal density distribution across the die for several products. A large variety of metal coverage distributions can be observed.

From the above experiments it can be concluded that even though different product may be manufactured in the same process, their layouts can show substantial differences in sensitivity.





**Figure 5.26** Metal coverage for different products (M1, M2)

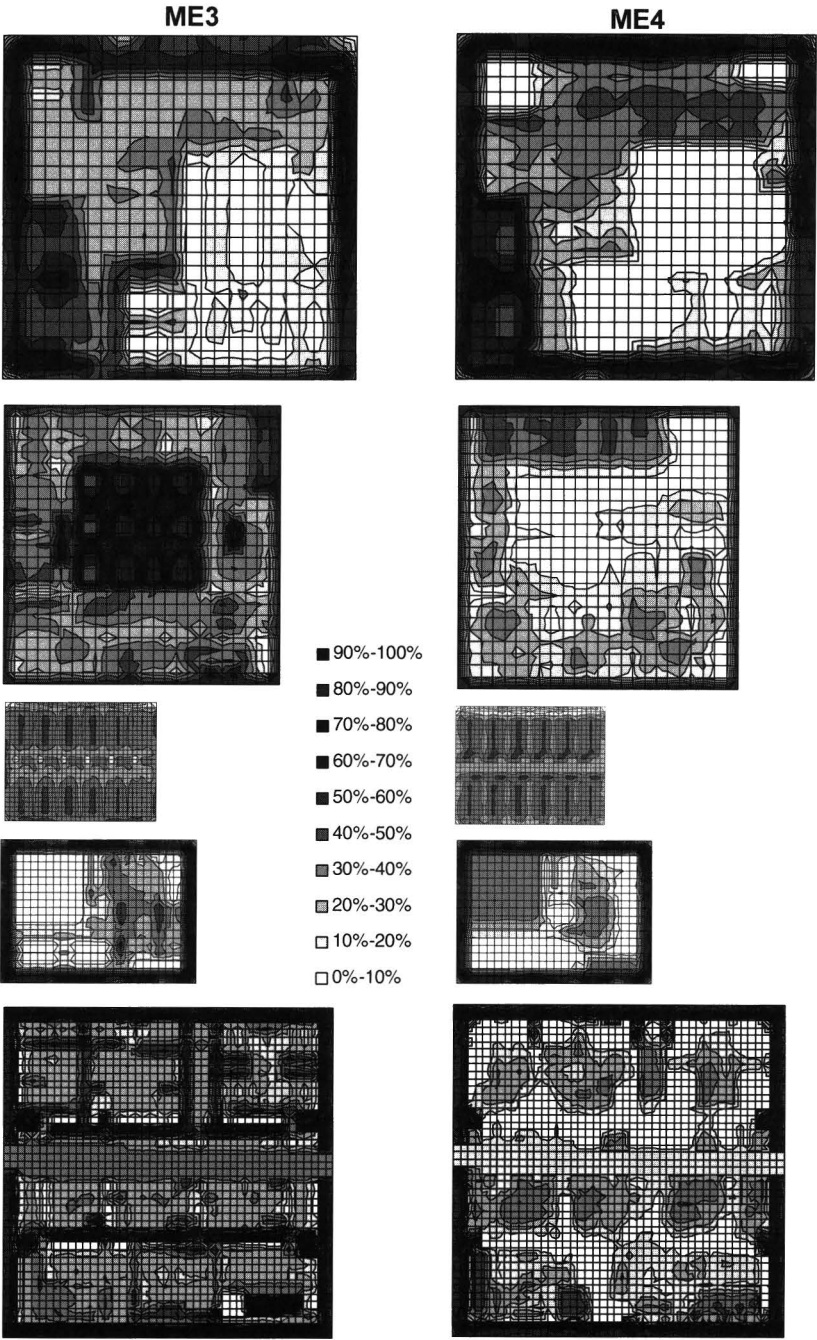


Figure 5.27 Metal coverage for different products (M3, M4)

### 5.5.3 Yield prediction and sensitivity analysis

In manufacturing environments yield models are used for product cost prediction or for tailoring the quantity of wafers that need to be manufactured to the market demand. It is obvious that inaccuracy in these predictions lead to financial losses. In that respect adequate yield models are a primary factor in the success or failure of the manufacturing line or product. When more products are being fabricated at the same time, usually an average  $D_0$  is calculated to report a fab's performance. When a semiconductor fab produces a small range of similar, mature products, this is often an adequate way of predicting yield. However, when the diversity of products is larger, the variations of levels of sensitivity to defects of the different products can be significant, as is shown in the previous section. In such cases it becomes more difficult to adequately fit the models. This immediately shows the shortcomings of using yield models that are only based on the average process defect density and chip area. When the new product of which the yield needs to be predicted is substantially different in terms of design style from the products that are used to determine the model parameters, the yield prediction may be far to optimistic or pessimistic resulting in subsequent financial consequences.

In order to account for product sensitivity without critical area extractions, a product sensitivity index,  $\Psi$ , can be added to the existing yield models to compensate for the different design styles. For this purpose the fault density can be rewritten as:

$$\lambda_{product} = \Psi_{product} \cdot A_{product} \cdot D_0 = \Psi_{product} \cdot \lambda \quad (5.21)$$

Depending on the application,  $\Psi$  expresses the design complexity in terms of for example the number of transistors or the ratio of areas of different design styles in the chip. In the latter case  $\Psi$  can be calculated as

$$\Psi_{product} = \frac{\sum_{i=1}^{i=K} \psi_i \cdot A_i}{A_{product}} \quad (5.22)$$

for a product containing  $K$  design styles, in which  $A_i$  is the area of the parts of design style  $i$ , and  $\psi_i$  is the sensitivity index for design style  $i$ , indicating the relative sensitivity of that design style. Table 5.4 shows an example of different design styles and the associated complexity factor  $\Psi$ .  $\psi_i$  for a particular manufacturing process can be obtained by fitting model (5.22) to test data of several products. Obviously, block level yields that are binned out separately after testing are needed.

Design style $i$	$\psi_i$
Standard cell logic	1
SRAM	1.5
DRAM	1.6
ROM	0.8
Analog	0.2
IO	0.3
MTP	1.9
OTP	5.3

**Table 5.9** Sensitivity correction factors for different design styles.

$\psi_i$  can be calibrated either by critical area extractions for each design style, or by fitting the predicted yields to the measured yield. The advantage of such a model is that an accurate yield prediction can be done taking design complexity into account without having to perform critical area calculations for all products. In addition such a model can be used for yield prediction in the early stages of the product design.

If a more detailed, layer level, yield prediction is needed, more design parameters can be extracted from the products and a MAE, such as MAPEX-II can be used to explain these yield differences. An example of such a yield prediction is shown in table 5.10 that lists the extracted design attributes that have been considered for this analysis. The total yield of a product is estimated by calculating the yield loss contributions of each design parameter  $i$ . The product yield is calculated as the product of all corresponding terms:

$$Y_p = \prod_{i=1}^n Y_i$$

for  $n$  design attributes.

In this calculation the Poisson model is used for the yield terms related to vias, contacts, landing pads and transistors. For the conducting layers a critical area yield model is applied. Where possible process dependent parameters such as defect size distribution and defect density parameters per design attribute are measured using shortloop test structures. See chapter 3. In order to fit the predicted product yields to the measured yields, a least squares algorithm is used in which the difference in predicted yield and measured yield is minimized by varying the process dependent yield model parameters such as defect densities. For that purpose a  $i$ -dimensional defect density space is defined in which each defect density parameter is allowed to vary within a specified window of the measured value. So, within this defect density space, parameters are varied so that for  $K$  products with  $i$  extracted design attributes

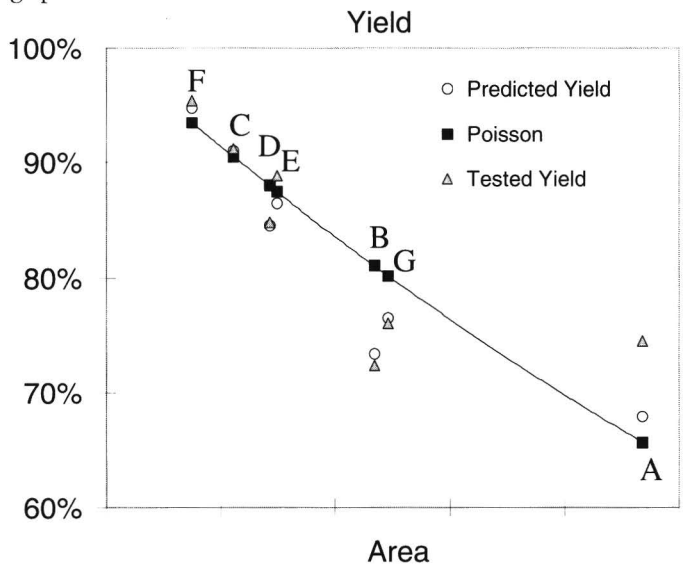
$$\sum_{k=1}^K \left( \prod_{i=1}^n Y_{ik} - Y_{mk} \right)^2 \rightarrow 0$$

in which  $Y_{mk}$  is the measured yield of product k. In this way it is possible to verify whether within the assumed defect density limits the differences in product yield can be explained based on the extracted design attributes. Figure 5.28 shows the predicted product yields and the measured yield as a function of die area. Also the yield as predicted by the Poisson model (taking only die size into account) is shown. Although very simple yield models have been used in this analysis, the yields predicted by taking more design attributes into account do agree much better with the measured values then the yields predicted by the Poisson model. The relatively low yields for products B and G can therefore be explained by their high sensitivity to defects in the metal layers. The predicted yield for product C is slightly higher than the tested yield. This observation leads to the conclusion that its yield loss can not be explained by the critical area yield models as applied in this analysis and that there is a parametric related yield loss problem.

Product name	A	B	C	D	E	F	G
Area	X	X	X	X	X	X	X
Yield impact							
N transistors	0.9990	0.9984	0.9997	0.9995	0.9991	0.9998	0.9985
P transistors	0.9992	0.9992	0.9997	0.9996	0.9995	0.9999	0.9985
Gateox area	0.9983	0.9983	0.9994	0.9991	0.9991	0.9993	0.9971
Total poly on active edge	0.9923	0.9922	0.9978	0.9963	0.9959	0.9985	1.0000
Total poly on locos edge	0.9985	0.9984	0.9996	0.9993	0.9990	0.9998	1.0000
Co_on_active	0.9928	0.9954	0.9980	0.9971	0.9967	0.9985	0.9955
Co_on_poly	0.9990	0.9991	0.9997	0.9994	0.9995	0.9999	0.9989
V1	0.9972	0.9973	0.9985	0.9981	0.9979	0.9989	0.9981
V2	0.9984	0.9987	0.9991	0.9990	0.9990	0.9994	0.9990
Stacked via on contact	0.9991	0.9992	0.9995	0.9994	0.9992	0.9997	0.9995
Stacked via2 on via1	0.9998	0.9998	0.9999	0.9998	0.9998	0.9999	0.9998
Poly	0.9080	0.9251	0.9751	0.9578	0.9547	0.9882	0.9075
ME1	0.9051	0.9301	0.9785	0.9668	0.9599	0.9864	0.9405
ME2	0.9044	0.9093	0.9732	0.9620	0.9539	0.9847	0.9391
ME3	0.9376	0.9600	0.9889	0.9841	0.9821	0.9943	0.9692
Total predicted yield	0.68	0.73	0.91	0.86	0.85	0.95	0.77
Total predicted Do	0.41	0.66	0.42	0.49	0.59	0.36	0.54
Do contribution							
N transistors	0.0011	0.0033	0.0014	0.0018	0.0032	0.0015	0.0031
P transistors	0.0009	0.0018	0.0012	0.0015	0.0017	0.0007	0.0030
Gateox area	0.0018	0.0036	0.0025	0.0029	0.0033	0.0045	0.0060
Total poly on active edge	0.0083	0.0167	0.0099	0.0124	0.0145	0.0102	0.0000
Total poly on locos edge	0.0016	0.0034	0.0018	0.0023	0.0034	0.0016	0.0000
Co_on_active	0.0077	0.0099	0.0090	0.0098	0.0116	0.0103	0.0091
Co_on_poly	0.0011	0.0019	0.0015	0.0019	0.0019	0.0009	0.0023
V1	0.0030	0.0057	0.0068	0.0064	0.0073	0.0071	0.0038
V2	0.0018	0.0029	0.0039	0.0033	0.0036	0.0040	0.0020
Stacked via on contact	0.0009	0.0017	0.0021	0.0020	0.0029	0.0023	0.0010
Stacked via2 on via1	0.0002	0.0004	0.0006	0.0006	0.0008	0.0007	0.0005
P1	0.1030	0.1664	0.1134	0.1447	0.1631	0.0796	0.1972
ME1	0.1066	0.1549	0.0979	0.1134	0.1439	0.0919	0.1247
ME2	0.1073	0.2033	0.1222	0.1301	0.1661	0.1033	0.1277
ME3	0.0688	0.0873	0.0501	0.0538	0.0634	0.0386	0.0635

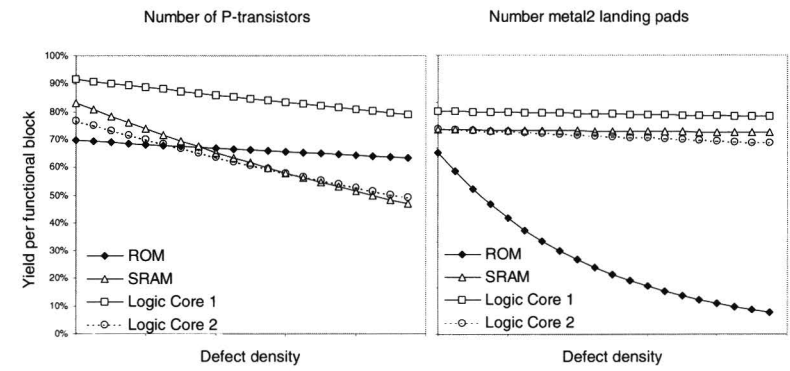
Table 5.10 Yield and Do impact matrix for the different products.

It was shown that there was a design sensitivity to a transistor parameter for this product. The high yield of product G relative to the predicted yield is explained by a somewhat lower test coverage of the test program during that period of time. The above example therefore shows that once there is a high level of confidence in the extracted design parameters and the corresponding yield models, the MAPEX-II system is not only capable of predicting yields of products but can also distinguish defect related yield loss from parametric yield loss or suggest test coverage problems that otherwise would not have been remarked.



**Figure 5.28** Yield measurements and predictions.

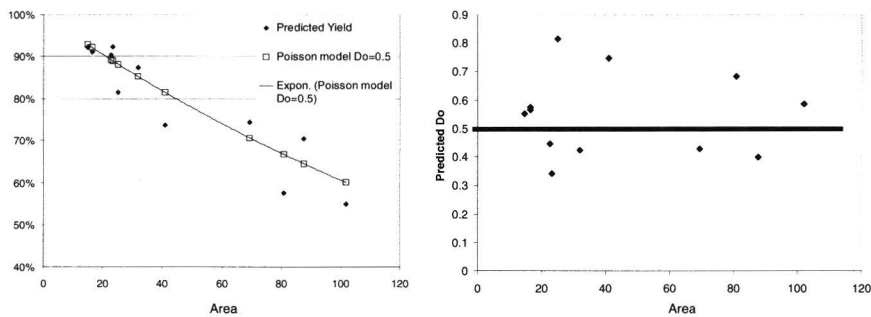
Another example of how the MAPEX-II system can be used to assess the difference of (parts of) products is the generation of yield sensitivity analyses for each design attribute. Figure 5.29. shows part of such an analysis for a product that contains ROM, SRAM and two logic cores.



**Figure 5.29** Sensitivity analysis of different blocks within a product.

For this analysis the partial yields of the different blocks in the product are calculated for different defect density levels for P-channel transistors and metal-2 landing pads. To obtain these figures the system calculates the partial yields for the different blocks while increasing the defect density level for only one design attribute. Other defect levels are kept constant. This sensitivity analysis shows that the ROM block in the product is much more sensitive to yield loss for landing pads than the SRAM or the logic cores. Increase defect levels in P-channel transistors will cause abnormal yield loss in the SRAM and logic core2.

Figure 5.30a shows the MAPEX\_II predicted yield as a function of die area of 12 products. The same design attributes as listed in table 5.10 were used. Figure 5.30b shows the corresponding Do values for these products.



**Figure 5.30** Yield predictions and the corresponding  $D_p$  values for 12 products within one process.

A large spread in predicted Do is observed. From this it can be concluded that different design styles design can cause a large spread in the sensitivity of different products in the same process. It is therefore not reasonable to assume that an average process Do can be used to predict the yield of every product. Each product shows specific sensitivities, and in order to predict yield, costs and the required number of lot to be started for a particular product, all relevant design attributes have to be considered.

## **5.6 DfM for test development**

For reduction of PPM levels all possible failure mechanisms and their relative importance need to be evaluated. Opens or soft failures such as resistive vias are examples of failure mechanisms that are only partly covered by the conventional test strategies based on stuck-at models. A ranked list of failure mechanisms for a product can be used to assess whether the implemented test algorithms deliver sufficient test coverage. Also for the reduction of test times it is necessary to be able to assess the efficiency of different test algorithms with respect to their capability to detect multiple failure mechanisms.

## **5.7 Current R&D needs for the implementation of DfM**

Although in most semiconductor industries some of the DfM methodologies discussed in previous sections are used, they are not yet well established and embedded in the organization in a structural way, nor are there adequate industrial tools available. Below the needs for further implementation of the DfM methodology for future technologies are discussed.

### **Yield modeling**

As mentioned in chapter 2, the existing yield models are quite adequate to describe defectivity related yield loss. However for advanced technologies, beyond 0.25 $\mu\text{m}$ , many subtle design-process marginalities play an increasingly important role. In order to implement DfM methodologies successfully, more sophisticated yield models that describe these marginalities are needed. Examples of failure mechanisms for which new yield models are required are:

- Yield loss due to non-planarity after CMP
- Yield loss due to local or global density variations within the die or within wafer.
- Yield loss due to lithographic and patterning deficiencies (OPC)
- Yield loss because of performance related issues. For example operating frequency, cross-talk, leakage, power.



### **Yield model parameter extraction**

As discussed in chapter 3, not only the yield models themselves are needed, but also the capability to characterize the process defects in order to be able to calibrate the yield model parameters. For future technologies this is becoming an increasingly difficult task. In particular in-line inspection techniques have serious disadvantages such as recipe and product dependence of the measuring sensitivity and the difficulty of determining killing potential of defects that are found. Therefore an adequate test structure approach is needed (section 3.3).

As test structures need to reflect the sensitivities that are present in the products that are manufactured, this methodology is also becoming an increasingly complex task. New manufacturing processes are developed in order to integrate orders of magnitude more functionality on the same silicon area. Consequently, not only the number of transistors, but also the number of other design attributes is increasing rapidly. Conventional test structures are no longer adequate in terms of their critical area. Silicon real estate that is available for test structures and test costs are limiting the number of failure mechanism that can be characterized.

Although the above is true for many design attributes, here an example is given for the characterization of via resistance:

Products that are designed in future technologies will have many tens of millions of vias. Therefore the probability of fail for vias should be very low in order to achieve reasonable yield on the product. In order to verify such low probability of fail, the via test structures need to have at least a number of vias that is comparable to the number of vias in a product. However, for conventional test structures, the number of vias in a string is limited by the measurement resolution. Therefore more sophisticated test structures need to be developed that are able to characterize the resistance of very large numbers of vias with sufficient resolution and with minimum test costs.

### **Design characterization tools**

Since in future technologies the number of different failure mechanisms will increase, there will be a need to characterize designs for a larger number of design attributes. Layout characterization tools are needed that are able to extract thousands of layout parameters within reasonable time limits. Being able to implement such new extraction algorithms at a fast rate is very important. (section 3.4)

### **Design tools**

DfM tools that are needed in the design flow can be divided into two categories: design analysis tools and design enhancement tools. Design analysis tools help a designer visualizing problem areas in his design and quantify manufacturability effects. Design enhancement tools are tools that are able to automatically optimize for yield without user interference.

In order to minimize resistance of designers to embed DfM in the already complex and time consuming design flow, modifications for yield should be transparent to the end user. DfM should therefore be built in the relevant design

tools. For future technologies the software algorithms for compactors, synthesis tools or routers, need to be able to handle more complex constraints that not only take into account functionality, but also manufacturability. (section 2.4.1. and 5.3.1.). Design tool vendors that are able to embed manufacturability constraints in their tools will definitely have a strategic advantage over other those that don't.

**Cost modeling**

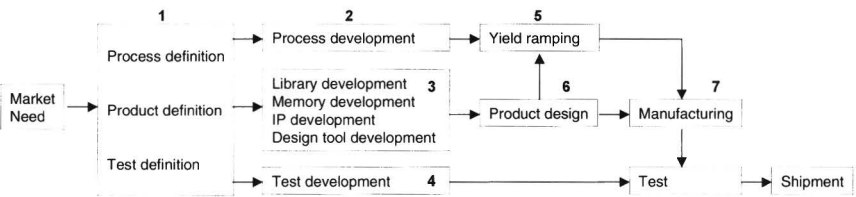
Modification to the design or manufacturing of a product in order to optimize yield can only be justified if financial impact from the yield increase is larger than the extra engineering efforts. This is particularly the case at the early stages of the lifetime of a manufacturing process. In later stages the process is improved and defect levels go down until they will reach a level where the yield improvement from DfM may not justify the extra design effort anymore. Although for some modifications the benefits may appear obvious, in other cases this is not so clear. In order to make the tradeoff in such cases, cost models that take into account the whole business process of manufacturing ICs (design, processing, fab logistics, testing, packaging etc) are needed.

**Organisation**

Embedding of a DfM methodology in a manufacturing or design organization can be difficult because the organizations responsible for the implementation are not the same organizations that are benefiting from it. In most cases DfM implementers are not held directly accountable for the benefits of DfM. Therefore DfM can only work if top management supports it. Top management is necessary to sanction the time and effort for developing DfM.

**5.8 Conclusions**

figure 5.11 shows a simplified business process for the development of new technologies. The numbers indicate where DfM can play a role of importance. Table 5.3 summarizes some examples of DfM opportunities and their benefits.



**Figure 5.32** *Points in the VLSI business processes where DfM can have beneficial impact*

		Examples of DfM opportunities	Benefits
1	Project definition	<ul style="list-style-type: none"><li>▪ Tradeoffs for process architectures (e.g. # of metal layers)</li><li>▪ Process and design FMEA</li><li>▪ Feasibility study of migration paths</li></ul>	<ul style="list-style-type: none"><li>▪ Risk assessment of the projects</li><li>▪ Product-process fit evaluation</li><li>▪ Yield ramping targets and requirements</li></ul>
2	Process development	<ul style="list-style-type: none"><li>▪ Set yield goals</li><li>▪ Setting defect density targets per process module that are needed to achieve yield goals</li><li>▪ Test structure design / monitoring</li><li>▪ Fault modeling and yield modeling</li></ul>	<ul style="list-style-type: none"><li>▪ Clear picture of what targets need to be achieved in the manufacturing process to meet yield requirements for first product.</li><li>▪ Test structures to verify those targets</li></ul>
3	Yield ramping	<ul style="list-style-type: none"><li>▪ Yield ramping methodology: deconvolution of yield loss into yield loss pareto</li><li>▪ Adequate test structures and yield vehicles</li><li>▪ Setting defect density targets per process module that are needed to achieve yield goals</li></ul>	<ul style="list-style-type: none"><li>▪ Clear yield loss pareto</li><li>▪ Faster yield ramp</li></ul>
4	Manufacturing	<ul style="list-style-type: none"><li>▪ Product yield prediction</li><li>▪ Product introduction risk analysis</li><li>▪ Yield analysis</li></ul>	<ul style="list-style-type: none"><li>▪ Insight into product specific sensitivities (predictability)</li><li>▪ Faster product yield ramp</li></ul>
5	IP development	<ul style="list-style-type: none"><li>▪ Library development for yield</li><li>- Yield constrained compaction</li><li>- Yield constrained synthesis</li><li>▪ Memory development for yield</li><li>- yield constrained design</li><li>- redundancy approach</li><li>- Assessment of different libraries and memories form different sources</li></ul>	<ul style="list-style-type: none"><li>▪ High yielding building blocks</li></ul>
6	Product design	<ul style="list-style-type: none"><li>▪ Design tools for yield optimization</li><li>- wire spreading, via doubling, OPC, tiling etc..</li><li>▪ Design for yield guidelines</li></ul>	<ul style="list-style-type: none"><li>▪ Optimal product – process fit</li><li>▪ Product yield optimization</li></ul>
7	Test	<ul style="list-style-type: none"><li>▪ Fault modeling</li><li>▪ Test Methodology</li></ul>	<ul style="list-style-type: none"><li>▪ Lower PPM levels</li><li>▪ Better test coverage</li></ul>

**Table 5.11** *DfM opportunities and benefits*

In this chapter it is shown that embedding of DfM into the IC development process can have substantial benefits. As design and process technology development progresses, the number of complex failure mechanisms that are determined by design-process marginalities will increase drastically. The ability of a company to rapidly take into account these failure mechanisms in design, manufacturing and test developments in a flexible way will determine whether a company will be able to address the rapidly changing market demands. By embedding DfM into the organization, predictable yield and reliability can be assured by design and process control rather than expensive analysis and re-engineering. Products will reach the market sooner because they are designed first time right. Therefore DfM methodologies are an important factor in the semiconductor industry that needs to be able to deliver high yielding products at a fast rate. DfM may make the difference between being competitive or not succeeding in the market place. Using DfM as a common language between design, manufacturing and test is therefore a definite strategic advantage.

References

- [1] W.Maly, “high levels of IC Manufacturability: One of the Necessary Prerequisites of the 1997 SIA Roadmap Vision”. IEDM 1998.
- [2] Ramon Bakerjian, “Tool and Manufacturing Engineers Handbook”. Vol.6 “Design for Manufacturability”. ISBN. 0-87263-402-7.

---

# **Chapter 6**

## **Summary**

# 6

## Summary

Due to the enormous costs associated with VLSI manufacturing today, the ability to perform a fast yield ramp has become vital for any semiconductor company. In a manufacturing environment, understanding of failure mechanisms and the related yield models is not only important for volume planning, but also for bringing yield to an acceptable level as fast as possible. Yield learning is a complex activity that can only be achieved if engineers are able to quickly identify the failure mechanisms and, more importantly, when they are able to assess the impact of these failure mechanisms on product yield. Only then they are able to set priorities in possible improvement actions. This thesis describes methodologies that clarify the relationship between defects in manufacturing processes, the sensitivity of products to these defects, and yield.

Yield modeling also plays an important role in the design of VLSI products. Several experiments in this thesis show that adequate yield modeling capabilities during all design stages, may even determine the economic feasibility of a product. Presently, however there is little incentive for designers to take into account yield considerations during the already complex design process. Often this is caused by ineffective transfer of information on the vulnerabilities of the manufacturing process to the design environment. Design guidelines are a too simplistic representation of reality, or they are not embedded in design algorithms. On the other hand, non-robust designs are the result of the designer's assumption that manufacturing yield is the sole responsibility of the IC manufacturer. It is shown in this thesis that this assumption can no longer be justified for present and future deep submicron technologies. For a designer to be able to take into account yield, a methodology is needed to assess the impact of different failure mechanisms on different design solutions. This thesis describes such a methodology.

In *chapter 2* an overview is given of the existing yield models that have been developed over the past few decades. The benefits and drawbacks of these yield models for different manufacturing and design applications are also discussed. Furthermore, experiments are done in order to validate new yield models that use

design and process parameters that are relatively easy to obtain. The new models are valuable for both design and manufacturing applications.

In *chapter 3* it is shown that next to the yield models themselves, new yield model parameter extraction methodologies for both the product design and manufacturing process are needed in order to achieve useful and accurate yield prediction. The test structure based approach for obtaining process related yield model parameters described in this chapter plays an important role in achieving this goal. Practical applications of this methodology for both the design and manufacturing environments are described.

Since yield is not only a function of the failure mechanisms that are introduced by the manufacturing process, but also of the sensitivity of the product to these failure mechanisms, characterization of the product design is an essential part of any yield prediction system. Therefore *chapter 3* also describes design characterization methods. The development, implementation, and use of a state of the art design characterization system in an industrial environment is described. Product characterization experiments show significant differences in sensitivities between IC's that are manufactured in similar technologies and have similar functionality. Several experiments show that the ability to extensively characterize design differences is invaluable for the analysis of product specific yield loss.

In the experiments, described throughout this thesis, that were conducted in order to understand product yield, it became clear that in many cases the following train of thought needed to be followed repeatedly:

1. It is necessary to *understand and characterize the failure mechanism* in terms of layout sensitivity and statistical distribution across wafers and lots. Either modeling or a test structure based approach, can be used.
2. *Assessment of the impact of the failure mechanism on the product yield.* Often this comes down to developing a yield model specific to the failure mechanism and developing a corresponding design characterization method. Once these objectives are achieved, the yield loss with respect to the failure mechanism can be evaluated for any design.
3. *The yield model can then be used as a guideline to develop a methodology for robust design.*

To demonstrate the effectiveness of the above approach, it was put through the test using the plasma damage phenomenon as an example. In *chapter 4* it is shown that in advanced IC manufacturing technologies plasma damage increasingly influences manufacturing yield and reliability. Both design and process have to be considered to solve the complicated problem. It is shown that the existing methodologies for robust design are too simplistic and inadequate to address the real problems the designer is facing. Only an all-encompassing approach that takes into account both the process conditions and the product design results in the ability to realize charging robust products.

Extensive experimental data shows a clear yield impact of plasma damage. Therefore different charging mechanisms and their relationship with yield loss are studied using measurements on specially designed plasma damage test structures. The experimental results of these test structures are used to determine the layout dependence of the different charging mechanisms. A new charging yield model is developed and an effective methodology for the characterization of the layout dependence of charging damage of products is proposed, opening the way to charging robust design.

Finally, in *chapter 5*, the obtained results are generalized. It is shown that for better manufacturability it is necessary and possible to better integrate design, process and test development. This activity is generally known as Design for Manufacturability or DfM. It is described how a DfM methodology can be embedded in the development of ICs. Several experimental results in the design and manufacturing domains show the beneficial effect of DfM on yield. DfM enables a predictable, high yield, and ensures the expected time to market. Therefore, embedding DfM in the development of deep submicron technologies results in a clear strategic advantage.

# Chapter 7

## Samenvatting



# 7

## Samenvatting

Door de enorme kosten die met de fabricage van hedendaagse geïntegreerde schakelingen (IC's) gepaard gaan, is het van groot belang om met grote snelheid een hoge opbrengst te kunnen realiseren. Het begrip van fout mechanismen en de daarbij behorende opbrengst modellen is dan ook niet alleen van belang voor de planning van het productie volume, maar ook voor het snel op peil brengen van de opbrengst (yield ramping). Deze laatste activiteit is een erg complexe, en kan alleen tot een goed einde worden gebracht als de betrokken ingenieurs in staat zijn fout mechanismen snel te identificeren en, nog belangrijker, als ze in staat zijn de invloed van deze fout mechanismen op de opbrengst van producten te begrijpen. Alleen dan kunnen er prioriteiten worden gesteld ten aanzien van de vele mogelijke activiteiten die tot opbrengst verbetering leiden. Dit proefschrift beschrijft methodes die het mogelijk maken het verband tussen het fabricage proces, het product ontwerp en de opbrengst duidelijk te krijgen.

Modellen die fabricage opbrengst beschrijven spelen ook een belangrijke rol bij het ontwerpen van IC's. De experimenten die in dit werk worden beschreven laten zien dat het gebruik van deze modellen tijdens alle niveaus van ontwerp abstractie zelfs bepalend kunnen zijn voor het uiteindelijke financiële succes van een product. Tegenwoordig is de bereidheid voor ontwerpers om de opbrengst mee te nemen in hun toch al ingewikkelde ontwerp proces echter erg klein. Vaak wordt dit veroorzaakt door de inefficiëntie waarmee de informatie over de kwetsbaarheden van het productie proces wordt doorgegeven aan de ontwerp omgeving. Bovendien zijn de richtlijnen die gebruikt worden voor het ontwerp vaak een te simplistische representatie van de werkelijkheid, of ze kunnen niet worden ingebed in de gebruikte ontwerp algoritmes. Daarnaast zijn niet-robuste ontwerpen het resultaat van de aanname dat een de opbrengst puur fabricage aangelegenheid is. In dit proefschrift wordt op meerdere manieren aangetoond dat deze aanname niet langer kan worden gehandhaafd in hedendaagse en toekomstige IC technologieën. Om het voor een ontwerper mogelijk te maken om de invloed van de verschillende fout mechanismen op zijn ontwerp, en dus op de

opbrengst, te evalueren, is er een nieuwe methodologie nodig. In dit proefschrift wordt een dergelijke methodologie uiteengezet.

In hoofdstuk 2 wordt een overzicht gegeven van de verschillende opbrengst modellen die in de afgelopen decennia zijn ontwikkeld. De voor- en nadelen en de verschillende toepassingen van deze modellen worden besproken. Verder, wordt een grootschalig experiment beschreven waarin nieuwe opbrengst modellen worden ontwikkeld en geverifieerd. De nieuwe modellen gebruiken eenvoudig te bepalen ontwerp- en proces parameters en blijken waardevol te zijn voor zowel ontwerp- als fabricage van IC's.

In hoofdstuk 3 wordt aangetoond dat voor een waardevolle en nauwkeurige voorspelling van de opbrengst, naast de opbrengst modellen zelf, ook een nieuwe parameter extractie methodologie nodig is. De benadering die in dit proefschrift wordt voorgesteld is gebaseerd op teststructuren en speelt hierbij een belangrijke rol. Praktische toepassingen van deze methodologie voor zowel de ontwerp als productie omgeving worden beschreven.

Omdat opbrengst niet alleen een functie is van de fout mechanismen in het productie proces, maar ook van de gevoeligheid van het ontwerp, is de karakterisatie van IC ontwerpen van essentieel belang. Daarom worden in hoofdstuk 3 methodes om ontwerpen te karakteriseren beschreven. Ook de ontwikkeling, implementatie en het gebruik van een ontwerp karakterisatie systeem in een industriële omgeving worden beschreven. Verschillende experimenten laten grote verschillen ten aanzien van gevoeligheid zien tussen verschillende producten die in hetzelfde productie proces gemaakt worden. De mogelijkheid om product ontwerpen te kunnen karakteriseren blijkt van grote waarde te zijn bij het verklaren van product specifieke opbrengst verliezen.

In de voor dit werk uitgevoerde experimenten werd het herhaaldelijk duidelijk dat dezelfde rode draad gevolgd diende te worden om de opbrengst van IC's te kunnen voorspellen:

1. *Begrip en karakterisatie van de foutmechanismen* in termen van ontwerp gevoeligheid en statistische verdeling over de plak of lot. Zowel modellering, als test structuren kunnen hiervoor worden gebruikt.
2. *Inschatting van de invloed van het foutmechanisme op de opbrengst.* Vaak komt dit neer op de ontwikkeling van een opbrengst model, specifiek voor het fout mechanisme, en de ontwikkeling van een corresponderende ontwerp karakterisatie methode.

Als aan 1. en 2. is voldaan, kan de opbrengst voor elk ontwerp worden voorspeld.

3. *Het opbrengst model kan vervolgens gebruikt worden voor de ontwikkeling van een methodiek voor robuust ontwerp.*

Om de effectiviteit van bovenstaande, stapsgewijze aanpak te demonstreren werd deze toegepast op het ‘plasma schade’ fenomeen. In hoofdstuk 4 wordt experimenteel aangetoond dat in geavanceerde IC fabricage technologieën, plasma schade van invloed is op de opbrengst en betrouwbaarheid van circuits. Er wordt aangetoond dat de bestaande methodes voor robuust ontwerp te simplistisch zijn en niet in staat zijn de plasma schade problemen het hoofd te bieden. Alleen een benadering waarbij zowel het proces als het design worden meegenomen resulteert in een oplossing voor het ontwerp van producten die bestand zijn tegen de invloeden van plasma productie processen.

Daarom worden allereerst de verschillende opladings-mechanismen bestudeerd met behulp van speciaal ontwikkelde test structuren. De resultaten van deze structuren worden vervolgens gebruikt om de gevoeligheid van circuits voor dit mechanisme als functie van het ontwerp te bepalen. Een nieuw plasma schade opbrengst model wordt ontwikkeld en een effectieve manier voor de karakterisatie van de ontwerp afhankelijkheid van plasma schade wordt voorgesteld, waarmee plasma robuust ontwerp mogelijk is geworden.

Als laatste worden in hoofdstuk 5 de besproken resultaten gegeneraliseerd. Er wordt aangetoond dat voor betere ‘maakbaarheid’ van IC’s integratie van ontwerp, proces en test ontwikkelingen noodzakelijk is. Deze activiteit wordt ook wel aangeduid met DfM (Design for Manufacturability). In dit hoofdstuk wordt beschreven hoe een DfM methodologie ingebed kan worden in de ontwikkeling van geavanceerde IC’s. Verschillende experimenten laten een positief resultaat op de opbrengst zien. DfM maakt het mogelijk met hoge opbrengst te produceren en verzekert de beoogde ‘time to market’. Daarom biedt de implementatie van DfM in geavanceerde IC technologieën een duidelijk strategisch voordeel.

---

# List of abbreviations

ARDE	Aspect ratio dependent etch rate
CMP	Chemical mechanical polishing
CVD	Chemical vapor deposition
CVI	Charging vulnerability index
D <sub>0</sub>	Average process defect density
DA	Design attribute
DfM	Design for manufacturability
Dp	Product dependent defect density
DRC	Design rule check
FIB	Focused ion beam
FM	Failure mechanism
GDS	Layout file format
IC	Integrated circuit
IddQ	Quiescent (steady state) leakage current
IP	Intellectual property
LIL	Local interconnect
MAE	Manufacturability assessment environment
MAM	Multiplexed antenna monitoring test structure
MIMIC	Test structure mimicking a real product layout
OPC	Optical proximity correction
OTP	One time programmable embedded memory
PoF	Probability of failure
PPM	Parts per million (fail rate)
PVD	Physical vapor deposition
PYRA	Product yield risk assessment
RIE	Reactive ion etch
ROM	Read only memory
SEM	Scanning electron microscope
SRAM	Random access memory
STI	Shallow trench isolation
VLSI	Very large scale integration
YEM	Yield evaluation monitor

---

# List of publications

P. Simon, W. Maly, D. K. de Vries, E. Bruls, "*Design Dependency of Yield Loss Due to Tungsten Residues in Spin on Glass Based Planarization Process*", Proc. of ISSM, San Francisco, California, Nov. 1997.

W.Maly, H.T.Heineken, J.Khare, P.K. Nag, P.Simon and C.Ouyang, "*Design-Manufacturing Interface: Part II – Applications*", Design Automation and Test Conference, Europe, 1998 (DATE'98), pp.557-562, Paris, France, Feb.23-26, 1998.

J.R.M. Luchies, P.Simon, F. Kuper and W. Maly, "*Relation between Product Yield and Plasma Process induced Damage*", Proceedings of the P2ID conference, Honolulu Hawaii, pp. 7-10, June 1998.

P.Simon, W.Maly, J.R.M.Luchies and R.Antheunis, "*Multiplexed Antenna Monitoring Test structure*", Proceedings of the P2ID conference, Honolulu Hawaii, pp.205-208, June 1998.

P. Simon and W. Maly, "*Identification of Plasma Induced Damage Conditions in VLSI Designs*", proc. ICMTS, Stockholm, March 1999, pp.1-6.

P. Simon, et al. "*Layout based manufacturability assessment and yield prediction methodology*", Conference on In-line Characterization techniques for performance and Yield Enhancement in Microelectronic Manufacturing, Edinburgh, Scotland, May 1999, pp.282-288

P. Simon, J.R.M.Luchies and W. Maly, "*Antenna Ratio Definition for VLSI Circuits*", Proceedings of the P2ID conference, Monterey California, pp.16-20, May 1999.

P. Simon, J.R.M. Luchies and W. Maly, "*Identification of Plasma Induced Damage Conditions in VLSI Designs*", IEEE Transactions on Semiconductor Manufacturing, VOL. 13, NO.2, May 2000.

Y. Fei, P. Simon and W. Maly "*New Yield Models for DSM Manufacturing*", IEDM 2001.

---

# About the Author

Paul Simon was born in Voorschoten, The Netherlands. He received his M.Sc. degree in Electrical Engineering from the Technical University of Delft, The Netherlands. His master's thesis deals with auto-calibration of silicon Hall plate magnetic sensors.

In 1995 he joined the Product Engineering Department of MOS4YOU wafer fab of Philips Semiconductors in Nijmegen, The Netherlands. His work involved yield modeling and development of yield ramping methodologies for advanced CMOS technologies.

At MOS4YOU he also led the Design for Manufacturability (DfM) Group and, since 1996, he carried out a Ph.D. study on DfM and yield modeling for deep sub-micron design.

He has published several papers on auto-calibration of silicon Hall plates, yield modeling, plasma charging damage and DfM.

Since April, 2001, Paul is working on 0.13 and 0.10  $\mu\text{m}$  CMOS technologies as a yield & industrialization engineering manager at the 300mm wafer fab of Philips Semiconductors and ST Microelectronics in Crolles, France.

## Stellingen

Behorende bij het proefschrift

# Yield Modeling for Deep Sub-Micron IC Design

Paul Simon

La Terrasse, 20 December 2001

1. Door de steeds strengere time to market eisen die aan basis cellen zoals SRAMs worden gesteld, zal de behoefte aan complexe simulatiemodellen en de daarbij behorende parameters in een vroeg stadium van de ontwikkeling afnemen. Het zal belangrijker worden om in staat te zijn robuust te ontwerpen met eenvoudige modellen en gebrekkige model parameter sets.
2. Het 'per default' toepassen van minimale afmetingen in IC ontwerpen getuigt van de verregaande ontkoppeling van het IC ontwerp process en de fabricage.
3. Voor een beter begrip van de opbrengst van IC's zijn er geen nieuwe yield modellen meer nodig, maar eerder methodes om model parameters te karakteriseren. [dit proefschrift, hfst. 2,3]
4. Voor een betere controle over de opbrengst van IC's zou het beter zijn als de vertaling van een IC ontwerp van een hoger abstractieniveau naar layout (gedeeltelijk) in de fabriek zou plaatsvinden.
5. De effectiviteit van diodes voor het beperken van het verlies in opbrengst door plasma schade wordt overschat.  
[dit proefschrift, hfst. 4]
6. Het concept 'yield' zou tot een van de 'key performance indicators' moeten behoren van IC ontwerpsoftware.  
[dit proefschrift, hfst. 5]
7. De ontwikkeling van test structuren voor 'yield ramping' verdient meer aandacht.
8. Zonder deviatie van het te doen gebruikelijke is vooruitgang onmogelijk.
9. Voor een gelukkig leven is het hebben van heldere doelen noodzakelijk.
10. Tijdens een conflict is het toppunt van arrogantie als de behoefte aan vergelding sterker is dan de behoefte aan begrip.
11. Het ongevraagd geven van 'goed bedoelde' raad geeft geen blijk van respect.
12. Ondanks het feit dat een platte organisatiestructuur flexibeler lijkt, is een hiërarchische organisatie in staat sneller te reageren op externe veranderingen.
13. Het in dezelfde periode combineren van een baan, een promotie onderzoek en het krijgen van kinderen vertoont sterke overeenkomst met spitsroeden lopen, jongleren en vuur spuwen tegelijk: alleen het op juiste wijze stellen van prioriteiten levert een bevredigend resultaat op voor minstens een van de drie.