RELIABILITY WEAROUT MECHANISMS IN ADVANCED CMOS TECHNOLOGIES

Alvin W. Strong
Ernest Y. Wu
Rolf-Peter Vollertsen
Jordi Suñé
Giuseppe La Rosa
Stewart E. Rauch, III
Timothy D. Sullivan

IEEE Solid-State Circuits Society, *Sponsor*
IEEE Press Series on Microelectronic Systems
Stuart K. Tewksbury and Joe E. Brewer, *Series Editors*

A JOHN WILEY & SONS, INC., PUBLICATION

RELIABILITY WEAROUT MECHANISMS IN ADVANCED CMOS TECHNOLOGIES

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Lajos Hanzo, *Editor in Chief*

R. Abari	T. Chen	B. M. Hammerli
J. Anderson	T. G. Croda	O. Malik
S. Basu	M. El-Hawary	S. Nahavandi
A. Chatterjee	S. Farshchi	W. Reeve

Kenneth Moore, *Director of IEEE Book and Information Services (BIS)*

IEEE Solid-State Circuits Society, *Sponsor*

Technical Reviewers

Paul Ho, University of Texas
Bill Knowlton, Boise State University
Pat Lenahan, Penn State University

RELIABILITY WEAROUT MECHANISMS IN ADVANCED CMOS TECHNOLOGIES

Alvin W. Strong
Ernest Y. Wu
Rolf-Peter Vollertsen
Jordi Suñé
Giuseppe La Rosa
Stewart E. Rauch, III
Timothy D. Sullivan

IEEE Solid-State Circuits Society, *Sponsor*
IEEE Press Series on Microelectronic Systems
Stuart K. Tewksbury and Joe E. Brewer, *Series Editors*

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2009 by the Institute of Electrical and Electronics Engineers, Inc.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey. All rights reserved.
Published simultaneously in Canada

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, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available:

ISBN 978-0471-73172-6

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface xiii

1 INTRODUCTION 1

Alvin W. Strong

1.1 Book Philosophy 1
1.2 Lifetime and Acceleration Concepts 2
1.2.1 Reliability Purpose 2
1.2.2 Accelerated Life 3
1.2.3 Accelerating Condition 8
1.3 Mechanism Types 9
1.3.1 Parametric or Deterministic Mechanisms 9
1.3.2 Structural Mechanisms 11
1.3.3 Statistical Mechanisms 12
1.3.4 Infant Defects 14
1.3.5 Operating Life Defects 15
1.3.6 Wearout 16
1.4 Reliability Statistics 16
1.4.1 Introduction 16
1.4.2 Assumptions 16
1.4.3 Sampling and Variability 18
1.4.4 Criteria, Censoring, and Plotting Points 22
1.4.5 Definitions (Normal) 25
1.4.6 Exponential Distribution 38
1.4.7 Smallest Extreme Value and Weibull Distributions 41
1.4.8 Lognormal Distribution 45
1.4.9 Poisson Distribution 49
1.5 Chi-Square and Student t Distributions 52
1.5.1 Gamma and Chi-Square Distributions 52
1.5.2 Student t Distribution 53

1.6 Application 54
1.6.1 Readouts Versus "Exact" Time-To-Fail 54
1.6.2 Additional Types of Censoring 55
1.6.3 Least-Squares Fit and Application 56
1.6.4 Chi-Square Goodness of Fit Application 64
1.6.5 Maximum Likelihood Estimation (MLE) 66
1.6.6 Closure 66
References 67

2 DIELECTRIC CHARACTERIZATION AND RELIABILITY METHODOLOGY 71
Ernest Y. Wu, Rolf-Peter Vollertsen, and Jordi Suñé

2.1 Introduction 71
2.1.1 Application and Fabrication of Silicon Dioxide-Based Dielectrics 73
2.1.2 Failure Modes of Gate Oxide and Reliability Requirements 76
2.1.3 Impact of Oxide Scaling 81
2.2 Fundamentals of Insulator Physics and Characterization 85
2.2.1 Capacitance–Voltage Characteristics 85
2.2.2 Carrier Tunnelling and Injection Mechanisms in MOS Structures 95
2.2.3 Oxide Voltage (Field) and Electron Energy at Anode 110
2.2.4 Determination of Oxide Thickness 115
2.3 Measurement of Dielectric Reliability 124
2.3.1 Measurement Methods 124
2.3.2 Designs of Stress and Test Structures 144
2.3.3 Physical Observations of Dielectric Breakdown 148
2.3.4 Considerations for Oxide Breakdown Detection 151
2.4 Fundamentals of Dielectric Breakdown Statistics 154
2.4.1 Weibull Function and Poisson Statistics 154
2.4.2 Area Transformation 160
2.4.3 Weibull Versus Lognormal Failure Distributions 166
2.4.4 Estimation of Weibull Parameters 168
2.4.5 Methods for Determination of the Weibull Shape Factor (Slope) 177

2.4.6 Modeling Bi- or Multimodal Weibull Distributions 178
2.5 Summary and Future Trends 192
References 193

3 DIELECTRIC BREAKDOWN OF GATE OXIDES: PHYSICS AND EXPERIMENTS 209
Ernest Y. Wu, Rolf-Peter Vollertsen, and Jordi Suñé

3.1 Introduction 209
3.2 Physics of Degradation and Breakdown 210
3.2.1 Oxide Degradation 210
3.2.2 The Role of Electric Field and Carrier Energy in the Degradation and Breakdown of Gate Oxides 218
3.2.3 Percolation Model for the Breakdown Statistics 222
3.2.4 Three-Dimensional Analytic Model for Oxide Breakdown Statistics 223
3.2.5 Thickness Dependence of Oxide Breakdown 226
3.3 Physical Models for Oxide Degradation and Breakdown 231
3.3.1 The Thermochemical Model 232
3.3.2 Hole-Induced Breakdown Models 236
3.3.3 Anode Hydrogen Release Model 249
3.4 Experimental Results of Oxide Breakdown 259
3.4.1 Voltage Dependence 260
3.4.2 Temperature Dependence 272
3.4.3 Interrelationship of Voltage and Temperature Dependence 276
3.4.4 Polarity Dependence 280
3.4.5 Degradation and Breakdown Under AC Stress Conditions 284
3.4.6 Gate Oxide Reliability Projection 287
3.5 Post-Breakdown Phenomena 295
3.5.1 Review of Post-Breakdown Experimental Observations 296
3.5.2 Modeling the Post-Breakdown Statistics 306
References 314

4 NEGATIVE BIAS TEMPERATURE INSTABILITIES IN pMOSFET DEVICES **331**
Giuseppe LaRosa

4.1 Introduction 331
4.2 Considerations on NBTI Stress Configurations 333
4.3 Appropriate NBTI Stress Bias Dependence 335
4.4 Nature of the NBTI Damage 339
4.5 Impact of the NBTI Damage to Key pMOSFET Transistor Parameters 341
4.5.1 Impact of the NBTI Damage to pMOSFET Physical Parameters 342
4.5.2 Relation Between Key Physical and Electrical pMOSFET Parameters 344
4.6 Physical Mechanisms Contributing to the NBTI Damage 350
4.6.1 Interface Traps Generation 351
4.6.2 Positive Charge Defects Generation/Activation 355
4.7 Key Experimental Observations on the NBTI Damage 363
4.7.1 pMOSFET NBTI: Worst MOSFET Bias Temperature Condition 363
4.7.2 Role of Cold Holes 366
4.7.3 Dependence on Gate Oxide Electric Field 368
4.7.4 Dependence on Stress Temperature 372
4.7.5 Time Evolution 375
4.7.6 Recovery Phenomena 378
4.7.7 Impact of NBTI Recovery to NBTI Stressing/Testing Methodologies 383
4.7.8 Dynamic NBTI 390
4.8 N_{it} Generation by Reaction–Diffusion (R–D) Processes 395
4.8.1 Modeling of N_{it} Generation by Reaction–Diffusion Processes 397
4.8.2 R–D Kinetics Controlled By an Arrhenius Diffusion Process 401
4.8.3 R–D Kinetics Controlled By a Dispersive Diffusion in SiO_2 407
4.8.4 N_{it} Repassivation Phase 409
4.8.5 Dynamic NBTI 411
4.9 Hole Trapping Modeling 412
4.10 NBTI Dependence on CMOS Processes 417

4.10.1 Hydrogen Species 417
4.10.2 Nitrogen 419
4.10.3 Fluorine 422
4.10.4 Boron 424
4.10.5 NBTI Sensitivity to BEOL Charging 426
4.11 NBTI Dependence on Area Scaling 428
4.12 Overview of Key NBTI Features 431
References 434

5 HOT CARRIERS 441
Stewart E. Rauch, III

5.1 Introduction 441
5.2 Hot Carriers: Physical Generation and Injection Mechanisms 443
5.2.1 Electric Field in a MOSFET Pinch-Off Region 444
5.2.2 Lateral Electric Field in the Pinch-Off Region of an LDD MOSFET 450
5.2.3 Vertical Electric Field in the Pinched-Off Channel Region 453
5.2.4 High Field-Induced Carrier Heating and the Carrier Energy Distribution Function 454
5.2.5 Impact Ionization Phenomena 464
5.2.6 Primary Impact Ionization (1II) in a MOSFET in Saturation 469
5.2.7 Channel Hot Carrier Injection Mechanisms 472
5.2.8 Lucky Electron Model 476
5.2.9 Bulk Current at Low V_{dd} ($V_{dd} \leq$ or just above E_G/q): The Cross-Over Effect 481
5.2.10 Secondary Impact Ionization (2II) 481
5.2.11 Limits of the Lucky Electron Model 485
5.2.12 The Energy-Driven Model 486
5.2.13 Localized Self-Heating Effects 489
5.3 Hot Carrier Damage Mechanisms 491
5.3.1 Introduction 491
5.3.2 Interface States Generation 494
5.3.3 The Giant Isotope Effect 498

5.4 HC Impact to MOSFET Characteristics 499
5.4.1 I_d–V_{gs} Shifts 499
5.4.2 Localization of Channel Hot Carrier Damage in the Drain Region 505
5.4.3 CHC-induced Increase in Parasitic Drain Series Resistance 506
5.5 Hot Carrier Shift Models 508
5.5.1 Note About "Device Lifetime" 508
5.5.2 Lucky Electron Model–Peak I_{sx} CCHC 509
5.5.3 The Electron-Effective Temperature Model 510
5.5.4 Energy-Driven Model: nMOSFET CCHC 511
5.5.5 Modeling of PFET Conducting Channel Hot Carrier 513
References 514

6 STRESS-INDUCED VOIDING 517
Timothy D. Sullivan

6.1 Introduction 517
6.1.1 Overview 517
6.1.2 Conceptual Basis 519
6.2 Theory and Model 524
6.2.1 Relating Time-to-Failure to Void Size 530
6.2.2 Stress Contribution 533
6.2.3 Accelerated Testing for Stress Voiding 535
6.2.4 Alloying and Impurity Effects 538
6.2.5 Relating Resistance Changes to Void Growth 541
6.2.6 Factors that Complicate SV Data 543
6.3 Role of the Overlying Dielectric 546
6.4 Summary of Voiding in Al Metallizations 549
6.5 Stress Voiding in Cu Interconnects 550
6.5.1 Microstructure of Cu 554
6.5.2 Role of Dielectrics in Cu Voiding 555
6.5.3 Microstructural Effects 556
6.5.4 Structural Influences 558
6.5.5 Structures and Models for Cu Voiding 560
6.6 Concluding Remarks 562
References 563

7 ELECTROMIGRATION **565**

Timothy D. Sullivan

7.1 Introduction 565
7.2 Metallization Failure 566
7.3 Electromigration 567
7.3.1 Single-Component Metallization 571
7.3.2 Layered Metallizations 574
7.3.3 Short-Length Effect 578
7.3.4 Multilevel Metallizations 580
7.3.5 Incubation Time 584
7.3.6 Electromigration in Cu Lines 586
7.3.7 Electromigration Testing 588
7.4 General Approach to Electromigration Reliability 589
7.4.1 Projection Methodology Statistical Considerations 592
7.4.2 Normal Distribution 594
7.4.3 Lognormal Distribution 597
7.4.4 Bimodal Distributions 603
7.4.5 Three-Parameter Lognormal Fit to a Single Distribution 605
7.5 Thermal Considerations for Electromigration 607
7.5.1 Self-Heating 607
7.5.2 Wafer-Level Electromigration 614
7.5.3 Indirect Joule Heating 616
7.6 Closing Remarks 617
References 617

Index **619**

PREFACE

As consumers today we care a great deal about the useful life of a car, a computer, or a pacemaker. Folks of an earlier century were concerned about the life of their carriages and we will return to this later. Manufacturers have always needed to understand the length of the useful life of their product. In addition to potential warranty costs, there are customer satisfaction issues and there may even be liability and ethical ramifications. Reliability, then, is justifiably of great concern to all. It must answer the questions: What is the useful lifetime of a given product and how can one verify that lifetime? In a CMOS world, the operating criteria for products can be very different. Some are used rarely and so may require only a few minutes of useful life. Singing greeting cards and yodeling stuffed animals would fall into this category. Other CMOS products, the pacemaker for example, would be an unmitigated disaster if it failed within minutes of implantation. CMOS applications in space can have both very long expected operational lifetimes as well as very severe environmental conditions. The wide range of CMOS applicability makes the reliability questions all the more interesting, and challenging.

The purpose of this book is to present in one place the physics, stress and analysis techniques, and models necessary to correctly determine the time of wearout for the major reliability mechanisms associated with CMOS technology. Given these tools and the product application and specifications, the engineer will be able to accurately predict when wearout will start to occur for any given product application. The engineer will also be able to verify that there are no surprises lurking to cause an early failure. Most previous books covering CMOS reliability have focused on only one or two of the technology mechanisms, have generally been more product application specific, and have not delved as deeply into the physics of the mechanisms as we have done here.

Reliability Wearout Mechanisms in Advanced CMOS Technologies has been written at a beginning graduate, or senior undergraduate, level and assumes some solid state physics background. It is designed to teach the physics of the major CMOS reliability mechanisms, the impact those mechanisms have on the device and circuits, and how to calculate that impact. The book assumes that the engineer has little or no reliability education or experience. The engineer that masters this book should have a good understanding of CMOS reliability physics, be able to design and conduct appropriate reliability experiments, analyze data, and accurately make the resultant reliability projections.

The book is divided into seven relatively independent chapters. The first chapter, the Introduction, discusses the assumptions necessary for any reliability work, describes the various types of CMOS reliability mechanisms, and at a very basic level introduces the statistics necessary for CMOS reliability analysis. The

second chapter, Gate Dielectric Characterization and Methodology, describes the techniques available to understand the properties of the dielectric in question and discusses the failure distribution models for that dielectric. The third chapter, the Dielectric Breakdown of Gate oxides, describes the physics behind dielectric breakdown and provides the experimental and analytical steps required to perform dielectric breakdown projections. The fourth chapter, the Negative Bias Temperature Instability, introduces transistor device behavior as it impacts device reliability; discusses the device configurations and process variations that are important to the effect; describes the physics that controls the degradation mechanism; and finally gives experimental procedures to measure, analyze, and project NBTI lifetimes. The fifth chapter, Hot Carriers, describes the sensitive operational configurations for the HC effect and why they are important, discusses the physics and models of the effect, how the effect degrades the transistor performance, and describes how to measure HC degradation and project device lifetimes that are limited by that degradation. The sixth chapter, Stress-Induced Voiding, moves to the interconnect levels and introduces the theory and models that apply when a constrained system undergoes temperature changes; describes the impact of those temperature changes to the metal layers, vias, and their interfaces; and presents the analysis techniques to estimate the lifetime of the metallurgy. The seventh and last chapter, Electromigration, addresses the lifetime of current-carrying metal components; discusses the physics of failure for those components; and gives the experimental procedures, models and analysis techniques for projecting lifetimes limited by electromigration.

Because of the depth and breadth of CMOS technology reliability itself, we do not discuss electrostatic discharge, latchup, radiation-induced soft error rates, package reliability, or the reliability of the package and chip interactions.

The following poem returns us to those folks of that earlier century who were also interested in reliability — the reliability of a carriage, as recorded by Oliver Wendell Holmes.

Alvin W. Strong

Essex Junction, Vermont
July 2009

The Deacon's Masterpiece
or
The Wonderful "One-Hoss Shay"

Have you heard of the wonderful one-hoss shay,
That was built in such a logical way
It ran a hundred years to a day,
And then of a sudden it — ah, but stay,
I'll tell you what happened without delay,
Scaring the parson into fits,
Frightening people out of their wits, –
Have you ever heard of that, I say?

Seventeen hundred and fifty-five.
Georgius Secundus was then alive,
Snuffy old drone from the German hive.
That was the year when Lisbon-town
Saw the earth open and gulp her down,
And Braddock's army was done so brown,
Left without a scalp to its crown.
It was on that terrible Earthquake-day
That the Deacon finished the one-hoss shay.

Now in building of shaises, I tell you what,
There is always a weakest spot, –
In hub, tire, felloe, in spring or thill,
In panel or crossbar, or floor, or sill,
In screw, bolt, throughbrace, — lurking still,
Find it somewhere you must and will, –
Above or below, or within or without, –

And that's the reason, beyond a doubt,
That a chaise breaks down, but doesn't wear out.

But the Deacon swore (as deacons do,
With an "I dew vum," or an "I tell yeou")
He would build one shay to beat the taown
'N' the keounty N all the kentry raoun';
It should be so built that it couldn' break daown:
"Fer," said the Deacon, "t's mighty plain
Thut the weakes' place mus' stan' the strain;
N the way t' fix it, uz I maintain, is only jest
'T' make that place uz strong uz the rest."

So the Deacon inquired of the village folk
Where he could find the strongest oak,
That couldn't be split nor bent nor broke, –
That was for spokes and floor and sills;
He sent for lancewood to make the thills;
The crossbars were ash, from the the straightest trees
The panels of whitewood, that cuts like cheese,
But lasts like iron for things like these;
The hubs of logs from the "Settler's ellum," –
Last of its timber, — they couldn't sell 'em,
Never no axe had seen their chips,
And the wedges flew from between their lips,
Their blunt ends frizzled like celery-tips;
Step and prop-iron, bolt and screw,
Spring, tire, axle, and linchpin too,
Steel of the finest, bright and blue;
Throughbrace bison-skin, thick and wide;
Boot, top, dasher, from tough old hide
Found in the pit when the tanner died.
That was the way he "put her through,"
"There!" said the Deacon, "naow she'll dew!"

Do! I tell you, I rather guess
She was a wonder, and nothing less!
Colts grew horses, beards turned gray,
Deacon and deaconess dropped away,
Children and grandchildren — where were they?
But there stood the stout old one-hoss shay
As fresh as on Lisbon-earthquake-day!

EIGHTEEN HUNDRED; — it came and found
The Deacon's masterpiece strong and sound.
Eighteen hundred increased by ten; –
"Hahnsum kerridge" they called it then.
Eighteen hundred and twenty came; –
Running as usual; much the same.
Thirty and forty at last arive,
And then come fifty and FIFTY-FIVE.

Little of of all we value here
Wakes on the morn of its hundredth year
Without both feeling and looking queer.
In fact, there's nothing that keeps its youth,
So far as I know, but a tree and truth.
(This is a moral that runs at large;
Take it. — You're welcome. — No extra charge.)

FIRST OF NOVEMBER, — the Earthquake-day, –
There are traces of age in the one-hoss shay,
A general flavor of mild decay,
But nothing local, as one may say.
There couldn't be, — for the Deacon's art
Had made it so like in every part
That there wasn't a chance for one to start.
For the wheels were just as strong as the thills
And the floor was just as strong as the sills,
And the panels just as strong as the floor,
And the whippletree neither less or more,
And the back-crossbar as strong as the fore,
And the spring and axle and hub encore.
And yet, as a whole, it is past a doubt
In another hour it will be worn out!
First of November, fifty-five!
This morning the parson takes a drive.
Now, small boys get out of the way!
Here comes the wonderful one-hoss shay,
Drawn by a rat-tailed, ewe-necked bay.
"Huddup!" said the parson. — Off went they.

The parson was working his Sunday's text, –
Had got to fifthly, and stopped perplexed
At what the — Moses — was coming next.
All at once the horse stood still,
Close by the meet'n'-house on the hill.
First a shiver, and then a thrill,
Then something decidedly like a spill, –
And the parson was sitting upon a rock,
At half past nine by the meet'n'-house clock, –
Just the hour of the earthquake shock!

What do you think the parson found,
When he got up and stared around?
The poor old chaise in a heap or mound,
As if it had been to the mill and ground!
You see, of course, if you're not a dunce,
How it went to pieces all at once, –
All at once, and nothing first, –
Just as bubbles do when they burst.

End of the wonderful one-hoss shay.
Logic is logic. That's all I say.

1

INTRODUCTION

Alvin W. Strong

1.1 BOOK PHILOSOPHY

This CMOS technology reliability book has been written at a beginning graduate level or senior undergraduate level and assumes some solid state physics background.

The book is divided into seven relatively independent chapters consisting of an introduction, gate dielectric characterization, gate dielectric physics and breakdown, negative bias temperature instability or just NBTI reliability, hot carrier injection or hot electron reliability, stress-induced voiding or stress migration reliability, and electromigration reliability. The chapters describe the reliability mechanisms and the physics associated with them. They then take that understanding as the framework to build the bridge between the accelerated mechanism and the product mechanism.

For a CMOS reliability course or understanding focused only on one of the mechanisms, the authors expect that the material covered would include most of the first chapter and that focus chapter.

Several mechanisms are occasionally considered with reliability mechanisms, but these are not included here. Examples of these include latch-up [1], electrostatic discharge (ESD) [2, 3], and the radiation-induced soft-error rate (SER) [4].

Reliability Wearout Mechanisms in Advanced CMOS Technologies. By Strong, Wu, Vollertsen, Suñé, LaRosa, Rauch, and Sullivan
Copyright © 2009 the Institute of Electrical and Electronics Engineers, Inc.

1.2 LIFETIME AND ACCELERATION CONCEPTS

It is a fact of life that every human-devised system has a finite lifetime before the catastrophic failure of the system occurs. However, most systems have a reasonably well-defined lifetime, and the catastrophic failure, or wearout, occurs well past that expected lifetime. That system has met our expectation and the customer is satisfied. Wearout is best thought of in terms of all of the systems or subsystems failing within one or two orders of magnitude in time. For example, a computer system with an expected lifetime of 10 years should experience no significant wearout before 10 years. However, all of the systems could be expected to wear out sometime between 20-plus years and 200-plus years.

1.2.1 Reliability Purpose

The purpose then of reliability is to ensure that the life of the system will be longer than the target life and that the failure rate during the normal operating life of the system will be below the target failure rate. The reliability of the product must be known when the product is sold so that the operating-life warranty costs can be quantified and customer satisfaction protected. Ensuring these objectives are met means that each failure mechanism must be quantified so that its impact during normal operating life can be predicted and the time at which it starts to cause the system to wear out can be predicted as well.

The length of time one has to do the reliability stressing and make the predictions is dependent on the state of the program. As a new technology is being developed, reliability engineers should be generating reliability data to help guide the program in the appropriate design, cost, and reliability tradeoffs. This work may occur over the course of several months to a few years. However, feedback on any given experiment needs to be given as quickly as possible. Once the technology is ready for implementation, it would typically undergo a "qualification" of no more than three months in duration. If a problem is discovered after qualification, that is, during manufacturing, it is all the more crucial to give feedback quickly.

The concept of an accelerated life is necessary for reliability stressing to have meaning. That is, it must be possible to find some condition or conditions that will allow one to shrink a 10-year product life down to a three month period, or less, so that the reliability of the system can be investigated and guaranteed in that three months. The conditions used to accelerate a given mechanism usually cannot be applied to the whole system (in our case, the semiconductor product chip). In this case, a test structure must typically be built that will replicate the behavior of the element in the product chip, but allow one to apply an accelerating condition. Hence, with the concept of accelerated life, we also need to posit the concept of a representative test structure.

It must be noted that in all of the above discussion, the product is assumed to operate perfectly when it is first turned on, for example, at time zero.

Once the reliability of each element has been investigated, understood, and modeled, an additional step should be feedback for the next design pass so that the

product team can design in reliability. A simple example of this design for reliability would be to use minimal groundrules only where necessary.

1.2.2 Accelerated Life

An accelerated life concept must include several features to be useful. In addition to the requirement that it can be used to accelerate a particular reliability mechanism, it must be possible to quantify how much that condition actually accelerates the reliability mechanism. It must be possible to build a bridge between the accelerated stress conditions and the use condition so that it is possible to quantify the degree or amount of acceleration. This quantification is also necessary to ensure that no mechanism is introduced with the accelerating condition that does not exist at the use condition of the product. One must understand if the behavior of the mechanism is uniform and consistent from the use condition to the accelerating conditions.

Once an appropriate accelerating condition has been determined, the acceleration between the stress conditions and the use condition can be determined. First, data from at least two different values of accelerating conditions are measured. The cumulative fails from those conditions are then plotted on the y axis, versus time on the x axis. This plot is done on a set of axes that have been transformed in such a way that the resulting plot is a straight line. The methodology for doing this for the most common distributions used in semiconductor reliability is discussed in detail later in this chapter. The two distributions that are most commonly used in semiconductor technology reliability are the two-parameter lognormal and Weibull distributions because each of these is very flexible and can be used to describe many different types of behaviors. These distributions provide a functional form with which a distribution can be characterized so that it can then be treated analytically. The details of these distributions, and their axes transformations, are the topics of Section 1.4. A simple, although somewhat unrealistic, example would be a distribution whose cumulative failures were linear with the log of time. If the cumulative fails were to be plotted against time, a nonlinear curve would result. However, a transformation of the x axis by taking the log of the time and then plotting the cumulative failures against that log of time would result in a straight line. These transformations are necessary so that the distributions and the slopes remain invariant across all accelerating conditions and down to the use conditions. All transformations have been made for the example in Figure 1.1 so that the plot is linear. Two sets of voltage data are shown plotted on the top left of Figure 1.1. At least three different values of each accelerating condition are preferred but only two are shown on Figure 1.1 for simplicity. The data from these accelerated conditions are used to calculate an acceleration factor, which is in turn used to calculate the acceleration time between the lowest accelerated condition and the use condition.

One uses the lowest accelerated condition to minimize projection error. Obviously if any new mechanism is introduced due to the accelerated condition, or a nonlinearity in the expected mechanism is introduced, this acceleration time

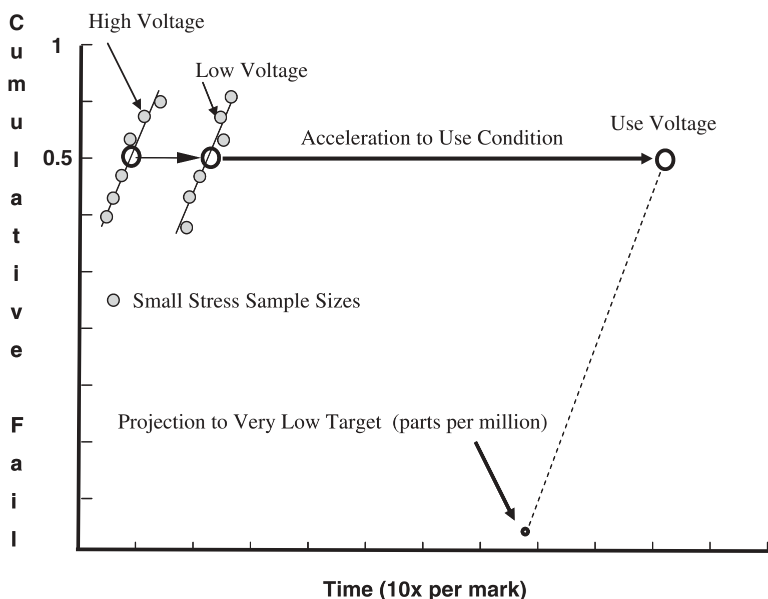

Figure 1.1. Acceleration example with two voltage conditions with the first projection to use condition at a 50% life and the final projection to low (ppm) failure rate target.

would not be valid. The transformation equations for the x and y axes of the plot will depend on the particular reliability mechanism in question and the probability distribution function that is most appropriate for that mechanism. The sample sizes for the stresses are typically very small, whereas the wearout target is typically expressed in terms of parts per million (ppm) or less. This means that once the acceleration between the stress conditions and the use condition is calculated, a second projection must then be made from that value, which typically is at about the 50% fallout point of each of the accelerated curves, to a very small percent fallout for that second projection curve. Note that the 50% value is used only as a convenient example here. The exact value at which the acceleration calculations will be made will depend on the distribution that is to be used and is discussed in detail later in this chapter. The intent here is to give a broad overview and avoid losing the reader in the detail. This extrapolation is done using the same slope found during at the stress conditions, since the axes have been transformed so that they remain invariant across all of the conditions of interest. Thus, an error in the acceleration factor causes the use condition to be incorrectly located in time, and any errors in determining the correct slope causes the projection to the small fraction fail target to have an additional error. Often this last error, the error due to an incorrect slope determination, can cause the largest error in the resultant projection. It should be highlighted that we are not speaking of graphing errors here since the calculations can all be performed using

computer software. If done graphically, those errors would be in addition to the errors mentioned previously.

Three plots are shown in Figure 1.1. The two plots on the left show the two different accelerated voltage conditions with all other conditions held constant. The dotted line plot on the right shows the projection from the 50% failure time for the use condition to the failure time associated with the target failure rate for that mechanism. One other potential factor that is not shown in Figure 1.1 is a test-structure scaling factor. This factor will be discussed in each of the chapters for which it is applicable. Although there are many accelerating conditions as shown in Section 1.2.3, by far the two most common accelerating conditions are voltage and temperature. The minimum experimental design that can yield a voltage acceleration factor is the two conditions as shown in Figure 1.1. For example, assume that the stress voltages in Figure 1.1 are 4.37 V and 4 V. The lines on the left side of Figure 1.1 represent fits to data taken at these accelerated voltage conditions. The slopes of the two stress conditions are shown as equal. The slope fit would normally be accomplished by a fitting program, which could force the best fit to all of the accelerated curves simultaneously. For this case we will assume an Eyring acceleration model [5] applies that has the form $Acc = t_2/t_1 = \exp\{(\Delta H/k)\{(1/T_2) - (1/T_1)\}\} \exp\{-\beta_V (V_2 - V_1)\}$. For the acceleration due to voltage, $Acc_{VOLTstress} = t_2/t_1 = \exp\{-\beta_V(V_2 - V_1)\}$, where V is voltage, t is time, and β_V is the voltage acceleration factor for this Eyring model. The temperature acceleration model by itself has the form $Acc_{TEMPstress} = t_2/t_1 = \exp\{(\Delta H/k)\{(1/T_2) - (1/T_1)\}\}$ where k is Boltzmann's constant and is also known as an Arrhenius model. These models will be discussed in more detail throughout this book but are introduced here to give the reader an early qualitative introduction to the acceleration concepts. Observation of the first two curves will reveal a time difference or acceleration of about $30\times$. The large circles in Figure 1.1 represent a mean life of the hardware under stress and as mentioned above are used as a convenient example. As will be discussed later, the points at which the acceleration calculations will be made are the most accurate values for the distribution under consideration. If the voltage used for the first curve on the left is $V = 4.37$ and the voltage for the second curve is $V = 4$, then β_V may be calculated, given $Acc_{VOLTstress} = 30$, as $\beta_V = (\ln Acc_{VOLTstress})/(V_1 - V_2) = 9.2$. This value for β_V is then used to project from $V = 4$ to the $V = 2$ use condition as $Acc_{VOLTuse} = \exp\{9.2 \times (4-2)\} = 10^8$. Having made this calculation, one then needs to consider whether or not the value calculated is reasonable based on comparable data both from the reliability analyst's prior work, as well as literature values. A similar procedure would be used to calculate any acceleration including, for example, a temperature acceleration. This example should give the reader a better understanding of the actual process of stressing and then projecting to use conditions using acceleration concepts. Obviously the stress conditions must be appropriately chosen and the experiment appropriately designed to achieve useful results. Note that if too small of difference is used between two accelerating conditions then the experimental error and the statistical variation in the two sets of data may cause enough overlap of data such that the acceleration factor between the two sets of data cannot be calculated. On the other hand if the difference between

the two sets of data is too large, the failure times of the lower condition may be longer than the time designated for the stress. The discussion of the extrapolation to very small failing percentage targets will commence in Section 1.4.5.

We now return to a more general discussion of acceleration. The mean life from Figure 1.1 is plotted against one of the accelerating conditions in Figure 1.2. Figure 1.2 presents a picture of the progress of the state of the art of reliability stressing across the last 20-plus years. Each circle represents a change of approximately 40 × in time.

More than 20 years ago, all reliability stressing was done with the reliability test structures wire-bonded onto a die carrier or module. This structure, which was contained within the package, was then put into a stress apparatus, which typically applied stress temperature and voltage between weeks and months, depending on the mechanism under investigation. The readouts were made at preset values, typically on the order of two or three times per decade. The test structures had to be removed from the stress apparatus and physically transported to a tester for each readout. This stressing is represented in Figure 1.2 by the second circle from the left. Note that the left most circle represents the useful life of the structure, typically 10 years. For mechanisms like ionic contamination, which will relax unless the voltage is continuously applied, it was necessary to have large batteries connected to keep the hardware at stress voltage while transporting the hardware

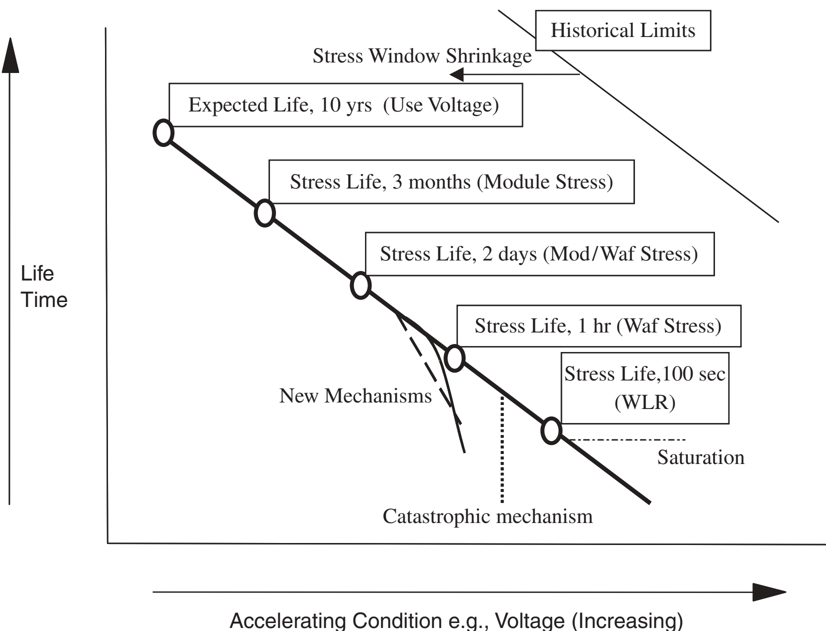

Figure 1.2. Lifetime projection curve with each large circle representing the 50% life for the accelerating or use condition and showing various nonlinearities that can compromise a highly accelerated stress.

to the tester and awaiting test. Even for mechanisms that do not relax when not under bias, this method has the disadvantage that no data can be obtained for the first few weeks after the hardware becomes available because the wafers are being built onto the chip carriers or modules and this build typically takes several weeks. The advantage of this method, even today, is that the stress equipment is relatively inexpensive and the test equipment is general enough to be used for readouts for several mechanisms. This amortizes the test equipment across all of these stresses and further decreases the cost of ownership.

However, for accuracy and simplicity, it is desirable to use the same equipment for the stress application and the readouts. This also minimizes handling damage and human error. Data acquisition improvements during the past 20 years have allowed the detection of exact times-to-fail even for the long three-month stresses. Now whether the stress is a long, three-month stress or a very short stress, the exact times-to-fail can be obtained with the same equipment.

In addition, advances in the state of the art in hot carrier stressing, in dielectric stressing, and in electromigration stressing have moved the leading practice to the far right two points on Figure 1.2, that is, to stresses of hours and minutes. Typically, stressing with seconds of duration is used in conjunction with at least one additional stress of longer duration. For example, in the case of dielectric stressing, optimally three voltages are stressed with the shortest stress duration having a median fallout on the order of 10 to 100 sec and the longest (lowest voltage) having a duration of 1000 to 10000 sec. This has been practiced for the last 5 to 10 years for dielectric stressing but as the state of the art thickness approaches 1 nm, it may actually become necessary to return to the relatively long stresses of several months. In the case of electromigration, the quantitative bridge for stressing on the order of seconds was only demonstrated a few years ago [6–8].

One of the obvious points that should be explicitly made is that for a three-month stress, the extrapolation to a 10-year use life is only a factor of 40. While for a 100 sec stress, the extrapolation is a factor of more than 1E6. Much more care concerning the projection error must be exercised when one is extrapolating six orders of magnitude, than when one is extrapolating only a little more than one order of magnitude.

Also one has to investigate very carefully whether any change in the accelerating condition of the mechanism in question has occurred or can occur under any reasonable set of conditions. This is depicted graphically in Figure 1.2 as new mechanisms, which may occur above a certain stress level. If any such mechanisms exist, they may be either linear or nonlinear as shown, and they would preclude exceeding that stress level since no straightforward model or bridge to use conditions would be possible in that case. Another possibility is that the accelerating or stress condition saturates above a certain level. That is, a further increase in the accelerating condition causes no resultant decrease in the lifetime. Again, a mechanism of this nature would limit the accelerating condition to a value no higher than just below its saturation value, and even there, the physics would need to be well understood.

The final concept that Figure 1.2 attempts to depict is that the window available for stressing is shrinking as the technology features continue to shrink. In the past, a very significant margin existed in many of the mechanisms. Often a calculated acceleration would demonstrate that the stress had gone well beyond a 10-year life but no wearout for that mechanism had been observed. Dielectric stressing is an excellent example of this. For 12 nm oxides, dielectric stressing on the order of three months could not detect any indication of wear out, and the stress focus was just on the extrinsic or defect part of the curve. Today, models are constructed to understand whether fractions of nanometers can be shaved from the oxide thickness and still meet the end of life targets.

1.2.3 Accelerating Condition

Acceleration concepts have been discussed and we now turn our attention to accelerating conditions. What types of external forces can be applied to the semiconductor test structure in such a way as to cause the end of life, say 10 years, to be reached in a time period that is significantly shorter than that 10-year life. In principle, the shorter the stress time the better, as long as one can still bridge to the use conditions. Examples of accelerating conditions for semiconductors are shown below. This extensive list includes all of the common accelerating conditions and a list of pertinent mechanisms with chapter references where applicable.

- Voltage (DC)
  - Dielectric breakdown (3.4)
  - Electromigration {indirectly} (7.3)
  - Hot carrier (5.2)
  - Temperature bias stability (4.3–4.5)
  - Interconnect opens and shorts (7.3)
  - Ionic contamination
  - Energetic particle-induced soft error mechanisms
  - Variable retention time mechanisms
  - Leakage mechanisms
- Voltage Change (AC)
  - Conducting hot carrier mechanisms (5.2)
- Temperature
  - Dielectric breakdown (3.4)
  - Electromigration {indirectly} (7.3)
  - Stress migration (6.2–6.5)
  - Interconnect shorts and opens (6.2–6.5, 7.3)
  - Hot carrier (5.2)
  - Temperature bias stability (4.3–4.5)

  - Ionic contamination
  - Variable retention time mechanisms
  - Leakage mechanisms
- Temperature Change
  - Interconnect opens
- Temperature Change Rate
  - Interconnect opens
- Current Density
  - Dielectric breakdown (3.2–3.4)
  - Electromigration (7.3)
- Humidity
  - Corrosion
- Humidity and Pressure
  - Corrosion
- Harsh environment
  - Corrosion
- Mechanical pull tests
  - Mechanical strength of interconnects and adhesives
- Radiation
  - Some dielectric breakdown concerns
  - Soft error rate (SER) for certain flash memory

Note: The SER effect does not get worse with time for most CMOS devices.

1.3 MECHANISM TYPES

1.3.1 Parametric or Deterministic Mechanisms

A parametric or deterministic mechanism is defined, albeit somewhat arbitrarily, as any mechanism that impacts all identical structures nearly equally. A stress for this type of mechanism will always cause the parameter under question to shift. And, even if many samples are stressed, the shifts will all be very close to the same value assuming all of the stressed structures are identical. For this reason, very small sample sizes can successfully be used to characterize a parametric mechanism. Most of the variation of the shifts observed for parametric mechanisms is caused by variations of the controlling parameters and not by random statistical variation.

The hot carrier (HC) mechanism is one example of a parametric mechanism. While a field effect transistor (FET) is turning on or off, the gate current has a peak value resulting from channel hot electron injection. These electrons gain enough energy to surmount the Si/SiO_2 interface without suffering energy-losing

collisions in the channel. The electrons are trapped and result in FET performance degradation. This mechanism is uniform and parametric in the sense that for a set of FETs that are all structurally identical, the shifts resulting from the above stress will be almost identical across all of the devices stressed; that is, the shifts will be determined by their parameter values, not by the random variation. In practice, if chips from several wafers or lots are stressed, variation will be seen but that variation will be a function of slight differences in the structures of the FETs across the wafers and lots.

Electromigration is an example of a mechanism that has aspects of a parametric mechanism. A current flowing through a line will cause atomic motion in that line. If that line is aluminum, significant atomic motion will occur at higher current densities and will cause the line resistance to increase and ultimately open. This is fundamental to the structure and the metallurgy. For high enough current densities, electromigration will always occur for that aluminum line. It is not caused by a defect although it can be exacerbated by a defect. Although the physics cannot be changed, sometimes it is possible to mitigate the problem. For example, if redundant layers of certain other metals are used in conjunction with aluminum, the sandwich line structure will increase in time-zero resistance if the overall cross-sectional area of the line remains constant, but electromigration typically will only cause a resistance increase and not an open under the same high current-density stress. For some metallurgies, no electromigration will occur even at higher current densities. However, it must be pointed out that in the case of electromigration, there are also aspects of a random mechanism because the grain structure of the line is random. And this randomness is true for metallurgy that is identical in processing. Typically, larger sample sizes are necessary when stressing mechanisms that have a greater degree of randomness.

Obviously it is crucial to understand the fundamental physics for a parametric mechanism. Once the physics is understood, strategies can be put into place to mitigate the effect or to eliminate the problem by structural or operating-point changes. Sometimes mitigation is possible and sometimes it is not. For the electromigration example, tungsten is sometimes used for the lower levels of wiring where the distances are small and the higher time-zero line resistance is tolerable. For the longer wiring levels, the resistivity of tungsten is too large and aluminum or copper must be used and other strategies invoked to decrease the impact of electromigration.

Once the physics is understood, so that all of the controlling parameters are identified and each of their impacts quantified, it is possible to address elimination and mitigation strategies. To be able to quantify the impact of a given parameter, it is usually necessary to characterize the impact of that parameter on a test structure where individual control of all of the terminals is possible. If, for example, the physics of the mechanism is related to a parasitic edge transistor in parallel to the bulk transistor, the decision must be made as to whether to change the process to eliminate the parasitic transistor, or to simply mitigate its impact on the circuit. A problem may occur only at one extreme of the normal processing window or set of biases and tolerances. HC is one example since it is worst at the shortest channel

lengths for a given set of stress conditions. In some cases the strategy may be to run the process to a tighter manufacturing limit. Because this type of mechanism equally affects all structures with the identical process, only a few structures need to be investigated to reasonably well characterize a parametric mechanism. However, these devices under tests (DUTs) must all be structurally identical.

From this previous discussion it should be obvious that it is necessary to investigate parametric mechanisms at all salient process window extremes to ensure that no undesired effects occur. Again, each process window investigation point requires only a small sample size.

Below are some examples of parametric mechanisms and one or two strategies for controlling or eliminating the effect. In most of the cases, there are other strategies that could also be invoked. Applicable chapter references are shown.

- *Hot carrier*: design point change, e.g., lower operating voltage the device experiences
- *Bias temperature stress or (negative bias temperature instability)*: design point change, e.g., decrease operating voltage
- *Ionic contamination*: discovery and removal of contamination source
- *Stress induced leakage current*: design point change, e.g., decrease operating voltage or thicken gate oxide
- *Electromigration*: design point change, decrease current density
- *Soft error rate (radiation induced)*: design point change, increase critical charge of pertinent cells or decrease charge collection efficiency. This is not discussed further in this book

1.3.2 Structural Mechanisms

Structural mechanisms are those mechanisms for which the fails physically occur in the same place. The distinction here from the structurally induced parametric fails is that these fails are only a function of a structural artifact. Although these definitions are all somewhat arbitrary, they help in understanding the sample size differences recommended in the later chapters. Usually significant failure analysis is required to determine that a particular failure type has a structural, systematic signature. Often this signature only occurs at one of the process extremes so that it does not occur on every wafer or lot. Sometimes it is even more difficult to identify because not only does it only occur at one process extreme, it may also require a certain set of process biases and/or tolerances to align in just a ‘‘right’’ way for the failure to occur. This may take the form of one part of the wafer having an acute susceptibility, or it may be tool dependent. In some of these cases, it may appear random, while in fact, the fail is part of a manufacturing defect or process window tail. This can usually be avoided if a large enough sample is investigated and if at least part of that sample comes from the salient process extremes. If the failure analysis then identifies a particular feature failing more than once, that feature should undergo very careful scrutiny.

Sampling is very important since the problem may not impact all lots or wafers or die equally. The sampling must gauge all process variations unless one process extreme can be identified as the worst case for the given mechanism. A minimum of three manufacturing lots is recommended with one produced at the identified critical extreme. For this type of mechanism, sampling for random statistical variation is less important than sampling for the pertinent process extremes.

Once this type of mechanism is understood, it can often be mitigated with a strict application of statistical process control (SPC). However, no structural fails should be acceptable within the normal process limits. Otherwise this would represent a technology weakness that, if accepted, would likely result in an inordinate number of customer failures even with tight SPC. It is always better in the long term, to fix a problem rather than to try control it. Fixing the problem can take the form of structural modifications or a redefinition of the process limits. Especially for hardware made later in a program, the normal exercise of SPC, once the process line is full of hardware, may eliminate the possibility of a problem.

Process improvements made during manufacturing can inadvertently introduce new structural mechanisms. An effective method to avoid this is to sample a large number of chips and wafers looking for changes even in time-zero characteristics. Changes in the time-zero characteristics will not always flag a change in a reliability mechanism, but a change in the time-zero characteristic should be carefully investigated especially if a significant database exists for the normal properties of the parameter. Wafer-level reliability (WLR) is an even better gauge as to the impact of process improvements on reliability. And in fact, occasionally the time-zero properties have been changed through process changes only to make the reliability worse in a direct tradeoff between yield and reliability. All of the examples for this type of mechanism are very technology/process dependent.

1.3.3 Statistical Mechanisms

Statistical mechanisms are defined as those mechanisms that are primarily random. Thus the more susceptible area to a particular statistical mechanism, the more likely that mechanism will cause a chip fail. The occurrence of the fail will be totally random within that susceptible area. This is in contrast to a structural fail, which will always occur at a given feature within a structure. It is also in contrast to the parametric or deterministic mechanism for which the process variation or process extreme will have a larger impact on the result than does the random statistical variation within a given process point. One must be careful at this point because the statistical mechanisms are also caused by fundamental physics unless the discussion is limited to defects. The distinction is more focused on the impact that the random statistical variation has on the investigation of the mechanism.