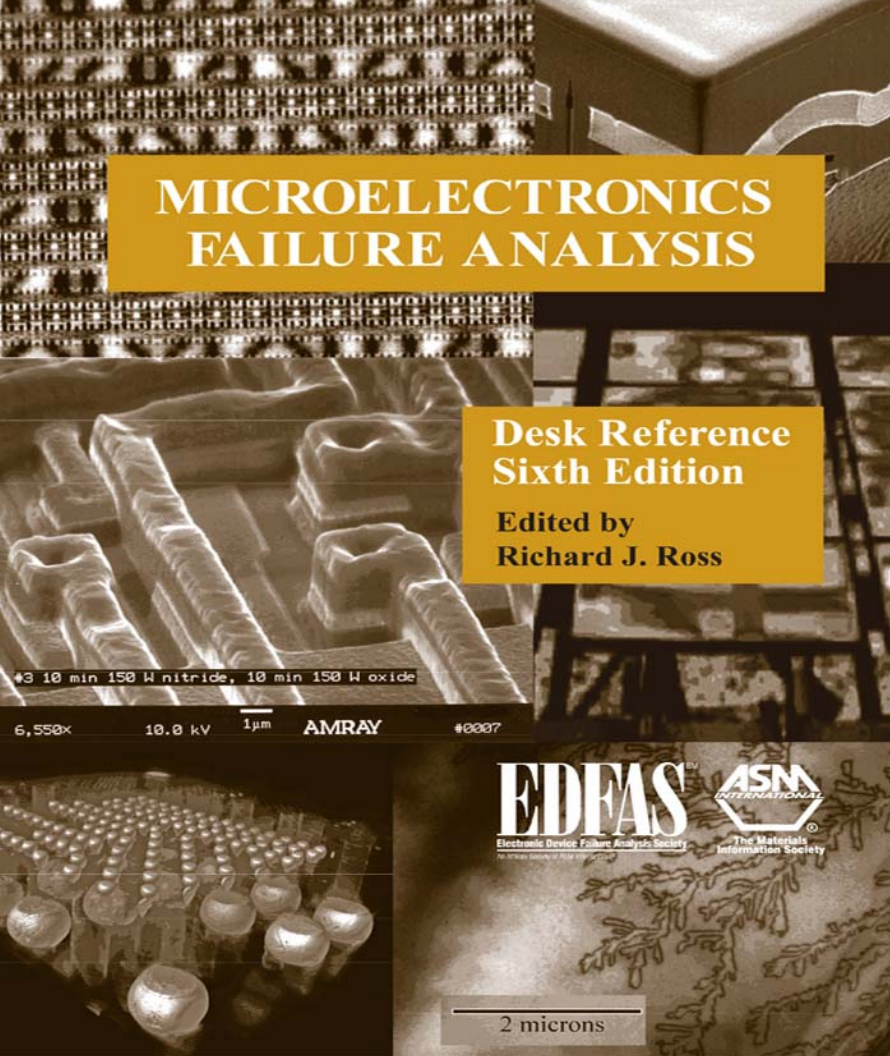


MICROELECTRONICS
FAILURE ANALYSIS
Desk Reference
Sixth Edition
Edited by
Richard J. Ross
#3 10 min 150 W nitride, 10 min 150 W oxide
6,550x
10.0 kV
1µm
AMRAY
#0007
EDFAS
Electronic Device Failure Analysis Society
ASM INTERNATIONAL
The Materials Information Society
2 microns

Microelectronics Failure Analysis

Desk Reference
Sixth Edition

Edited by
Richard J. Ross

Published by
ASM International®
Materials Park, Ohio 44073-0002
www.asminternational.org

Copyright © 2011
by
ASM International®
All rights reserved

No part of this book 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 written permission of the copyright owner.

First printing, October 2011

Great care is taken in the compilation and production of this book, but it should be made clear that NO WARRANTIES, EXPRESS OR IMPLIED, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, ARE GIVEN IN CONNECTION WITH THIS PUBLICATION. Although this information is believed to be accurate by ASM, ASM cannot guarantee that favorable results will be obtained from the use of this publication alone. This publication is intended for use by persons having technical skill, at their sole discretion and risk. Since the conditions of product or material use are outside of ASM's control, ASM assumes no liability or obligation in connection with any use of this information. No claim of any kind, whether as to products or information in this publication, and whether or not based on negligence, shall be greater in amount than the purchase price of this product or publication in respect of which damages are claimed. THE REMEDY HEREBY PROVIDED SHALL BE THE EXCLUSIVE AND SOLE REMEDY OF BUYER, AND IN NO EVENT SHALL EITHER PARTY BE LIABLE FOR SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES WHETHER OR NOT CAUSED BY OR RESULTING FROM THE NEGLIGENCE OF SUCH PARTY. As with any material, evaluation of the material under end-use conditions prior to specification is essential. Therefore, specific testing under actual conditions is recommended.

Nothing contained in this book shall be construed as a grant of any right of manufacture, sale, use, or reproduction, in connection with any method, process, apparatus, product, composition, or system, whether or not covered by letters patent, copyright, or trademark, and nothing contained in this book shall be construed as a defense against any alleged infringement of letters patent, copyright, or trademark, or as a defense against liability for such infringement.

Comments, criticisms, and suggestions are invited, and should be forwarded to ASM International.

ISBN-13: 978-1-61503-725-4
ISBN-10: 1-61503-725-X
SAN: 204-7586

ASM International®
Materials Park, OH 44073-0002
www.asminternational.org

Printed in the United States of America

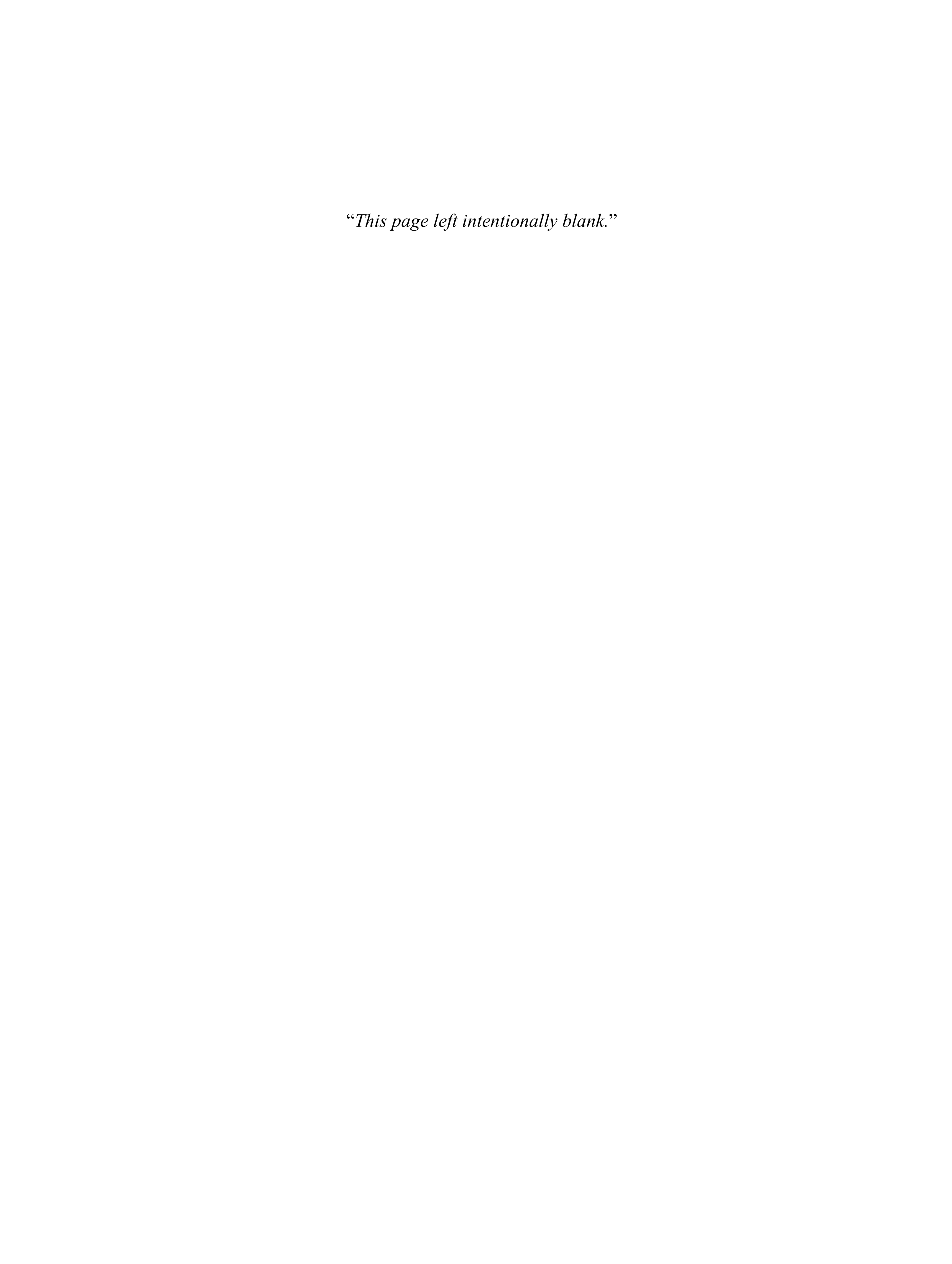
Editorial Board

Editor-In-Chief

Richard J. Ross, *Consultant*

Editors/Section Champions

Vijay Chowdhury, *Evans Analytical Group*
Dermot Daly, *Xilinx*
Dave Dozor, *IR Labs*
George Gaut, *Qualcomm*
Cheryl Hartfield, *Omniprobe*
Leo G. Henry, *ESD/TLP Consulting*
Becky Holdford, *Texas Instruments*
Kultaransingh (Bobby) Hooghan, *FEI Corp*
Martin Keim, *Mentor Graphics*
Larry Kessler, *Sonoscan*
Steven Maher, *Oklahoma Christian University*
Richard J. Young, *FEI Corp*.
Thomas Zanon, *PDF Solutions*



"*This page left intentionally blank.*"

Contents

Preface to the Sixth Edition xi

Section 1: Introduction

The Failure Analysis Process 1
M. Steven Ferrier

Section 2: Failure Analysis Process Overviews

System Level Failure Analysis Process: Making Failure Analysis a Value Add Proposition in Today's High Speed Low Cost PC Environment 16
Michael Lane, Roger Bjork, Jeff Birdsley

Board Level Failure Mechanisms and Analysis in Hand-Held Electronic Products 23
Sridhar Canumalla, Puligandla Viswanadham

Failure Analysis Flow for Package Failures 34
Rajen Dias

Chip-Scale Packages and Their Failure Analysis Challenges 40
Susan Xia Li

Wafer Level Failure Analysis Process Flow 49
J.H. Lee, Y.S. Huang, D.H. Su

Failure Analysis of Microelectromechanical Systems (MEMS) 52
Jeremy A. Walraven, Bradley A. Waterson, Ingrid De Wolf

Failure Analysis and Reliability of Optoelectronic Devices 78
Robert W. Herrick

Solar Photovoltaic Module Failure Analysis 99
G.B. Alers

DRAM Failure Analysis and Defect Localization Techniques 104
Martin Versen

Failure Analysis of Passive Components 111
Stan Silvus

Section 3: Failure Analysis Topics

Reliability and Quality Basics for Failure Analysts 121
Steven Hoffman, Chris Henderson

Electronics and Failure Analysis 128
Jerry Soden, Jaume Segura, Charles F. Hawkins

Submicron CMOS Devices 149
Theodore A. Dellin

Analog Device and Circuit Characterization 159
Steve Frank

Screening for Counterfeit Electronic Parts 171
Bhanu Sood, Diganta Das

Section 4: Fault Verification and Classification

An Overview of Analog Design for Test and Diagnosis 181
Stephen Sunter

An Overview of Integrated Circuit Testing Methods 190
Anne Gattiker, Phil Nigh, Rob Aitken

Diagnosis of Scan Logic and Diagnosis Driven Failure Analysis 199
Srikanth Venkataraman, Martin Keim, Geir Eide

Interpretation of Power DMOS Transistor Characteristics
Measured with Curve Tracer 209
Hubert Beermann

High-Volume Scan Analysis: Methods to Avoid Failure Analysis 218
Darrell Carder, Steve Palosh, Rajesh Raina

Differentiating between EOS and ESD Failures for ICs 225
Leo G. Henry

The Power of Semiconductor Memory Failure Signature Analysis 239
Cary A. Gloor

Section 5: Localization Techniques

Beam-Based Defect Localization Techniques 246
Edward I. Cole, Jr.

Electron Beam Probing 263
John T.L. Thong

Failure Localization with Active and Passive Voltage Contrast in FIB and SEM 269
Ruediger Rosenkranz

Fundamentals of Photon Emission (PEM) in Silicon – Electroluminescence for Analysis of Electronic Circuit and Device Functionality 279
Christian Boit

Picosecond Imaging Circuit Analysis – PICA 292
D. Vallett

Current Imaging Using Magnetic Field Sensors 301
L.A. Knauss, S.I. Woods, A. Orozco

Thermal Defect Detection Techniques 310
Daniel L. Barton, Paiboon Tangyunyong

Thermal Failure Analysis by IR Lock-In Thermography 330
O. Breitenstein, C. Schmidt, F. Altmann, D. Karg

Principles of Thermal Laser Stimulation Techniques 340
F. Beaudoin, R. Desplats, P. Perdu, C. Boit

Introduction to Laser Voltage Probing (LVP) of Integrated Circuits 349
Siva Kolachina

CAD Navigation in FA and Design/Test Data for Fast Fault Isolation 354
William Ng

Acoustic Microscopy of Semiconductor Packages 362
Cheryl D. Hartfield, Thomas M. Moore

Electronic Package Fault Isolation Using TDR 383
D. Smolyansky

Section 6: Deprocessing and Sample Preparation

Delayering Techniques: Dry Processes Wet Chemical Processing and Parallel Lapping 397
Kendall Scott Wills, Srikanth Perungulam

The Art of Cross Sectioning 417
B. Engel, E. Levine, J. Petrus, A. Shore

Delineation Etching of Semiconductor Cross Sections 437
S. Roberts, D. Flatoff

Special Techniques for Backside Deprocessing 440
Seth Prejean, Brennan Davis, Lowell Herlinger, Richard Johnson, Renee Parente, Mike Santana

Deprocessing Techniques for Copper, Low K, and SOI Devices 445
Huixian Wu, James Cargo

Section 7: Inspection

Optical Microscopy 457
John McDonald

Scanning Electron Microscopy 477
W. Vanderlinde

Ultra-High Resolution in the Scanning Electron Microscope 497
W. Vanderlinde

Transmission Electron Microscopy for Failure Analysis of Semiconductor Devices 506
Swaminathan Subramanian, Raghaw S. Rai

X-ray Imaging Tools for Electronic Device Failure Analysis 529
Steve Wang

Atomic Force Microscopy: Modes and Analytical Techniques with Scanning Probe Microscopy 536
J. Colvin, K. Jarausch

Section 8: Materials Analysis

Energy Dispersive X-ray Analysis 549
W. Vanderlinde

Analysis of Submicron Defects by Auger Electron Spectroscopy (AES) 561
Juergen Scherer, Patrick Schnabel, Kenton Childs

SIMS Solutions for Next Generation IC Processes and Devices 573
Gary Mount, Yung Liou, Han-Chung Chien

Section 9: Focused Ion Beam Applications

Focused Ion Beam (FIB) Systems: A Brief Overview 583
Kultaransingh (Bobby) Hooghan, Richard J. Young

Circuit Edit at First Silicon 594
Ted Lundquist, Mark Thompson

The Process of Editing Circuits through the Bulk Silicon 607
Nicholas Antoniou

Section 10: Management and Reference Information

Education and Training for the Analyst 612
Christopher L. Henderson

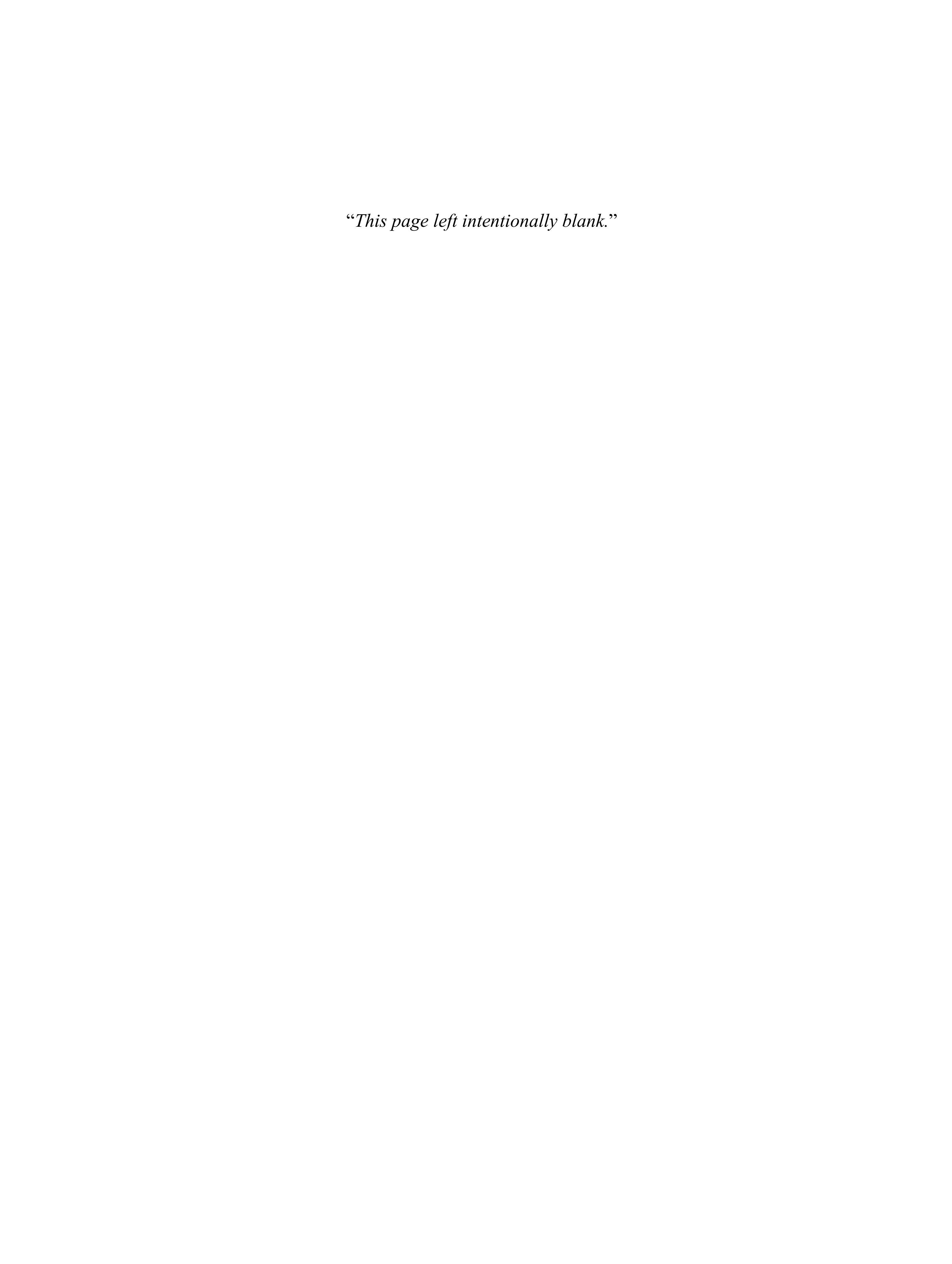
Management Principles and Practices for the Failure Analysis Laboratory 617
Richard J. Ross

Managing the Unpredictable – A Business Model for Failure Analysis Service 627
C. Boit, K. Scholtens, R. Weiland, S. Görlich, D. Schlenker

Failure Analysis Terms and Definitions 635
Ryan Ong

Author Index 651

Subject Index 653



“*This page left intentionally blank.*”

Preface to the Sixth Edition

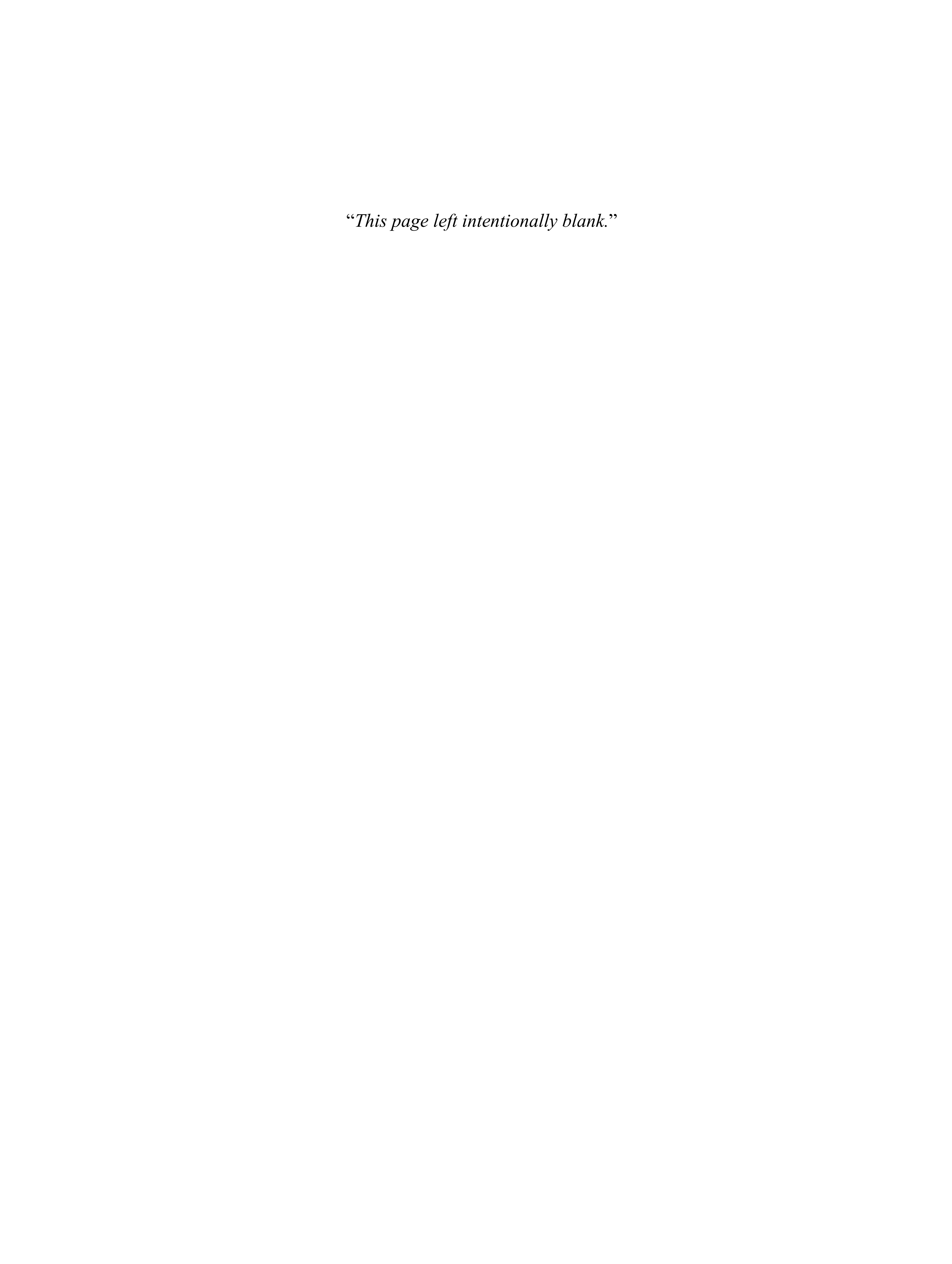
Richard J. Ross, Editor-in-Chief

As the semiconductor industry moves from the "micro" to the "nano" realm, the Failure Analysis community needs to be pro-active in maintaining its ability to verify, isolate, uncover, and identify the root-cause of problems. These problems may be discovered in design debug, product or technology development and qualification, fabrication, packaging, reliability stress, or, most unfortunately, in the field. New materials and ever-shrinking technology dimensions make it increasingly more challenging for the failure analyst and make it increasingly important to provide analysis with information, training, equipment, and materials to enable them to cope with these challenges and opportunities.

For over three decades, this work, "Microelectronics Failure Analysis Desk Reference" has been a key aide to analysts. It has been used as a textbook, a workbook, and a laboratory manual over that time and has undergone, now, six iterations of content selection and revision. The printed version has gone from 275 pages to over 600. Some of the methods and techniques which are included in this edition did not even exist when the first edition was published in the 1970s.

The work itself must change as well to reflect the challenges and opportunities of the times. This edition will exist in web-based online, DVD, and printed form to meet the diverse needs of the community. Some "old favorites" remain as their base technology and practice are still relevant and useful; others have become superseded by the technology. The use of color throughout the work is also introduced in this edition.

No undertaking of this magnitude is accomplished without the efforts of many. The Editorial Board and Section Champions recruited experts in the various specialized fields, nurtured and encouraged them, and drove to a schedule which, in the economic climate of the past few years, required understanding and revision. The staff at ASM International, particularly Kate Russell and Scott Henry were of immeasurable help and support. Thomas Zanon, EDFAS Education Chair, was tolerant, understanding, and supportive when the inevitable frustrations of time and effort appeared. Without the various authors, of course, this work does not exist and I am eternally grateful to each of them. Finally, I want to thank my family for their support for the time spent on the computer and the phone.



"*This page left intentionally blank.*"

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

The Failure Analysis Process

M. Steven Ferrier
Metatech Corporation • 1190 SW Sunset Drive • Corvallis, OR 97333
Telephone 541-752-2713 • Email metatech@comcast.net

Introduction. How does an experienced Failure Analysis engineer crack a tough analysis? What makes the difference between a new engineer's Failure Analysis approach and the seasoned, effective analysis method of a veteran? Can our industry capture a formulation of correct Failure Analysis methodology to accelerate the development of new engineers and aid the strong skills of experienced analysts?

More crucial, perhaps, is the underlying question, Are we able as analysts to think abstractly about the nature of our work? Or are we instead bound to a tool-focused, concrete way of thinking that has all the adaptability of the construction material of the same name? While our historical inability in the FA industry to abstract a general process from our daily work will probably not doom our discipline, this shortcoming will certainly rob us of a great deal of efficiency, flexibility and basic understanding of what we do in electronic device Failure Analysis. This author has the conviction that we are a better group of people than to allow this shortcoming to continue undetected and uncorrected for much longer.

While many or most Failure Analysis departments make the definition and application of specific analysis process flows a significant priority, the subject of general FA methodology remains a minor one in the literature as of this writing.[1-3] By way of a summary remediation, this tutorial paper presents a fully-general analysis formalization intended to make the failure analysis design efficient, rigorous and consistent from analysis to analysis and from failure to failure. The formalization amounts to a practical form of scientific method[4,5], customized for the failure analysis discipline, and enabling faster analysis to satisfy business constraints.

The analysis design methodology, or decision engine, can be formulated as six steps arranged in a cycle or loop, shown in Fig. 1. This may not look like a Failure Analysis process flow, which is appropriate, since it is not, strictly speaking, such a flow. Instead, these steps in this order form a *metaprocess*, that is, a process whose function is to create another process. Applied to semiconductor failure analysis, this general decision engine metaprocess creates a specific analysis flow for a specific failure based on the facts about the failure, the results of analysis operations, the body of knowledge about physical causes and their effects and, finally, inferential logic. This metaprocess is intended to work properly—that is, generate a correct unique failure analysis flow—for any failure analysis of a manufactured product coming from any of the semiconductor wafer and packaging processes predominant today. Using the approach described in this tutorial, the analyst of any experience level should become able to make analysis choices in ways that reflect mature failure analysis skills. This approach can thus accelerate the maturity

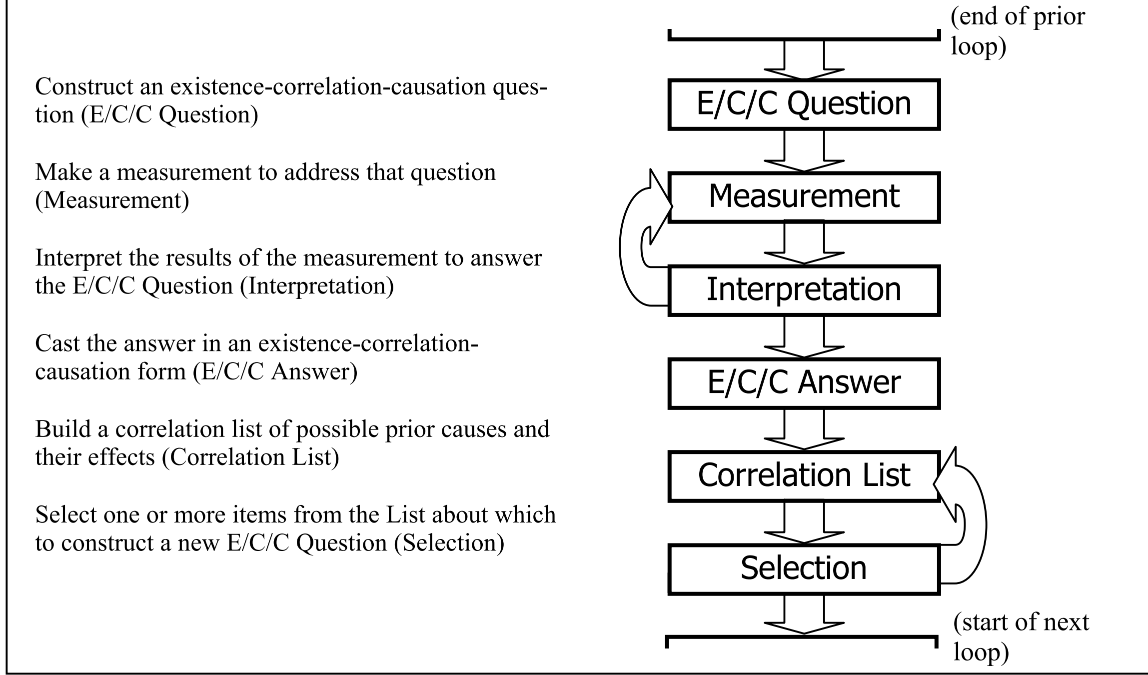


Figure 1 Scientific Method Metaprocess Loop for Failure Analysis (unfolded version)

of beginning analysts once they have the small amount of experience needed to use the metaprocess effectively. Some benefits of this methodology will be described in detail a little later in the tutorial. All examples of any kind in this tutorial, with one major exception described shortly, are hypothetical but possible.

The Metaprocess in a Nutshell. Very briefly, here is how this metaprocess produces a correct device analysis flow. All analysis starts with understanding the failure mode. We have a part and a complaint; does the failure mode exist on this part and match the complaint? We make a measurement to confirm the failure mode. If we do confirm the mode, it makes sense, rather than just jumping into a measurement, to ask ourselves what could cause the mode. Even if we have no data yet, we can still come up with a list of possible causes that may be substantial. The next step is to write down on our list *other* effects of the causes we just wrote down on our list. We then select one effect on our list and ask whether it exists.

The reader will notice that at this point we are back where we started: asking whether an effect exists. We began by asking whether the effect called the failure mode exists, and ended by asking whether an effect on our list existed. This makes it clear that the pro??cess we are using represents a loop. If we follow this loop, it will lead us from the failure mode to the root anomaly. Pursuing a rigorous version of this approach provides the correct analysis flow.

The Metaprocess in Real Life. To show this, we start with a real-life example of the use of this approach, drawn verbatim (excepting a few editorial excisions) from a published ISTFA paper[6]. After this practical example, we will provide some clear definitions of basic terms and then proceed to explain the six steps on their own merits.

Application of a formal Failure Analysis metaprocess to a stubborn yield loss problem provided a framework that ultimately facilitated a solution. Absence of results from conventional failure analysis techniques such as PEM (Photon Emission Microscopy) and liquid crystal microthermography frustrated early attempts to analyze this low-level supply leakage failure mode. Subsequently, a reorganized analysis team attacked the problem using a specific top-level metaprocess[a,4].

Using the metaprocess, analysts generated a specific unique step-by-step analysis process in real time. Along the way, this approach encouraged the creative identification of secondary failure effects that provided repeated breakthroughs in the analysis flow. Analysis proceeded steadily toward the failure cause in spite of its character as a three-way interaction among factors in the IC design, mask generation, and wafer manufacturing processes. The metaprocess also provided the formal structure that, at the conclusion of the analysis, permitted a one-sheet summary of the failure's cause-effect relationships and the analysis flow leading to discovery of the anomaly.

As with every application of this metaprocess, the resulting analysis flow simply represented an effective version of good failure analysis. The formal and flexible codification of the analysis decision-making process, however, provided several specific benefits, not least of which was the ability to proceed with high confidence that the problem could and would be solved. This tutorial describes the application of the metaprocess, and also the key measurements and cause-effect relationships in the analysis.

Manufacturing and Failure Mode. The yield failure mode, labeled V_{DD} Leakage, occurred in a specific printhead IC manufactured in a 1.0 um single-poly three-metal CMOS process. Although this is a relatively simple wafer process, the mode's characteristics conspired to counteract any real benefit from the process simplicity. Up to 80% of the IC's manufactured exhibited a slightly elevated (but definitely anomalous) supply leakage current in a standby con-

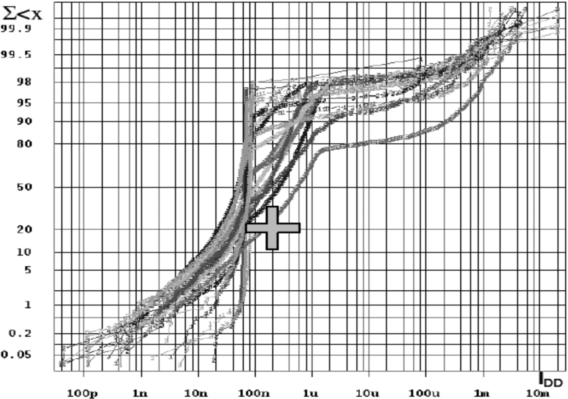


Figure 2 Cumulative distribution of failing design leakage behavior of wafer lots with only 20% of I_{DD} currents falling below 200nA.

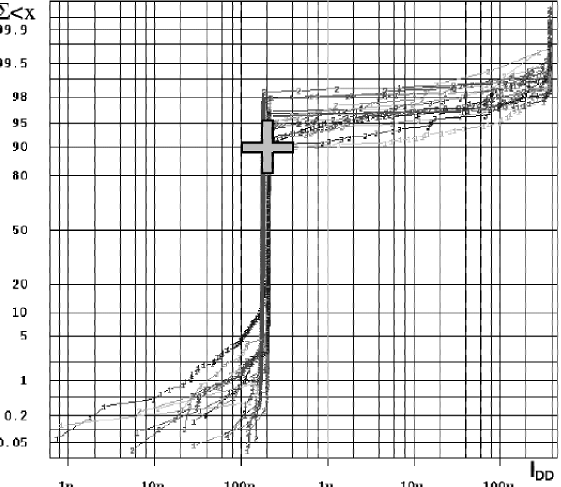


Figure 3 Cumulative distribution of leakage behavior of related but non-failing design with I_{DD} of over 90% of devices falling below 200nA.

---

a A metaprocess is a general process whose purpose and function is to create a specific process to suit specific conditions and requirements.

figuration compared to comparable IC designs. The anomalous leakage current revealed itself as a sloppy cumulative current distribution curve (Fig. 2) compared to the clean curves of non-failing designs (Fig. 3).

Another fact stood out at the outset as a clear clue. The leakage did not occur on the susceptible design when it was manufactured in an older wafer process with significantly heavier n-well doping. Early failure analysis indicated no correlating liquid-crystal transitions, a result consistent with the leakage levels of approximately 1-5uA. Other traditional analysis measurements (including a search for correlating photon emission) also proved unsuccessful. In spite of months of yield loss without a resolution, the team responsible for yield refused on principle to simply raise the test limit and redefine this product as 'good'.

Application of the Metaprocess. The problem was eventually returned to the IC's designers, and a task force assembled across design and manufacturing organizations to solve the problem. The ensuing analysis was guided chiefly by a cyclic metaprocess, diagrammed in Fig. 1 above. This metaprocess performed a critical three functions for the analysis effort.

Function 1. The metaprocess provided a constant focus on (a) the nature and details of the cause-effect network that the failure had created, (b) the analysis path, and (c) the location (status) of current analysis activity on that path. The metaprocess provided this function through the canonical form of the E/C/C Question, "Does <the selected specific possible prior cause> Exist, Correlate with <the known, observed failing behavior>, and Cause that behavior? For example, at one point the question became: "Does the possible gate oxide short exist, correlate with the known quadratic I_{DD} leakage, and cause that leakage?" Depending upon the answer to this question, one of the following three cases applied:

I. The analysis path is false (in this example, the possible cause, a gate short, did not exist);

II. The analysis path is near the main cause-effect relationship (the possible cause, a gate short, existed and correlated with the observed quadratic I_{DD} behavior);

III. The analysis path is at the main cause-effect relationship (the possible cause, a gate short, existed, correlated with the known I_{DD} behavior, and was related to the behavior as cause and effect by a specific physical law of transistor operation).

Function 2. The metaprocess directed focus evenly on both halves of any good analysis: measurement and prior cause postulation. The former tempts many analysts to neglect to focus on possible prior causes. The Correlation List step (middle bottom) forced consideration of possible causes for the observed behavior, and then focused analysis measurement design to the goal of detecting secondary or side effects of those possible causes[5]. This repeatedly drove innovation and inspired creative analysis approaches. Conversely, the Measurement step (middle top) disciplined the construction of the Correlation List to treat possible prior causes arising from the failure's anomalous behavior rather than from speculation.

Function 3. The metaprocess provided application points for methodical analysis logic. At some points during analysis, behavior that correlated with failure differed very little from behavior that did not. The metaprocess facilitated the use of logical principles able to accurately distinguish fine differences, such as the Contrapositive Test[b] and Inductive Proof[c].

Analysis Flow. The following section describes the overall analysis flow arising from the metaprocess, including some details of its application.

A. Failure Mode Verification and Characterization. In verifying failure mode, analysis began with a focus on the causes and effects operating in the failing IC, incorporating the first function informed by the metaprocess described earlier (providing a constant focus on the nature and details of the cause-effect failure-created network, the analysis path, and the location of current analysis activity on that path). The metaprocess loop was initiated, as Fig. 1 indicates, at the E/C/C Question step.

<u>E/C/C Question</u>: Does the reported V_{DD} leakage exist, correlate with the complaint, and cause this IC to be categorized as a failure?

<u>Measurement</u>: Under conditions matching the failed test, the die under analysis exhibited anoma-

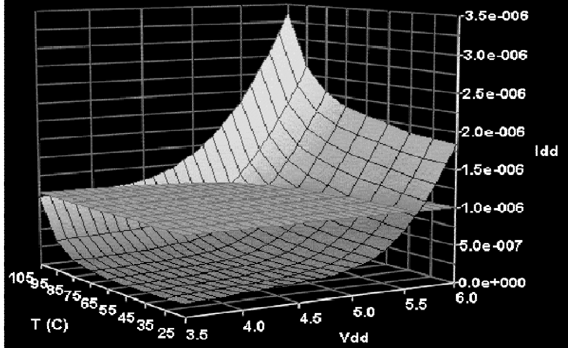


Figure 4 $I_{DD}(V_{DD},T)$ for a failing IC, with a superimposed flat 1.0uA test limit surface.

lous $I_{DD}(V_{DD},T)$ current (Fig. 4).

<u>Interpretation</u>: Voltage and temperature behavior of the anomalous current matched the complaint behavior.

---

b Contrapositive Test: The test of a statement P by seeking a valid consequent Q that is false. P=>Q means that any ~Q=>~P. In FA terms, if a sure effect is truly absent, its cause is absent.

c Inductive Proof: The proof of a statement P by sufficient demonstration of valid and true consequents Q_n without exception: for n>sufficient limit, P=>Q_n and all Q_n are found true. In FA terms, if the first predicted side effect is present, test for the next. If sufficient side effects are present without exception, declare the common cause of all side effects proved and therefore present.

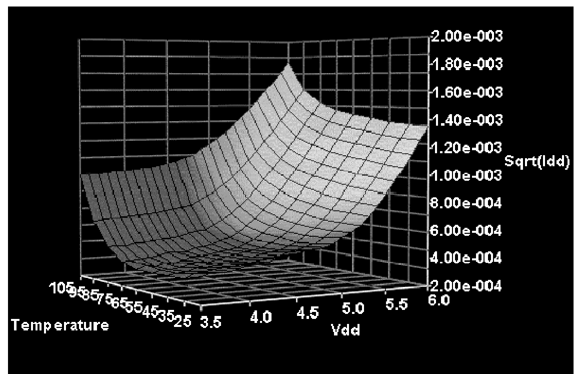


Figure 5 Plot of the square root of $I_{DD}(V_{DD},T)$ exhibiting a linear relationship to V_{DD} axis (note straight surface grid lines in lighter area).

E/C/C Answer: V_{DD} leakage existed, correlated with the complaint, and caused the IC to be categorized as a failure.

Possible Prior Cause of Known Failing Behavior (V_{DD} leakage)	Predicted or Known Correlating Side Effect of Prior Cause (at Left)
Transistor on-state	Quadratic $I_{DS}(V_{GS})$
	Negative $I_{DS}(T)$
Gate Oxide short	Quadratic $I_{DS}(V_{GS})$
	Negligible $I_{DS}(T)$
	Strong photon emission
Reverse-biased diode leakage	I_{DS}~$V^{1/2}$
Forward-biased diode leakage	I_F~e^V
Subthreshhold current	I_{DS}~e^V
Resistive anomaly	I=V/R (linear)
Latchup current	I >~ 1mA
	Latching effect (hysteresis)

Table 1 First Cycle Correlation List

B. Analysis Flow Generated by the Metaprocess. The main part of the analysis follows. After the E/C/C Answer that validated the failure mode, a Correlation List was next constructed. This step addressed two (non-E/C/C) questions:

1. What possible prior causes may create the confirmed V_{DD} leakage?
2. What other effects do each of these possible prior causes imply?

This step began by implementing the second function described earlier (directing focus evenly on both measurement and prior cause postulation). This was approached through brainstorming, through research into failure cause-effect relationships, and through developing a thorough understanding of the intentional, designed cause-effect relationships applied by the technology. The Correlation List for this analysis consisted (in part) of the items in Table 1. Note that the side effects represent the behaviors available for the next measurement(s), and yield actual prior causes.

Selection: Items from the Correlation List were selected for further analysis based upon a set of prioritization principles. Analysis then proceeded to the E/C/C Question step and subsequently through the next section of the metaprocess. Early $I_{DD}(V_{DD})$ results (as described) immediately ruled out several of the items on the Correlation List (such as resistive anomalies) using the logical principle of the Contrapositive Test. Since these were disproved, they were not pursued. The transistor on-state selection presented the most promise; therefore the analysis pursued that path.

E/C/C Question: Does an anomalous transistor on-state exist, correlate with the known V_{DD} leakage and its characteristics, and cause the V_{DD} leakage?

Measurement: Anomalous $I_{DD}(V_{DD},T)$ exhibited quadratic voltage dependence (Figs. 5 and 6).

Interpretation: Quadratic voltage dependence was observed to be consistent with the behavior of an MOS transistor which is turned on, and corresponded to the well-known dependence of the source-drain current I_{DS} on the gate voltage V_{GS} in saturation[7]:

$$I_{DS} = \mu\varepsilon_{ox}W/2Lt_{ox}\,(V_{GS}-V_t)^2 \qquad (1)$$

This correspondence supported the possibility that the V_{DD} leakage was caused by an anomalous transistor on-state. In fact, (1) describes a physical law that would permit the declaration of such an MOS structure (once found) to be a cause for the leakage.

Measurement: In answering the E/C/C Question, the measurement and interpretation steps may be repeated as many times as necessary. This next measurement (Fig. 7) indicated that the anomalous $I_{DD}(V_{DD},T)$ possessed positive temperature dependence. Reasons for this characteristic were not fully

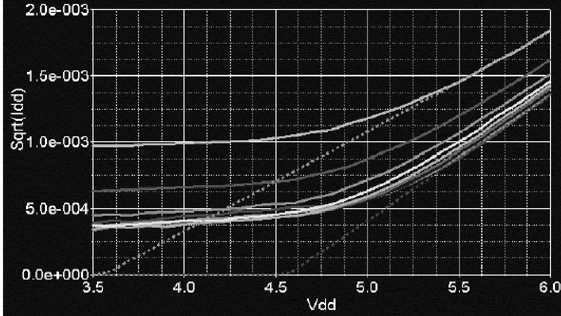


Figure 6 2-D plot of the data in Fig. 5, indicating the quality of linear fit to Sqrt(I_{DD}) in higher-V_{DD} regimes.

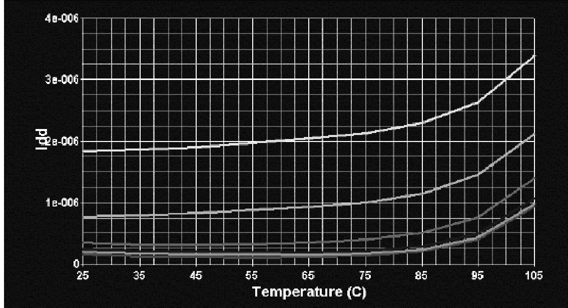


Figure 7 2-D plot of the data in Fig. 4, exhibiting positive leakage temperature dependence.

apparent during this phase of the analysis, but possible explanations arose near the end of the analysis from the observed root anomaly. Even more importantly in the short term, this characteristic helped localize the failure site (described shortly).

Interpretation: This observed temperature behavior was the opposite of that expected from a transistor. Normally, channel mobility μ, which dominates[8] the temperature dependence of the saturated MOS drain current I_D, decreases with increasing temperature[9]. Therefore, the structure that was causing the anomalous current to flow should not be solely an MOS transistor. Nonetheless, the quadratic I(V) behavior represented an indicator of MOS transistor involvement that could not be dismissed. This conflict was resolved by postulating a structure that included an MOS transistor, but also incorporated other components whose temperature dependence dominated that of the transistor. Such a structure could be postulated while leaving its particular components unspecified for the moment.

Measurement: Next, a new photon emission measurement was performed, which again exhibited no emission signal, even at higher voltages and temperatures that significantly increased the failing V_{DD} leakage.

Interpretation: System tests verified that the PEM systems used in these measurements were capable of detecting emission from forward-biased diode structures passing currents in the single-digit microampere range.

Recombination rates of MOS transistors were expected to fall below this range, and therefore any such emission would remain undetected. This negative result did, however, definitively rule out a gate oxide short (whose stronger emission should have been detected). The logical principle of the Contrapositive Test immediately demanded this conclusion.

E/C/C Answer: A circuit anomaly that includes a transistor on-state exists, correlates with the confirmed V_{DD} leakage, and causes it.

Correlation List: The Correlation List (Table 2) was next constructed. This list contains possible prior causes for the just-confirmed behavior (the anomalous transistor on-state), and also includes other side effects (effects other than observed V_{DD} leakage) of those possible prior causes. In clarification of the reset signal entry noted in Table 2, it was known from earlier work that the external reset pin could turn off the leakage on failing dice, which led to the reset state as a possible prior cause in the electrical domain. The Table also includes normal transistors in an abnormal on state due to abnormal electrical conditions, as well as abnormal (unintended) transistors.

Selection: The correlation of the reset pin state with the observed V_{DD} leakage was selected. This prior cause postulated that the test, which was expected to set the reset pin to an active state, actually set it to an inactive state thereby allowing the leakage to occur.

In order to summarize the next several metaprocess steps, an E/C/C Question was framed concerning the selected PPC (Possible Prior Cause); that is, the reset pin state's correlation to the leakage: does an incorrect reset condition exist, correlate with the leakage, and cause it? Following this, a measurement determined that the normal reset state (rather than the incorrect state) correlated with leakage. This ruled out incorrect reset pin setup during testing and, in terms of the E/C/C Answer, that this possible prior cause did not exist. (Note that this did not mean that the correlation of the reset pin state to the failure was ignored, but rather that a test error creating the leakage by an incorrect reset pin state was ruled out.)

From the metaprocess flow, the next step would be to make a new selection from the same Correlation List since the prior cause of the same known effect was still under examination. In order to pursue the remaining Correlation List items (as described), however, it was necessary to resolve the question of exactly which structure was causing the leakage.

The most important methodology question at this stage in any analysis is: how can more information about the cause-effect relationships involved in the failure be obtained? This question drives the most challenging and significant operation within typical analysis flows: localization. Localization is valued as the only effective way to gather extensive information about the effects created by a defect. The ultimate purpose of localization is to provide greater access to the cause-effect network immediately surrounding the defect.

During a localization operation, the analysis makes no progress down the cause-effect chain. Instead, the analyst spends time and effort working to place the known failure behavior characteristics at some location in the IC's physical structure. This means that the analysis remains at the same point in the cause-effect chain while the spatial correlation of the failure's known behavior is identified.

In this case, PEM and Liquid Crystal provided no signals for localization. The IC was too complex to perform electrical localization in the design's own cause-effect domains (blocks' effects on other blocks), and the network of metallization was too dense to easily permit Focused Ion Beam (FIB) cut localization of the leakage path (from the topmost die level).

The discipline imposed by the Correlation List step (demanding enumeration of all possible secondary effects of the possible prior causes) yielded a return at this point. The leakage's positive temperature coefficient (Figs. 5-7) should be local to the anomaly site. The analysts therefore sought to locate this characteristic in the die interior.

E/C/C Question: Does a local thermal effect exist, correlate with the observed global I_{DD} temperature characteristics, and cause them? With this question in mind, a local heater was built. Its tip diameter (defining its spatial resolution) was large at 1000 um, but this still represented a relatively small region of the IC.

Measurement: Using the topside airflow heater, a topside temperature rise was impressed at various points across the IC, and a very slight rise in I_{DD} leakage as a function of the heater position was observed. Considerable effort was required to separate the signal from the noise, and only a few data points per day resulted.

Possible Prior Cause (PPC) of Known Failing Behavior (On-Transistor)	Predicted or Known Correlating Side-Effect of Prior Cause at Left
Incorrect reset signal state during test setup causes normally-off transistor to be on	Reset forced to abnormal state should produce global leakage
	Reset forced to normal state should eliminate global leakage
Normal transistor otherwise wrongly activated	Electrical anomaly will appear in fabricated IC
	Observed global I(V) will occur at a local region in the IC
	Observed global I(T) will occur locally
Abnormal transistor created and active	Physical anomaly will appear in fabricated IC
	Observed global I(V) will occur at a local region in the IC
	Observed global I(T) will occur locally

Table 2 Second Cycle Correlation List

Interpretation: Although not all regions were examined due to the laborious process involved, results obtained suggested that approximately 90% of the IC area could be ignored, and that the analysis could focus on a relatively small region at the top of the die. A plot of the signal as a function of die location appears in Fig. 8. The region at the left of the figure is at the top of the die. This interpretation confirmed

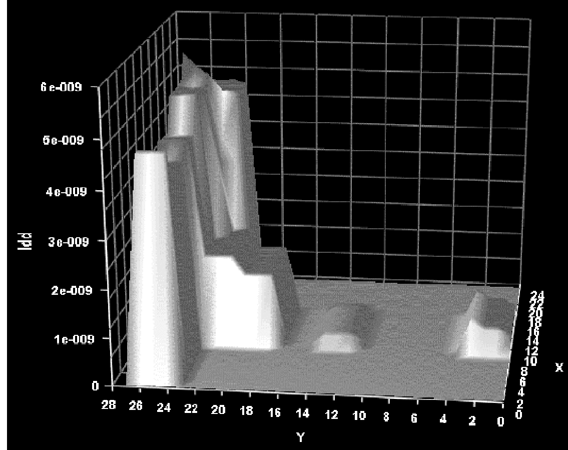


Figure 8 Δ $I_{DD}(x,y)$ generated by topside heating

that the effect named in the Correlation List (Table 2), a localized $I_{DD}(T,x,y)$, did exist and correlated with the global failure behavior. Since no physical law demanded that this thermal characteristic caused the observed leakage, only the first two sub-questions could be answered positively for this possible prior cause.

E/C/C Answer: A local thermal effect exists, correlates with the failure behavior, but does not cause it.

Since an effect (but not the prior cause itself) had been localized, the next step was to revisit or reconstruct the Correlation List. To proceed effectively at this point, however, improved access to the cause-effect network around the specific defect or anomaly was needed. To gain this access, material was built using a custom metal mask. Failing wafers built with this mask routed their (anomalous) supply current through a middle metallization sandwich layer. A FIB cut then confirmed the thermal map's results. A typical cycle through the metaprocess took shape as

Possible Prior Cause of Known Failing Behavior (On-Transistor)	Predicted or Known Correlating Side-Effect of Prior Cause at Left
Normal transistor otherwise wrongly activated	Electrical anomaly will appear in fabricated IC
	Observed global I(V) will occur in analog circuitry
	Observed global I(V) will occur in digital circuitry
	Observed global I(V) will occur in custom circuitry
	Observed global I(V) will occur in standard cell circuitry
Abnormal transistor created and active	Physical anomaly will appear in fabricated IC
	Observed global I(V) will occur at the transistor
	Observed global I(T) will occur at the transistor
	Observed global I(V) will occur in custom circuitry
	Observed global I(V) will occur in standard cell circuitry

Table 3 Final correlation list related to observed anomaly

described next.

Correlation List: Localization can be executed without much elaboration. It should be understood and developed, however, as a series of loops through the Fig. 1 metaprocess. To enable this, the Correlation List indicated in Table 3 was constructed. Localization then continued based upon the Correlation List items identified.

Selection: I_{DD} current path in analog circuitry section.

E/C/C Question: Does a current path in the analog circuitry section exist, correlate with global failure characteristics, and cause them?

Measurement: All but the bottom (more resistive) refractory metal layer was removed mechanically (by probe scrub) and measurements made for possible resulting voltage drops.

Interpretation: Results exhibited no voltage drops along any V_{DD} metallization path to the analog circuitry.

Measurement: Focused Ion Beam (FIB) cuts iso-

lated current paths between the external V_{DD} pin and analog circuitry. Global I_{DD} was then re-measured.

Interpretation: External current exhibited original characteristics as a function of voltage (anomalous current was not eliminated by the FIB cut).

E/C/C Answer: A current path in the analog circuitry does not exist.

At this point, current localization was pursued on multiple devices by FIB and other means until it was traced, first to a region of standard cells, then to a half-row within the region, and finally to two similar standard cells in the row (Fig. 9). This region was broadly consistent with the thermal map in Fig. 8. Correlation of the two cells to the leakage was demonstrated by tying the local implementation of the reset line of each cell to the low-current state. These

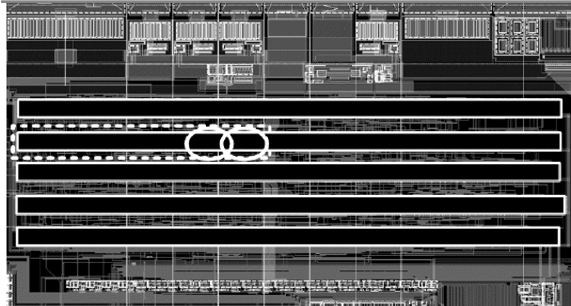


Figure 9 Standard cell region, half-row, and two cells containing leakage

FIB modifications produced approximately a 50% reduction of the anomalous I_{DD} leakage for each cell. To further localize the current path, the local reset within the cell was isolated before and after an inverter, and the post-inverter signal was found to correlate while the pre-inverter signal did not.

At this point the root anomaly was seen for the first time. The mask-generated layout in the area related to the post-inverter signal was inspected, and the structure depicted in Fig. 10 was observed. This figure has key structures annotated with three vertical polysilicon paths traversing an n+ guardring. These structures are not intentional (designed) transistors.

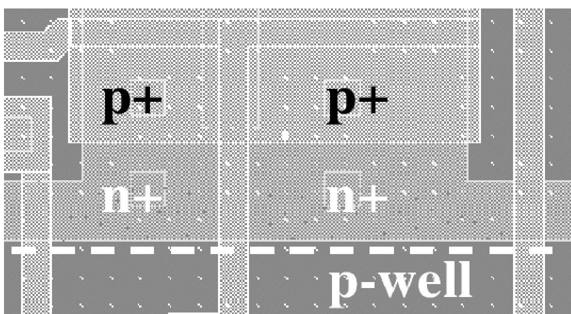


Figure 10 Key transistor-like structures common to both leaky standard cells

This structure stood out because it appeared to provide an unintended biased and transisting path from the p+ regions through a channel across the n+ guardring, and across the depletion region at the p-well boundary to the p-well region. Three anomalous factors (Fig. 11) interacted to create this path:

1. A mask generation algorithm created a very small extension of the p+ regions to the vertical polysilicon crossing the guardring (two instances, left and right);
2. An n+ guardring removed field oxide and created conditions for channel formation;
3. An apparent reduction in well doping expanded the depletion region to contact the channel area.

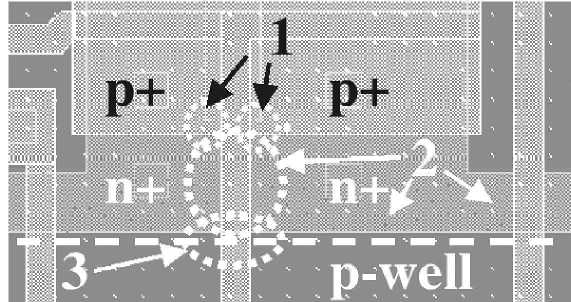


Figure 11 Components of anomalous leakage current path

At this point the failure was, to a large degree, considered localized. Since localization held the analysis substantially in place on the cause-effect network, however, it was appropriate to recall the analysis status in terms of the metaprocess loop. This review made it clear that some evidence now existed to address the E/C/C Question: "Does an abnormal transistor exist, correlate with the known V_{DD} leakage, and cause it?" That evidence suggested that such a transistor did exist, did correlate with the known V_{DD} leakage, and did cause it.

Here another opportunity to apply methodical logic arose, representing the third function of the metaprocess described earlier (providing application points for methodical analysis logic). Inductive Proof must be invoked to prove that a possible prior cause is an actual cause. When predicted side effects are always found, the decision to declare the PPC proved by induction rests entirely with the analyst. There is no objective criterion for the number of validated side-effect predictions. For some on the team, the number of validated predictions was simply not yet sufficient, and so additional side effects were sought that would remove any doubts about the anomaly's role.

Unfortunately, FIB access to further localize or exercise the local circuitry was not possible. The analysis therefore again turned to the Correlation List

Possible Prior Cause of Known Failing Behavior (On-Transistor)	Predicted or Known Correlating Side-Effect of Prior Cause at Left
Abnormal transistor created and active	Physical anomaly will appear in fabricated IC
	Transistor source, channel and drain occur in fabricated IC
	Other components create observed current threshold

Table 4 Second Cycle Correlation List, 1st Mod

(Table 4) to predict, select, and measure other secondary effects of the observed set of adjacent anomalies. The List made it clear that if a transistor source, channel, and drain existed in the region of the apparent anomalies, definite changes in the surface doping at regions 1 and 3 should exist.

The construction of this region virtually guaranteed that region 2 would exhibit characteristics of a doped channel, but it was apparent that much could be learned from comparing the first and third regions with the channel. (These statements all represent implications in support of inductive proof.) The Correlation List again led immediately to the E/C/C Question, "Do the transistor source, channel, and drain exist in the fabricated IC at the region of the anomaly, do they correlate with the observed failing behavior, and do they cause it?" A measurement was required to make these possible side effects visible.

Local high-resolution surface doping measurements of the region with Scanning Capacitance Microscopy (SCM)[10,11] on an Atomic Force Microscope (AFM) would serve the purpose. The SCM technique can provide very high spatial resolution mapping of semiconductor surface doping. This information, combined with precise AFM topographic measurements revealing field oxide extents, identified the ends of the parasitic channel, other local doping variations, and also depletion region widths.

These results (Measurement) are illustrated in Fig. 12, and confirmed the availability of a current path through the anomalies capable of producing the observed leakage current (Interpretation). The results satisfactorily demonstrated the "abnormal transistor wrongly activated" in the modified Correlation List. That is, they provided the last positive indicators demanded by the analysts' application of inductive

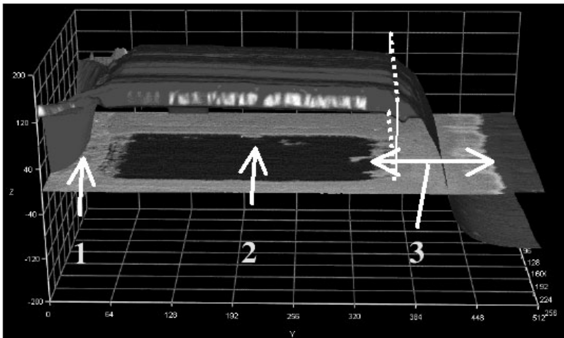


Figure 12 AFM surface (topographic) and SCM doping results (flat, doping shown through contrast) from suspicious region. Region 1: P+ region connects to channel through region that is at least depleted, and possibly inverted. Region 2: Channel region clearly reaches from the p-well at far right, extending leftward past the edge of field oxide (top topography surface edge delineated by vertical line extending to SCM data surface) to connect to the Region 3 channel (double-headed arrow at right).

proof to show that "an abnormal transistor structure at coordinates (x,y) exists, correlates with the original V_{DD} leakage, and causes it." (Final E/C/C Answer)

C. Corrective Action and Graphical Summary

This analysis successfully identified the anomaly causing the V_{DD} leakage. It remained to determine whether this anomaly also explained other initial failure characteristics. The low local power dissipation explained why no Liquid Crystal signal was seen. Photon emission was probably not observed due to equipment sensitivity to the mechanism. The high doping level of the old process explained why the same design fabricated in that earlier process exhibited no leakage. This probably occurred because the higher doping reduced the depletion region width (Fig. 12, Region 3) sufficiently to prevent contact between the depletion region and the channel. Finally, from Fig. 12 and the dependence on well doping, it is clear that the anomalous positive leakage temperature dependence is caused not through channel resistivity, but rather through characteristics of the depletion regions adjacent to the channel. During leakage, the greatest influence on current levels probably occurs in Region 1 of Fig. 12, where doping concentrations evidenced by SCM appear to be near intrinsic levels. Thermally generated carriers there will increase the carrier density in this region, lowering its resistivity and passing more current into the channel. Circuits modified to remove this structure exhibited no anomalous V_{DD} leakage, demonstrating that this analysis (summarized in Fig. 13) had identified the root anomaly.

Six Metaprocess Steps. Having completed the example of their application and use, let's look at the metaprocess from a more general, even an abstract, point of view. We must begin with a short description of the nature of cause and effect. Each phenomenon we observe is an effect of some cause—that is, it does not create itself. When we directly measure or observe an effect, we can call it a *known effect.* Any candidate for a cause of our known effect is a *possible prior cause*, or PPC. Other (initially unverified) effects of our PPC will be *secondary effects* in relation to the (primary) known effect. So in the above example, a possible prior cause of the known effect of V_{DD} leakage was an anomalous transistor in an on state. A secondary effect of that anomalous transistor was a quadratic I-V characteristic. The first known effect we encounter in any failure analysis is the failure's verified failure mode, our V_{DD} leakage in the example. We will term the first cause of failure in an electronic device the *root anomaly*. In our example, the root anomaly was the set of three regions created by the polysilicon routing. The cause of the root anomaly within the manufacturing process we will call the *process cause*. In our case above, the process cause was whatever permitted the anomalous transistor to be created. We don't use the term 'root cause', since it does not adequately distinguish between the first cause within the semiconductor device and the first cause within the manufacturing process. This approach and these concepts brought success in the above analysis example, but let us delve deeper into the specific application of these concepts to individual analysis decisions. To do this, we will start by

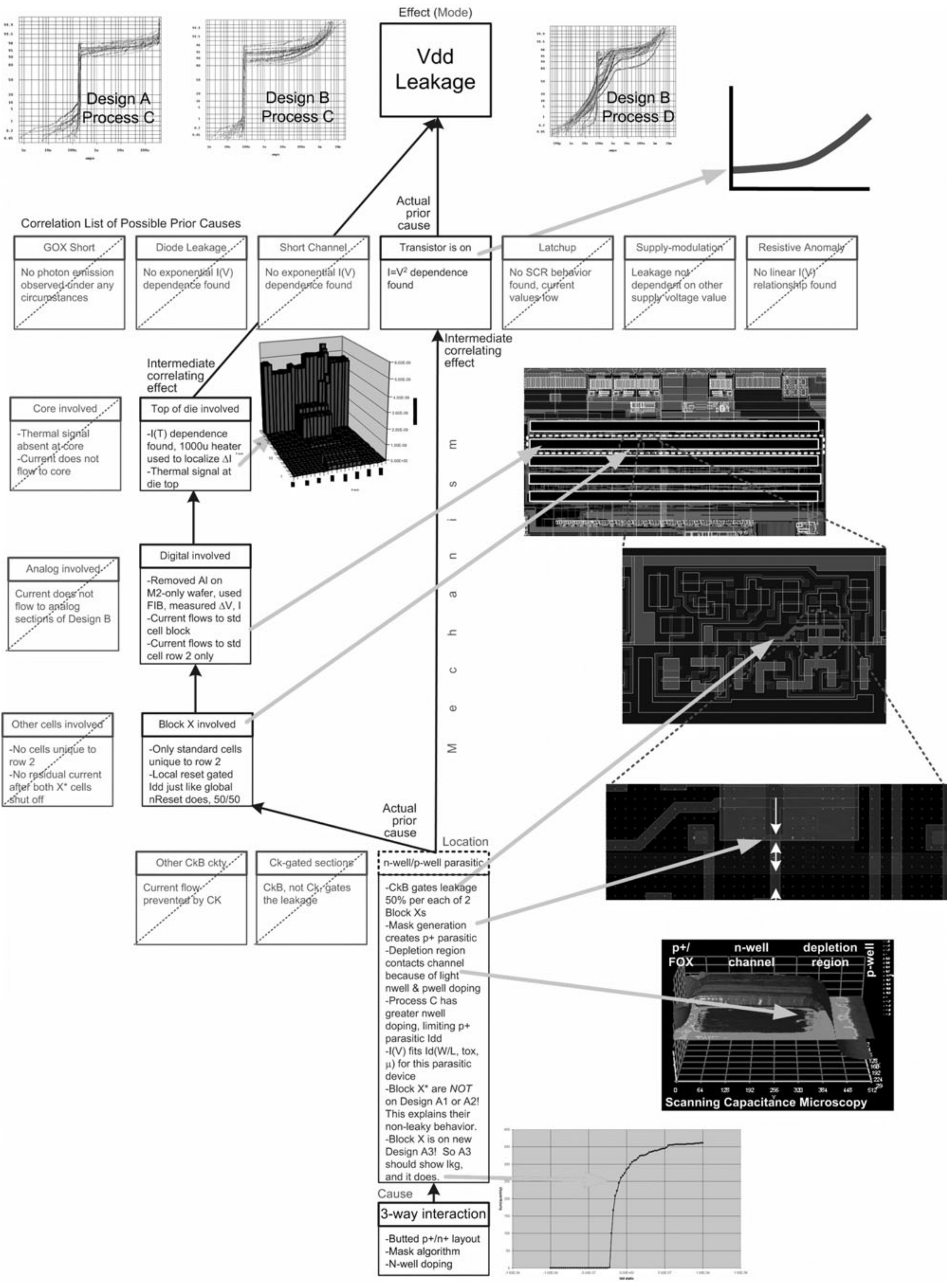
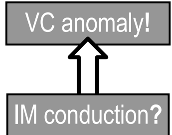


Figure 13 One-sheet full-analysis summary showing cause-effect relationships from root anomaly to failure mode, including intermediate cause-effect relationships and the analysis path that evaluated them.

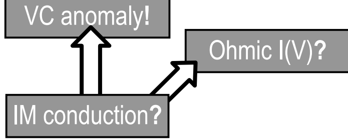
jumping right into the middle of a new, completely imaginary analysis.

E/C/C Question. Suppose we have found, through earlier analysis, a voltage contrast anomaly (that is, our failure differs from a good part in this way.) The

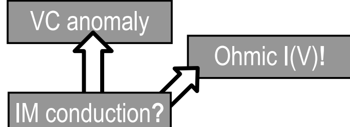


anomalous contrast--our known effect--could be a result of any of a number of possible prior causes, but let's select one: a conductive path between the metal trace showing the contrast and some other nearby metal trace. (The resulting short would change the trace's voltage, and therefore its contrast.) Our possible prior cause, then, is an intermetal conductive path, and the E/C/C Question step would lead us to construct this question: ***"Does an intermetal conductive path exist, correlate with our observed voltage contrast anomaly, and cause it?"*** Multiple examples of the use of the E/C/C Question appear in the first part of this tutorial. It will serve the reader well to review the above case study as this general description of the metaprocess describes each of the metaprocess's six steps.

Measurement. This second step uses analysis tools, observation, and logical reasoning to look for an effect of the possible prior cause--the intermetal conductive path--*other* than our known effect. In other words, what *else* would an intermetal conductive path do besides create our observed contrast anomaly? When we come up with such secondary effects, we measure them at this second step. We

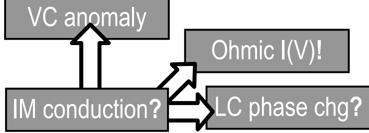


might, for example, measure I(V) current to adjacent lines and look for a linear relationship as an effect of an intermetal path. Such an ohmic I(V) characteristic between adjacent lines is a second, different effect of

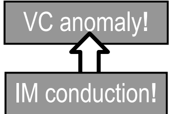


our possible prior cause.

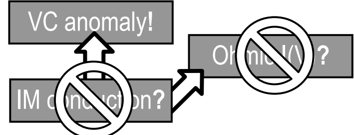
Interpretation. We next interpret our measurement results and determine what answer they give to our E/C/C Question. If we find the ohmic I(V) trace we just mentioned, we can keep our possible prior cause--intermetal conduction--as a *candidate* for the actual cause. We cannot yet conclude, however, that we have *found* the actual cause, since there are still



other possible prior causes which could explain both our observed contrast and also ohmic I(V) simultaneously. At this point, then, we loop back to the measurement step (see Fig. 1) and look for yet another secondary effect of the PPC. We may look for other



effects such as Liquid Crystal phase changes, current local to the output block or to the phase change, and so on, as we continue to loop. When we actually find enough of our predicted effects, we can conclude that

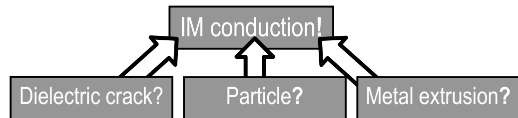


their common cause, intermetal conduction, exists.

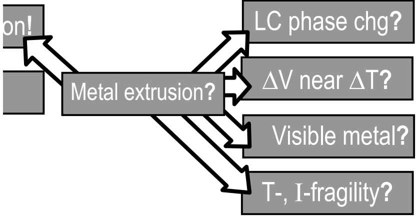
If on the other hand the ohmic I(V) trace is absent, then we immediately rule out an intermetal short, in accordance with a logic guideline called the Contrapositive Test, also discussed below.

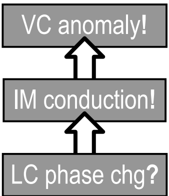
E/C/C Answer. After interpreting our measurement results and finding enough evidence for our prior cause, we formulate an answer that corresponds exactly in form to the question we posed.

If and when results support it, we a) conclude that the linear I(V) exists, b) demonstrate that the I(V) trace correlates with the node's leakage anomaly, and c) show by appropriate physical laws that any observed I(V) causes the leakage. We may find that our possible prior cause doesn't exist, or we may find that it exists but that it doesn't correlate with our known anomaly, or we may find existence and correlation but not find that our PPC causes the known anomaly. Or we may find all three, existence, correlation and causation. In this case, for our example we would answer, ***"An intermetal conductive path exists, correlates with our observed voltage contrast behavior, and causes it."*** This IM conduction now becomes our new known effect, since we have proved its existence and shown where it lies in the failure's cause-effect chain.



Correlation List. Now that we have a new known effect, we next create a brainstorm list of PPC's of the intermetal short. We might include lateral or vertical dielectric cracking, intermetal particles, oxide pinholes, electromigration-induced metal extrusions, metal overheating, and so forth. We also--and this is



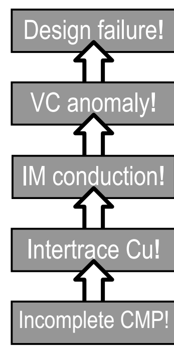


very important--generate an exhaustive list of other secondary effects of each of these possible prior causes. For example, along with our metal extrusion item we add to the Correlation List expected secondary effects including local liquid crystal phase change, further local voltage changes, visible metal anomaly, fragility with temperature, fragility with high current, along with any other such effects we can think of. One thing to note is that visual evidence of an anomaly has no special place in this methodology: it falls equal in importance to any other equally-attested effect. *A well-built correlation list will include at least ten secondary effects of each PPC.* If it proves difficult to come up with many secondary effects, this indicates, among other things, the need for analysts to strengthen their understanding of the physics of device failure. Our measurement step above, however, shows how important these secondary effects are to a successful analysis.

Selection. In the final step of our loop we select one of the Correlation List items and use it to begin a new Decision Engine cycle. We make the selection based in part on factors such as whether our choice must be investigated destructively, whether investigation of the PPC or effect is easy or difficult, and how likely the item is. In our example, we might select local liquid crystal phase change as our effect, since it is likely, relatively easy to measure, and non-destructive. We are free to choose more than one list item, each starting its own new Decision Engine cycle. This methodology enables us to increase our productivity in this way while giving us a tool to keep track of where we are in each analysis direction.

E/C/C Question (next loop iteration). At this point we prepare our next E/C/C Question about this selection: ***"Does local liquid crystal phase change exist, correlate with the observed intermetal conduction, and cause it?"*** The Engine cycle begins again with this step. But what has happened during these six steps plus one? We have moved from one known anomaly, a difference in voltage contrast between a good part and our failure, to the actual prior cause for that anomaly, an instance of anomalous intermetal conduction. Our single loop has taken us down one link in the unique chain of cause-and-effect that the failure contains.

Moving from Mode to Root Anomaly. This pattern of steps, then, demonstrably creates an analysis flow which proceeds efficiently back down the cause-effect chain by which the root anomaly creates the failure mode. If we simply apply the Decision Engine loop, our analysis flow unfolds, and the failure's cause-effect chain relationships are revealed link by link. From the initial failure mode (question, "Does a failure mode exist on this design and correlate with



the complaint") and its verification (answer, "the mode exists and correlates with the complaint") through intermediate questions (for example., "do hillocks exist, correlate with the observed intermetal conduction, and cause the conduction?") and their answers, to the root anomaly ("does incomplete CMP exist in the region of the conduction, correlate with the observed conduction, and cause it?"), this metaprocess creates the analysis flow which leads the analyst from the mode, or the first detected effect of the failure cause, back through intermediate cause-effect relationships to the root anomaly. In this way the metaprocess reveals the entire cause-effect chain from root anomaly to failure mode.

Analysis Flow Characteristics. Rather than delivering a prepackaged analysis flow which may or may not apply well to the failure at hand, this metaprocess produces the analysis flow one step at a time based on actual failure characteristics. This makes the unfolding analysis flow very responsive to the most current information about the failure. The metaprocess also directs the analyst to explore applicable cause-effect relationships at each point, and points to needed measurement equipment and techniques, whether available to the analyst or not.

These characteristics provide two benefits. First, they prioritize the analyst's focus correctly, placing greater importance on an understanding of good and failing device behavior than on technique or tool operation expertise. Second, they encourage the development of an analytical lab infrastructure which serves the objective of finding causes of failure, rather than the other way around. Many labs unwittingly make the toolset primary. The analysis process then becomes in large part a way of "grading" failures to determine whether or not they will reveal their secrets to a lab's existing analysis tool and technique set. Failures which do not yield to existing analysis capability often merely land in the "Cause unknown" bin rather than driving the lab's capability forward until their causes are found. The metaprocess described here makes the need for missing techniques both evident and quantifiable, enabling labs to improve their equipment set with confidence in the value of the techniques developed or purchased. (The metaprocess benefited the V_{DD} leakage analysis by prompting the development of the topside airflow heater whose results provided the first localization of the leakage.)

Second, the flow's characteristics make the formal

use of inferential logic easier, so as to help avoid analysis pitfalls and improve success. The flow does so by breaking the analysis process down into segments small enough that their relationship to formal logic principles becomes clear. The following section explains this for three of the most important logic concepts, namely logical inference, proof by induction and disproof by contrapositive testing.

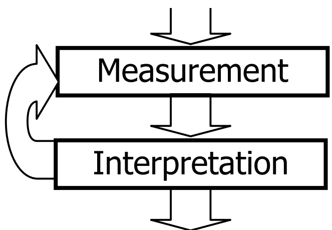


Figure 14 Metaprocess steps related to proving or disproving possible prior causes

Applying Formal Logic. Fig. 14 shows a segment of the full six-step loop. This segment relates to how we prove and disprove possible correlations in the course of a failure analysis. For example, if we have a functional failure mode which only occurs at positive supply voltages above 2.5V, we must look for possible prior causes which imply that specific voltage dependence. These may not be easy to identify. The apparent nonlinear relationship with supply voltage may suggest, however, that our failure mechanism involves active devices, and therefore may correlate with photon emission.

To investigate this possible correlation we identify implied characteristics of a common cause for the failure and the photon emission, then look for those characteristics. Simple but important logical implications--if-then statements--used here begin with the clause:

"If a common cause creates the functional failure mode and photon emission, then..."

and end with clauses such as,

"any observed photon emission will show an intensity change at 2.5V"; "temperature-dependent shifts in the failure onset will correlate with observed photon intensity change"; "emission pattern will correlate with the failing vector(s)"; "good devices will not show the observed correlations"

and so on. Each of these implications describes a cause-effect relationship; each possible effect provides something to measure.

After each measurement, an interpretation of the results indicates whether the predicted effect exists. As the list of observed effects grows, with no unsuccessful tests, the strength of the correlation grows with it, until the analyst can conclude that the common cause exists. After a sufficient number of measurements, each verifying a predicted effect, we consider the cause proved. This process is called proof by induction. (In actual analyses, we would be very specific about exactly what common cause--gate oxide leakage on the input of an and-or-invert block, threshhold voltage shifts in the region of a failing register, or poor transmission-gate design electrically close to the V_{DD} node--we wish to demonstrate.) Inductive proof--enough successful tests with no failures--represents a powerful tool for the analyst, in spite of its subjectivity.

Should we find on the other hand that a predicted effect of our possible prior cause is clearly and definitively absent, we conclude that the cause is absent. This needs to happen in principle only once, not repeatedly as with inductive proof. The logical principle by which we disprove such a prior cause is the contrapositive test. This test's principle states that if any cause positively implies a given effect, the real absence of that effect immediately disproves the existence of the cause.

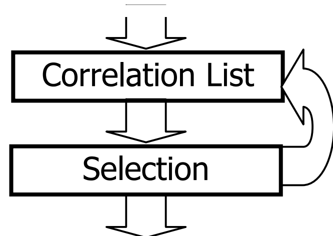


Figure 15 Metaprocess steps related to identification of PPCs

To reiterate, while it typically takes multiple iterations of measurement and interpretation to prove a correlation and its prior cause, in principle it takes the absence of only one predicted and sure effect to disprove a correlation or possible prior cause.

But our use of logic tools runs throughout the engine. On the other side of the decision engine loop are the Correlation List and Selection steps (Fig. 15.) Where the Measurement and Interpretation steps prove or disprove whether our possible prior cause is real, the Correlation List and Selection steps identify those PPCs in order to prove or disprove them. Logical principles applied at this step include, among others, logical implication, axioms of independent and dependent causes, and causation checking (identifying possible prior causes).

Avoiding False Paths. Logical implication, logical correlation or equivalence, inductive proof, and contrapositive testing represent four of at least twelve logical principles applicable at one or more steps in the decision engine loop. Space limitations prevent a full description in this tutorial, but others govern the logic of interactions, identifying failure characteristics by looking for patterns, handling multiple possible causes of a given effect, and justifying destructive steps. Each of these logical principles helps to identify false analysis paths and avoid all but brief trips down these paths. In the process, analyses succeed more often.

For example, inductive proof directs the analyst to test repeatedly, not just once, for a possible prior cause. Many effects such as photon emission and local temperature excursions have multiple possible prior causes, of course. These must be distinguished from one another during analysis by looking for their

other effects, as we described above in our description of the Interpretation step. To make a mistake in connection with this process means an unknowing

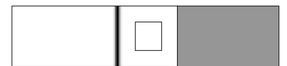


Figure 16 Voltage contrast change at contact

trip down a false path. Inductive proof leads the analyst to look for these false paths by finding the one prior cause whose effects all appear. In the process, the other possible (but not actual) prior causes are rejected. For example, suppose a particular signal never makes it from one block's output to the next input. A clear voltage contrast change on either side of the lone contact on the failing metal line (Fig. 16) leads immediately to the initial suspicion that the contact is open or high resistance. If inductive proof is not used, the suspicion may easily become a conclusion without further testing, which is a logically-incorrect, and risk-laden, approach. Alternatively, use of inductive proof would lead to more careful scrutiny of the downstream voltage contrast to look for harder-to-find evidence of the output signal, mechanical microprobing to test for DC current flow across the contact, and even photon emission tests to examine the behavior of the input stage to which the trace connects. If any evidence of low-frequency, low-amplitude post-contact signal exists, or photon emission in the input stage correlates with the failing

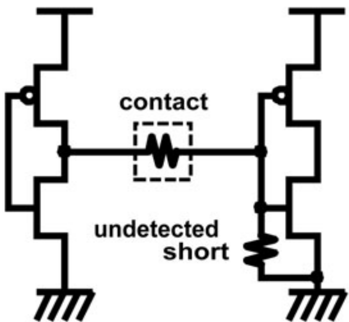


Figure 17 Alternative PPC for observed contact VC contrast change

condition, or especially if current flows normally through the contact, the analyst is immediately led away from the open contact prior cause, which is clearly in those cases a false path. A low-value short at the downstream input stage, creating a resistor divider with the contact, as illustrated in Fig. 17, may conceivably turn out to be more likely based on inductive proof's additional results.

But what might have happened if we had followed events in the pursuit of the incorrect open contact prior cause? A natural next step to evaluate an open contact would be a focused-ion-beam cross-section, clearly a destructive step. Other possible steps would produce a similar result. In fact, whatever the sequence of analysis operations chosen, the path eventually will include a destructive step. Therefore pursuit of the false path would sooner or later result in a destructive step performed prematurely and at the wrong location. This illustrates the point that destructive steps are generally not the cause, in and of themselves, of unsuccessful analyses. Some analyses, in fact, are composed almost entirely of destructive steps, with extremely successful results due in large part to proper and logical methodology[12]. Analyses fail rather because a false path is mistaken for the true path, and premature or misplaced destructive steps remove essential evidence for the actual cause-effect chain. The root anomaly then remains undiscovered in spite of all further analytical efforts. Good use of inferential logic within the metaprocess thus avoids improper destructive steps and increases the analysis success rate.

Analysis Speed. Execution speed may be the first concern that an analyst has with a scientific-method-based approach. This methodology addresses that concern first by efficiency. Unrecognized false paths waste more analysis time than any other time sink, not to mention the disastrous results if the analyst performs a destructive step. The remedy, testing for multiple effects of a PPC, costs far less in time than an unrecognized false path. Secondly, the method provides a level of organization that permits multiple decision engine cycles to operate in parallel without confusion, as described above. This means that the analyst can investigate several PPC's at a time and still keep track of the analysis easily. Finally, methodological insights into cause-effect versus effect-effect correlations and other characteristics of the main cause-effect chain enable the analyst to choose the fastest among several possible analytical routes to the root anomaly.

Uniform Analysis Representation. While providing the above benefits to analysis execution, speed and success, the metaprocess described here also generates a uniform, informative and compact way to represent an entire analysis flow. When each loop's Answer and Correlation List contents are written down, the analysis path revealing the main cause-effect chain, and also each investigated but disproved prior cause, appear in graphical form. A generic example of such a summary appears in Fig. 18. It may be profitably compared to the analysis summary of Fig. 13, which represents a specific instance of such a summary.

This representation lays a foundation for uniform presentation of analyses from case to case, analyst to analyst, and, for that matter, company to company. This in turn opens opportunities for better communication of analysis results and methods, for training of analysts that goes beyond informal or trainer-dependent approaches, and for database logging of analysis results to capture the essential aspects of any analysis.

Conclusion. Failure analysts who repeatedly find correct causes for difficult failures come to be known as clever analysts. While such reputations may be well deserved, cleverness of this sort can and should be formalized and taught to those not yet possessing

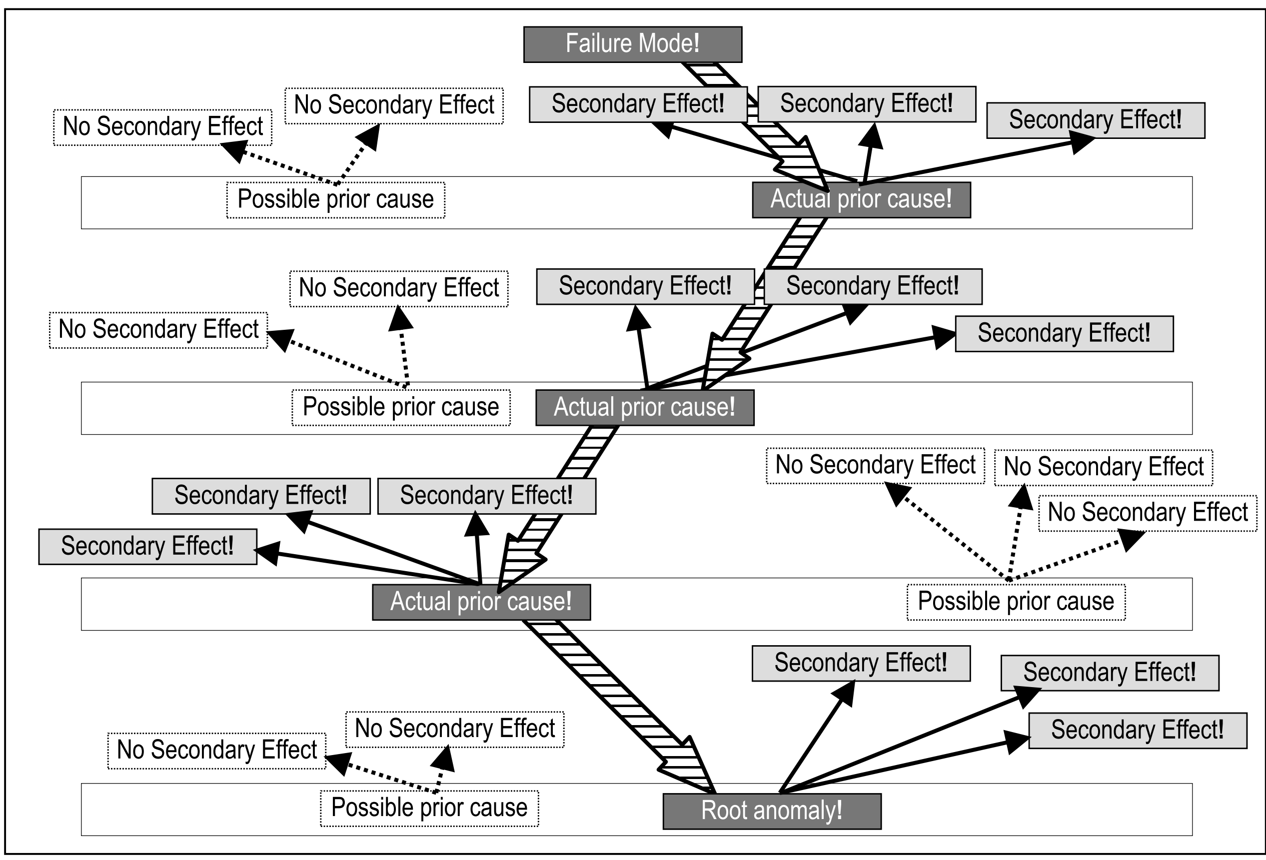


Figure 18 Generic diagram of analysis flow showing path to each actual prior cause through observed secondary effects, and elimination of incorrect possible prior causes by demonstrated absence of secondary effects. Note that each actual prior cause is itself an effect of an earlier cause in the chain. The prior cause which has no earlier cause in the failing integrated circuit is the root anomaly.

such a reputation. A well-constructed, careful analysis methodology provides a teachable formulation for the 'cleverness factor' (which consists chiefly, in this author's opinion, of discovering, understanding the nature of, and making use of the large network of cause-effect relationships which can lead to the failure's cause). It also provides a foundation for rigorous use of all available equipment during analysis. This tutorial has described a formal failure analysis methodology or Decision Engine which provides a speed-sensitive scientific approach for use in typical electronic-device manufacturing technologies. The approach is a metaprocess, a generic process which itself produces a specific process flow adapted to a specific failure analysis. It benefits the analyst by providing an effective custom analysis flow, keeping the analysis organized in the mind, making the approach as rigorous as required for the problem's difficulty, providing tools to optimize speed and analysis depth, and enabling a clear and standardized description of the analysis flow and outcomes.

References

1. Wagner, L., Failure Analysis of Integrated Circuits: Tools and Techniques, pp. 1-11. 1999: Kluwer Academic Publishers, Boston, MA.
2. Dicken, H., "A Philosophy of, and a Basic Approach to Failure Analysis," Electronic Failure Analysis Seminar Reference, pp. 13-25. 1998: ASM International, Materials Park, OH.
3. Pabbisetty, S., "Failure Analysis Overview," Electronic Failure Analysis Seminar Reference, pp. 3-11. 1998: ASM International, Materials Park, OH.
4. Ferrier, S. "A Standardized Scientific Method Approach for Failure Analysis Application," 2002 Proceedings of the International Symposium for Testing and Failure Analysis, pp. 341-347.
5. Ferrier, Steve, "Trying to See in the Dark," Electronic Device Failure Analysis News, May 2003, p. 25.
6. Ferrier, Steve, Martin, Kevin, and Schulte, Donald, Ph. D., Contributions of a Failure Analysis Metaprocess to Breakthrough Failure Analysis Results, 2003 Proceedings of the International Symposium for Testing and Failure Analysis, pp. 167-176
7. Gray, Paul R. and Robert G. Meyer, Analysis and Design of Analog Integrated Circuits, p. 64, eqn 1.175, John Wiley & Sons, New York, New York (1993).
8. Gray, Paul R. and Robert G. Meyer, Analysis and Design of Analog Integrated Circuits, p. 64,eqn 1.174, John Wiley & Sons, New York, New York (1993). Note that the temperature de-

pendence appears through k, which is directly proportional to the mobility

9. Grove, A. S., Physics and Technology of Semiconductor Devices, pp. 110, 347, John Wiley & Sons, New York, New York (1967).
10. Born, A., C. Hahn, M. Loehndorf, A. Wadas, Ch. Witt, R. Wiesendanger, "Application of Scanning Probe Methods for Electronic and Magnetic Device Fabrication, Characterization and Testing", J. Vac. Sci. Technol. B, 14, 3625 (1996).
11. C. C. Williams, "Two-dimensional Dopant Profiling by Scanning Capacitance Microscopy," Annu. Rev. Mater. Sci 29, 471 (1999).
12. Moss, J. M. et al., "Failure Analysis of Autoclave-Stressed SRAMs with Aluminum Fuses", Proceedings of the International Symposium for Testing and Failure Analysis, 1999, p. 293

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

System Level Failure Analysis Process: Making Failure Analysis a Value Add Proposition in Today's High Speed Low Cost PC Environment

Michael Lane
Intel Corporation, Oregon, USA
Roger Bjork, Jeff Birdsley
Dell, Texas, USA

Abstract

As an example of a viable system level failure analysis process, this paper describes a proven and effective model for addressing the Failure Analysis (FA) challenges that exist in today's Personal Computer (PC) environment.

The paper examines the factors and conditions that must be considered when assessing the FA scope and role, and its establishment as a critical value add requirement to maintain quality leadership and product performance expectations. The need for "Timely Actionable Information" becomes essential as part of the FA approach and a necessity to remain effective and competitive. This paper defines an approach by which this can be accomplished in the dynamic and high speed PC environment where product life cycles are consistently shrinking and the need for analytical techniques that yield "high confidence" answers quickly are becoming necessary.

The changing conditions of the PC environment are driving the PC Failure Analysis groups to respond and adapt by establishing an Integrated Failure Analysis framework. This framework may span multiple organizations or companies and must deliver proactive information in a timely manner throughout the product life cycle (PLC). The information is proactive in the sense that FA is now identifying issues and prioritizing which of those to analyze, versus the historical model of waiting for a request by another organization.

Introduction

Economic conditions and ongoing cost pressures are driving a need for Failure Analysis organizations to justify their role and highlight their contribution in today's rapidly evolving and changing low cost environment. In addition, quality and product performance expectations are becoming increasingly more challenging, requiring a highly engaged analytical role that is driving timely problem resolution and the support of ongoing continuous improvement efforts. The challenge facing Failure Analysis organizations is to establish a framework, processes, and techniques in order to meet these requirements. Further compounding this challenge are the multiple levels of integration required by the final product [Figure 1.] as well as the drive to outsource assembly and manufacturing services. All of these factors must be considered as part of the overall solution and require the need for an effective integrated Failure Analysis Model.

Integration Levels	Failure Analysis Activity	Owner
PC	Symptom Generation Fault Isolation to sub-system	System and sub-system Failure Analysis Group
M/B	FA & Root Cause Analysis Fault Isolation to component or Mfg Process	Motherboard Failure Analysis Group
IC	FA & Root Cause Analysis Fault Isolation to silicon or Mfg Process	Component Failure Analysis Group
Die	FA & Root Cause Analysis	Silicon Failure Analysis Group

Figure 1: Multi levels of Integration

The initial formation of an effective FA framework [Figure 2] needs to be based around a core team with knowledge of the final integrated product, the product use environment, and access to the required data to allow quantitative assessment of issues to be obtained.

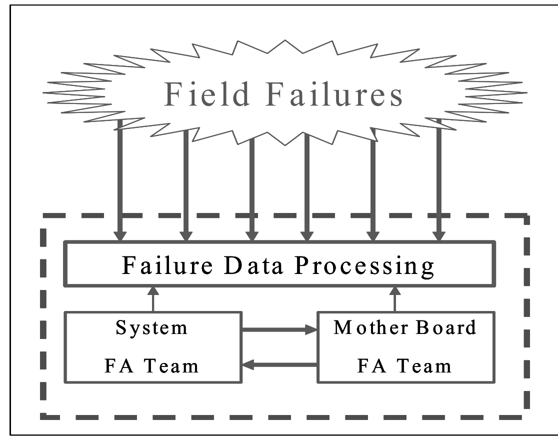


Figure 2. Initial FA Framework

The Personal Computer business, as with many consumer and enterprise level electronics businesses, is continuously being challenged to deliver greater

and greater performance and functionality. This is resulting in a more dynamic and constantly changing product portfolio, resulting in shorter product life cycles and faster product launches [Figure 3]. This in turn is driving faster issue resolution and quality improvements, resulting in extreme time pressure on Failure Analysis groups to establish an effective Root Cause Analysis.

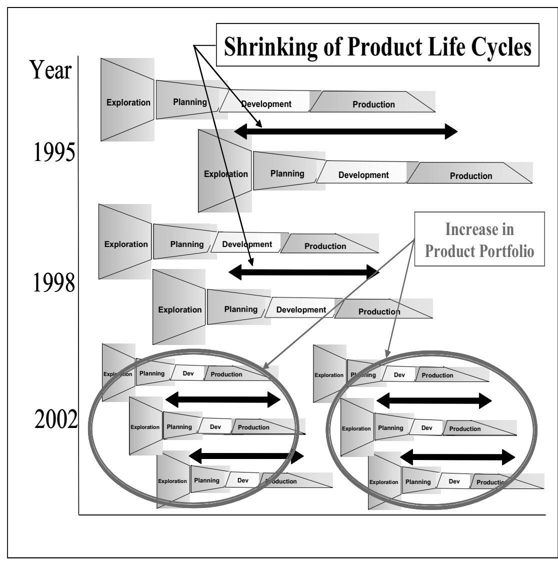


Figure 3. Shrinking Product Life Cycle

Actionable Vs Random Failure Analysis

The value add proposition for Failure Analysis in this environment is the ability to supply information to solve problems, resulting in cost savings. The current business climate is requiring Failure Analysis organizations to engage in the definition and consolidation of like failure symptoms and subsequently establish a "Failure Selection" methodology that supports a focused analytical effort in support of an improvement objective. The improvement objective can be in response to several different conditions, but are fundamentally in support of three main categories [Figure 4]: an Excursion Event; an Emerging Issue; or an Improvement Effort.

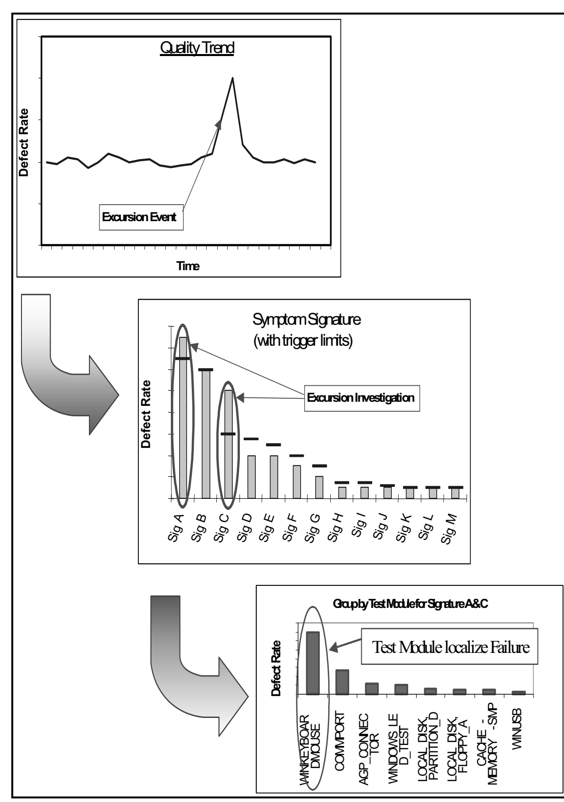


Figure4a Excursion Failure Selection

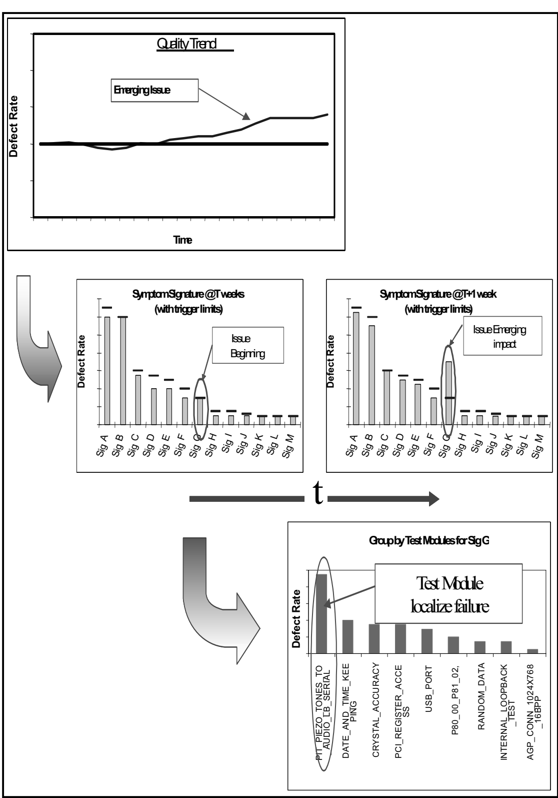


Figure 4b Emerging Issue Failure Selection

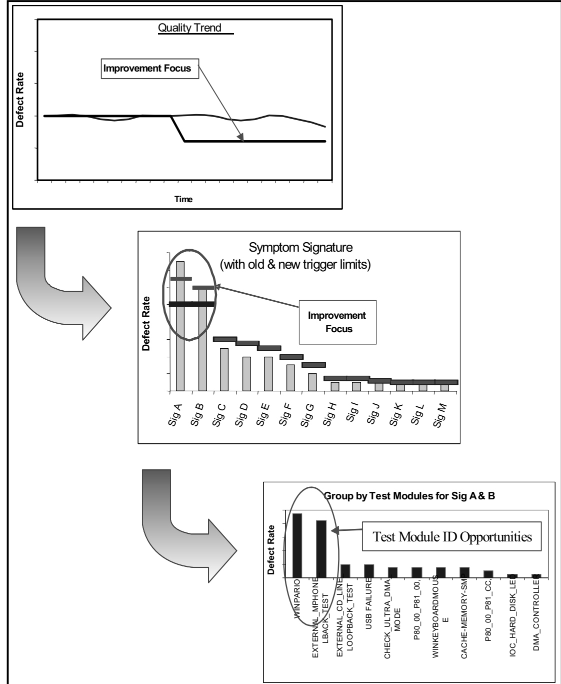
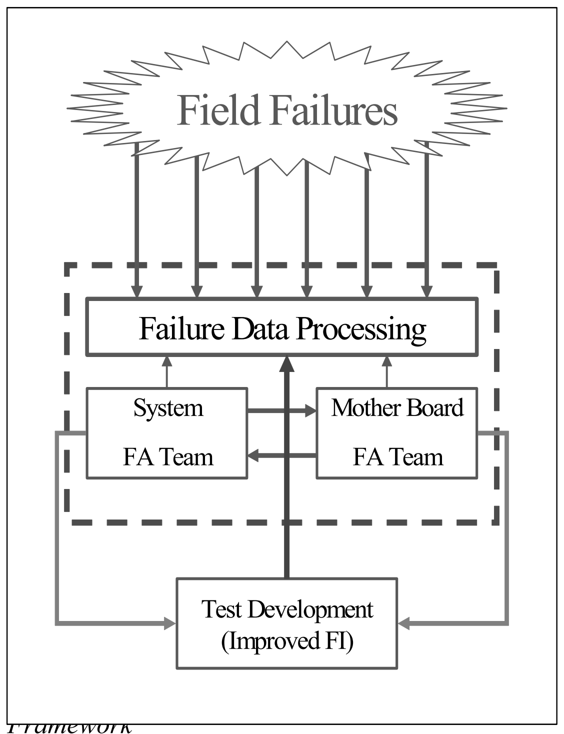


Figure 4c. Improvement Effort Failure Selection

To support these focus engagements it is required to develop a means to improve the odds of success in the selection of Failure Symptoms for failure analysis. This can be achieved through the integration of FA requirements into Product Testing solutions, which can enable a means to perform effective sorting that filters a targeted symptom and increases the odds of achieving the required result. Figure 4 exhibits how this is being achieved in support for the primary focused analytical effort mentioned earlier. This engagement can support the elimination of failures that are contributing noise to a unique and significant issue. This filtering can also assist in the isolation of a greater number of the same defects which opens up the analytical opportunities, i.e. can work destructive and non destructive analytical opportunities in unison which supports faster resolution.

To support this level of failure isolation, test development becomes a critical factor of the overall Failure Analysis Framework; this is best demonstrated in Figure 5

Failure Symptom to Failure Analysis
The “initial” failure symptom seen at a PC or system-level by the user must be isolated and associated to the affected sub-system (i.e. Hard Drive, Memory Module, Power Supply, Motherboard, etc.) with subsequent analysis required based on the sub-system specific Failure Symptom to drive the isolation of the next level of analysis. In the case of a Motherboard, this could be anything from the PCB to digital and/or analog circuitry. This can lead to individual sub-system functionality analysis and potentially the analysis of silicon. This hierarchy of analysis needs to be performed without the destruction of the failure signature in order to facilitate the next level of analysis.



Framework

Timely Failure Analysis

The compression of the product life cycle brings a criticality to the usability of failure analysis information. Actionable information is information that is statistically valid and comprehensive enough to drive effective process, component, or design changes. Achieving this validity in the shortest possible timeframe becomes critical to allow the information to be usable and acted upon. This timeframe needs to consider improvement implementation on the next product generation as well as the improvement interception on current product performance.

Failure Analysis Steps
The traditional role of Failure Analysis groups consists of two main steps: Fault Isolation and Failure Analysis (FI/FA). Fault Isolation is focused on defect localization and usually involves electrical

diagnostic techniques, while Failure Analysis involves physical and/or material analysis that establishes a failure mechanism, hence supporting failure root cause. These steps are usually sequential, and in the past were generally performed by the same FA group or organization. The drive to improve timeliness of information requires that steps be reconsidered as to how and when they are preformed. This problem is being solved through the development of the integrated Failure Analysis Team and the integration of failure analysis requirements into the Product Test solution. In addition, intelligent selection methods are being utilized to focus Failure Analysis on the areas that provide the most significant return to the business hence data systems and tools must deliver failure signature information to the solution providers as quickly as possible.

Product Testing, FA Engagement and Fault Isolation

The inclusion of Product Test as a part of the fault isolation process is changing the role of product test development groups to deliver solutions that not only validate the product quality, but also provide analytical direction. This is achieved by enhancing product test algorithms to output granular failure information that is utilized as the first step in the analysis process. This requirement is driving the need for a closer partnership between the Failure Analysis and the Test Development groups and the development of Design for Test (DFT) rules that enable the analytical factor to be supported. Effective characterization and sorting during the test process as well as careful evaluation of the data stream can support the isolation of significant events from random noise allowing the majority of the Fault Isolation activities to be completed even before the FA group have received the failed unit.

Root Cause Analysis

The need to perform detailed root cause analysis is well embraced by Failure Analysis groups today. The move from "what" is defective to "why" has been embedded in the FA way of thinking that supports design, engineering and material groups in the process of problem and issue resolution. Time constraints and failure opportunities are now challenging FA groups to look deeper into why something failed, to identification of what "triggered" the failure event in the first place. To address this challenge, failure analysis must take on the role of Failure Investigation. As in any investigative work, hypothesis development and testing becomes the key tools utilized in support of root cause analysis.

Failure Investigation

To support this approach, FA groups need to replicate the conditions at the time of failure and establish the critical parameters that were in existence at that time. This requires an assessment of the user's environment and details of the user's and product technical support's action. Based on this information, a defined Design of Experiments (DOE) can be developed that duplicates the failure event and supplies the necessary clue to what triggered the failure. Utilizing "Fault Injection" techniques, the engineering, design, or material group can apply the appropriate containment and long-term solution approach. This expansion of the Failure Analysis group's role has yielded significant improvement in both timeliness and effectiveness of the solution. This approach has enabled the implementation of solutions that have resulted in significant improvement within the current product cycle in addition to next generation product improvements.

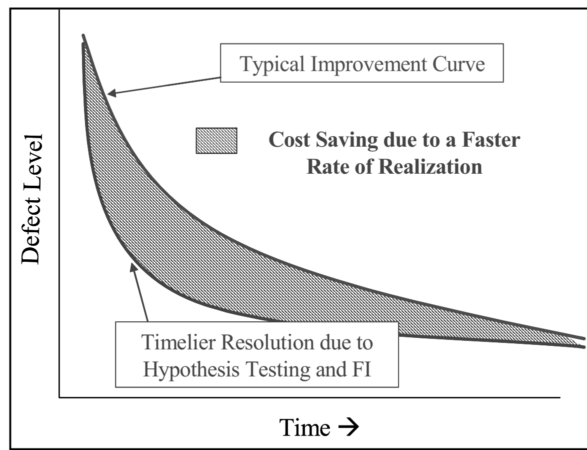


Figure 6. Hypothesis Testing and FI impact on Cost

Destructive Vs Non Destructive Analysis Techniques

Advancement in analytical capabilities has created several approaches that can be followed in support of the analysis of a failure. The probability of having several common failure opportunities is low requiring a systematic structured analytical approach [Figure 7]. The need to maintain the failure signature for as long as possible requires the utilization of analytical tools that do not destroy the failure. The development of capabilities such as X-ray, TDR, and Infrared Thermal Imaging has greatly assisted in pinpointing failures prior to the use of destructive techniques such as X-section analysis, component removal/remount, etc.

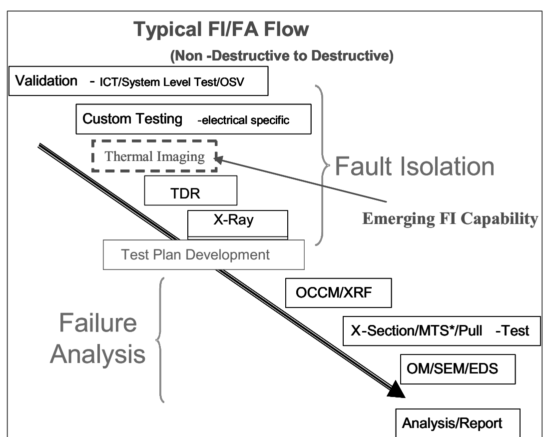


Figure 7. FI/FA Approach

Infrared Thermal Imaging [Figure 8] can be used as a first step in electrical failure isolation. It is a very effective non-destructive method for quickly isolating fault locations of resistive failures. IR thermal detection systems can produce a thermal profile of an entire motherboard or sub-assembly at one time. Applying image subtraction techniques to resulting profiles allows quick comparison of test assemblies against "golden" units isolating suspect areas for further, more in-depth analysis. This approach can be taken a step further in cases where thermal failure signatures are well understood. IR Thermal Imaging can provide confirmation of a suspected failure mechanism potentially eliminating the need for further analysis.

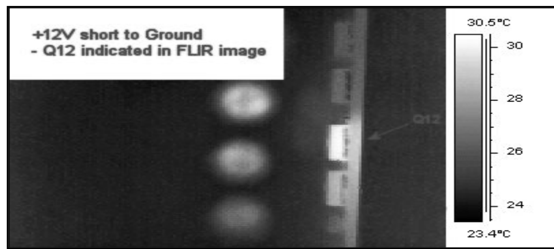


Figure 8. Infrared Thermal Image of a shorted FET.

X-ray has been widely used in the past as an effective tool for diagnosing solder shorts under BGA-type packages. However, lately X-Ray Laminography is becoming widely utilized for the detection of open or marginal joints. The utilization of this capability can quickly determine an effective analytical approach while mitigating the risk of incorrectly destroying the failure signature [Figure 9].

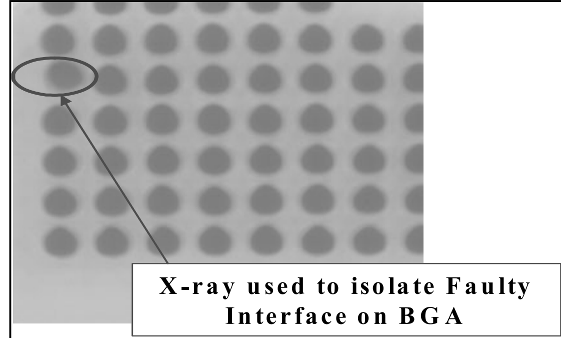


Figure 9. X-ray Laminography used for Non-Destructive Analysis of a BGA Joint

Time Domain Reflectometry (TDR) was developed primarily as a characterization and validation tool for package and component interconnects. However, it is establishing itself as a primary Failure Analysis capability that allows failure locations to be quickly isolated with very little preparation and reduces the risks associated with more conventional Failure Analysis methods [Figure 10]. These capabilities have allowed board level analysis to improve its detection rate and have greatly increased the probability of successfully developing a root cause and an actionable result.

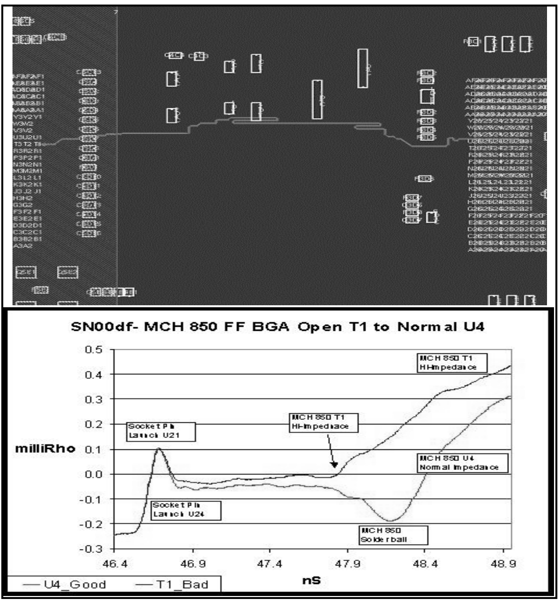


Figure 10 Time Domain Reflectometry (TDR) used for Non-Destructive Analysis of a BGA Joint.

The utilization of these techniques to identify root cause and develop actionable data allows the corrective action team(s) to respond in a timely and effective manner. The development of Corrective Action and the implementation of a solution

complete the cycle and deliver the result that was identified as part of the FA Framework [Figure 11].

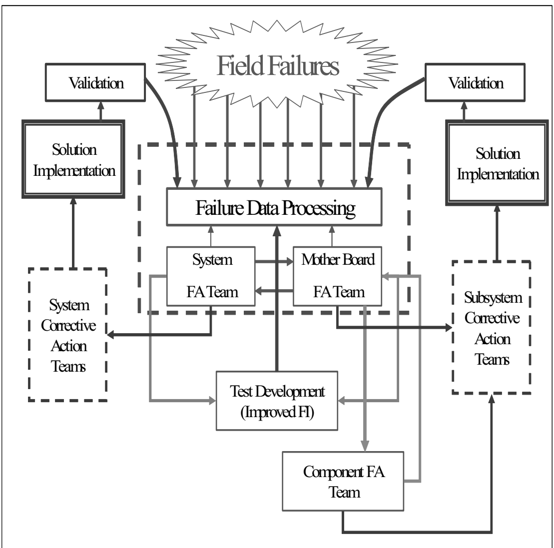


Figure 11. FA and CA Framework

Cost Factor

The implications of the bottom line and the ongoing cost pressures in the PC environment are driving the Personal Computing Business to investigate all aspects of the business model in the hunt for cost savings. Warranty costs are viewed as a prime source of cost savings opportunities even in the existing environment of continuously increasing quality and service expectations by customers. Warranty cost reduction is achieved most effectively through the utilization of an integrated failure analysis model [Figure 11].

Cost Saving Opportunities

Cost saving opportunities can be categorized in the following manner.

1. Cost Saving of Future Product Generations due to design manufacturing processes and/or material changes developed from analysis of the current generation product [Figure 12].
2. Cost Saving of Current Product Generation due to design, manufacturing and/or material changes based on timely actionable analysis and information resulting in an improved rate of savings realization [Figure 6].
3. Cost Saving across the installed base due to proactive service strategies based on analysis of the current product generation.

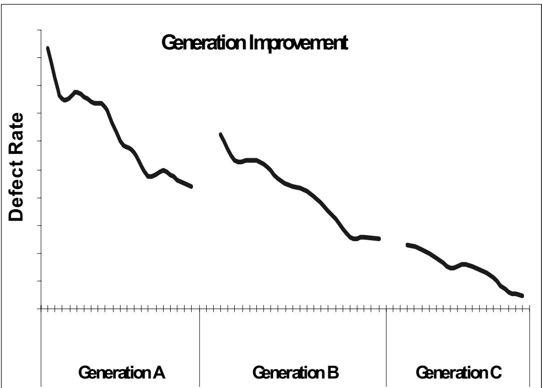


Figure 12. Defect Rate Generational Improvement.

With the development of "Timely Actionable Information", Failure Analysis groups are able to translate cost avoidance opportunities into Realized Cost saving. The cumulative savings achieved across all three cost categorized has fully justified the utilization of Failure Analysis groups and is quickly establishing FA as a value add role in the PC environment [Figure 13].

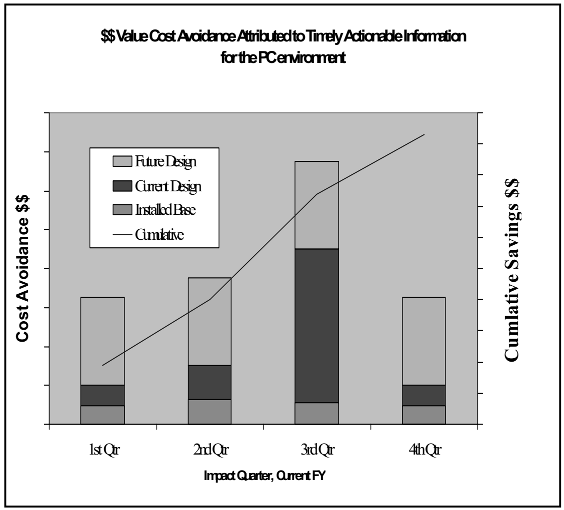


Figure 13. Cumulative Cost Saving by Product Generation

Summary

Through the establishment of an Integrated Failure Analysis framework and the utilization of unique processes and techniques, critical focus can be applied to a very dynamic and cost driven business in a fashion that has significant Value Add and impact to the bottom line. The need for emphasis on "Timely Actionable Information" is apparent and requires Failure Analysis Groups to step outside the

traditional boundaries and engage in areas of Product Test, Intelligent Failure Selection, Failure Investigation and Root Cause Analysis to enable resolution of failure trends in a timely and effective manner that maximizes the rate of cost saving realization. The application of Non-Destructive Analytical techniques to maintain the Failure Signature, and the engagement of Product Test Development Groups, to aid in early Fault Isolation, are critical factors to making this approach successful.

Generational Improvements and Defect Rate reduction results are confirming the Value Add this approach is bringing to the Personal Computing Business and is an example of how intrinsic and valuable FA groups are becoming as they step beyond the traditional FI/FA role.

Conclusions

The role of Failure Analysis groups has changed in response to the dynamic and increasing cost pressures of the Personal Computing business. Failure Analysis groups must establish integrated frameworks and working models that support increased effectiveness and responsiveness with the primary goal of delivering data and information that is timely and actionable and can quickly be used to drive issues to resolution. Results have shown that utilization of Product Test solutions for failure selection and early fault isolation allows issue resolution to impact current and future generation products and by so doing support the realization of cost avoidance opportunities.

Acknowledgements

The authors would like to thank the following people for their important contributions:
Rebecca Stuckman, Paul Hamilton, Pete Nicholas, Norman Armendariz, Christian Dieterle, John Woo, Khaled Jafar.

References

1. **"Parameter Extraction and Electrical Characterization of High Density Connector Using Time Domain Measurements"** S. Pannala, A. Haridass, M. Swaminathan, —IEEE Transactions On Advanced Packaging, Vol. 22, No. 1, February 1999.

2. **"Guideline for Measurement of Electronic Package Inductance and Capacitance Model Parameters"**—JEDEC Publication #123, JC-15 Committee, October 1995.

3. "**Measuring Controlled-Impedance Boards with TDR"** M.D. Tilden, "Measuring Controlled-Impedance Boards with TDR," – Printer Circuit Fabrication, February 1992.

4. **"Solder Joint Reliability of BGA, CSP, Flip Chip and Fine Pitch SMT Assemblies,"** McGraw Hill (1996)

5. **"Quality Control and Industrial Statistics,"** – 5th edition, Irwin (1986)

6. **"Case Studies of IR Based Rapid PC Motherboard Failure Analysis"** J. Swart, J. Woo, R. Zumwalt, J. Birdsley, Y. Taffese - ISTFA Proceedings, November 2002

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Board Level Failure Mechanisms and Analysis in Hand-held Electronic Products

Sridhar Canumalla and Puligandla Viswanadham
Nokia, Irving, TX 75039, USA.

ABSTRACT

Mobile communication and computing with hand-held electronic products have seen exponential growth in the past few years. Both customer expectations and end use environment are very different in comparison to desktop, server or office based electronic products. The generic failure analysis processes for such products and board level failure mechanisms related to aggressive mechanical and electrochemical environments are presented to highlight the unique mechanisms operative in hand-held electronic products. Specifically, mechanical drop and twist-related failures are discussed along with corrosion and electrochemical migration phenomena.

1. INTRODUCTION

Computing and communication have become mobile, indeed pervasive. Increasingly, these portable consumer electronics devices are characterized by a higher density of packaging and by higher levels of integration. Athough these sometimes possess the computing power and capabilities of desktop and office devices, they are vastly different from a technological and reliability perspective. Concomitantly, some of the failure mechanisms are also different than those encountered in conventional communication or office computing realms.

The technology and business drivers influencing the failure analysis needs of the mobile and handheld products are first discussed briefly. From the printed wiring board technology point of view, the trend is towards a) thinner layers, b) extremely high density circuit lines and spaces (100 µm line and spaces), and c) finer vias (75-100 µm) that are drilled with non-mechanical means. This generates a need for embedded devices and stacked multi-layer vias. In terms of component technologies, lower standoff and lower profile packages with thinner silicon are being used. As the printed wiring board (PWB) real estate demands grow, three-dimensional device stacking with 5 to 75 µm thick silicon will be used more and more. In addition, circuit card assemblies will be double sided to maximize the circuit card density. The industry will be challenged to eventually migrate to concepts such as system-on-package (SOP) and system-on-chip (SOC), through elimination of some levels of packaging.

Owing to their portability, the environment in which handheld electronic devices operate is often more severe in

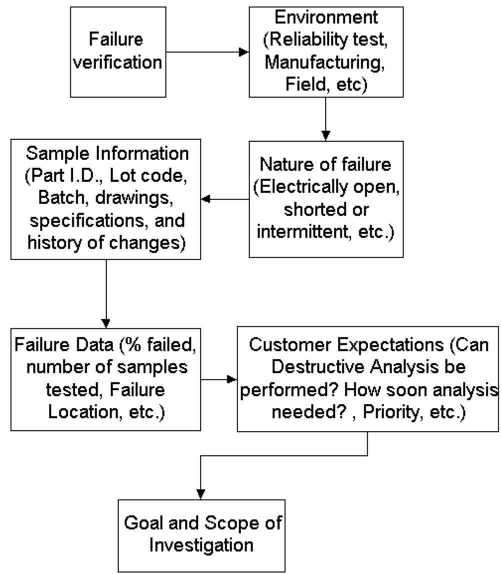


Figure 1. Pre-failure analysis process to determine the goals and scope.

terms of exposure to thermal excursions, humidity and corrosive environments. Additionally, portable products are much more prone to mechanical loads that include drop, bend, twist, etc. Further, owing to rapid developments in information processing technology, the product life of hand-held products is shorter than that of conventional office machines. The number of power ON cycles, the operational voltages, and other conditions will be different. Table 1 shows a typical comparison of the application conditions of portable hand-held devices with conventional desktop and automotive under the hood electronics.

Among the various loads that portable electronic hardware experience, the mechanical and humidity loads are perhaps more severe than others. Mechanical drops from heights of up to a meter and half on hard surfaces are not uncommon. Also, the products are much more prone to exposure to high humidity non-condensing and condensing atmospheres.

Table. 1. Comparison of typical application conditions of desktop, mobile, and automotive hardware (source JEDEC).

Desktop	Product Life (yrs)	Power ON Cycles /day	Power-on-Hours	Relative Humidity %	Enviroment Temperature Range/C	Operational Temperaure Range/C	Voltage V
Desktop	5	1-17	13,000	10-80	10-30	2-60	12
Mobile Terminal	5	20	43,800	10-100	-40-40	32–70	1.8–3.3
Automotive Under the hood	15	5	8200	0-100	-40-125	-40-125	12

In a high volume, low cost, competitive business environment, time taken to achieve sufficient product quality and reliability take on added significance. Cost efficient manufacturing and time to market need to be considered in addition to product quality owing to high customer expectations. In such an environment, the speed and effectiveness of failure analysis has a direct business impact and the analysis processes need to be optimized to meet business needs effectively.

2. FAILURE ANALYSIS PROCESS

The first step in the process of failure analysis often begins with defining the goal of the investigation and gathering background information. This process, sometimes known as pre-failure analysis, defines the scope of the failure analysis investigation. A typical sequence of steps in this pre-failure analysis process is shown in Figure 1. First, the statistical validity of the failure is examined to ensure that the effort spent in identifying a root cause is justified. Root cause analysis is sometimes described as the process in which the question “why?” is asked repeatedly until the cause of failure has been identified. This is sometimes impractical in enterprises that are not vertically integrated because the solution for the root cause for failure may reside a few steps up or down the supply chain. In such cases, root cause analysis may have to be redefined in terms of boundaries of direct influence, and the burden of further analysis transferred to the next level in the supply chain. Often, this means that first tier vendors will have to carry the responsibility of failure analysis even if *their* suppliers perform the actual analysis. For example, a consumer product maker may find that the root cause of failure is the poor quality of a component surface finish and refer the investigation to the component vendor. The component vendor may carry out further investigations and determine that the root cause of the poor surface finish is the poor incoming quality of the interposer laminates purchased from a second level vendor. Final root cause analysis may then depend on the investigations carried out at the

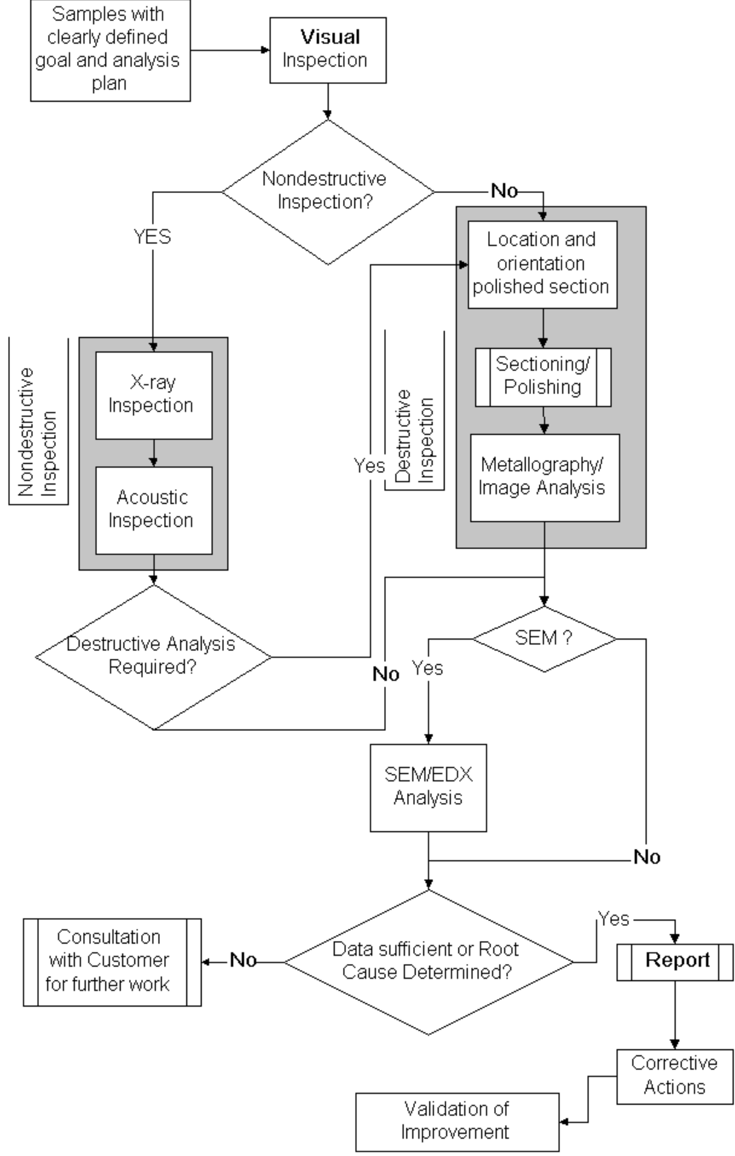


Figure 2. Process flow showing destructive and/or nondestructive analysis steps preceeded by visual inspection.

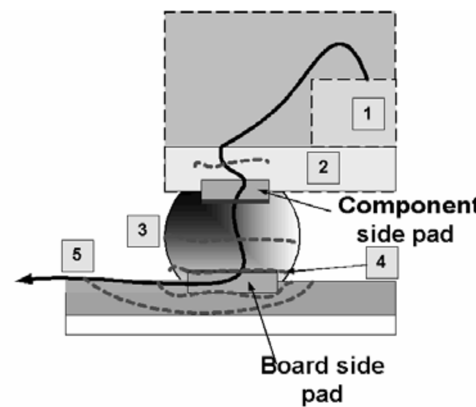


Figure 3. Simplified schematic of electrical interconnection from the Si die to multilayer PWB through different levels of packaging. Dashed line represents possible crack or open.

second level vendor's facility. It is not uncommon to find several levels of the supply chain involved in a failure analysis investigation, and the success of such a venture is dictated by good communication up and down the supply chain and the degree to which each level player understands the technologies and processes of other players in the value chain.

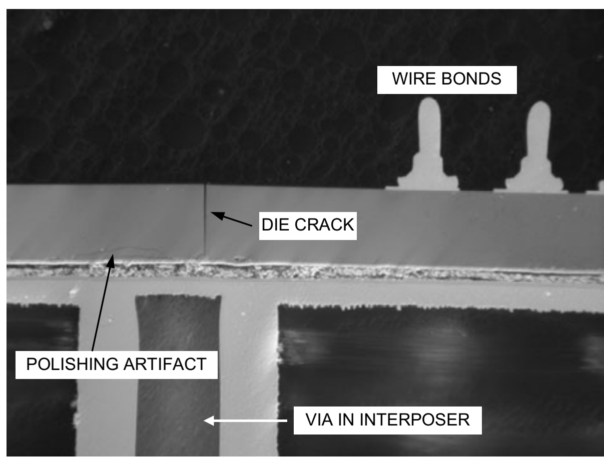


Figure 4. Die crack inside the package caused by PWB flexure

A basic failure analysis process is described in Figure 2. The tools involved in failure analysis include, but are not limited to, microscopes (optical, X-ray, acoustic, infrared, and electron), decapsulation and material removal tools such as ion milling, chemical analysis tools for volumetric and gravimetric analysis, spectroscopy tools (UV/Visible, infrared) and electron beam analysis tools (Auger, X-ray wavelength and energy) and laser induced mass spectroscopy. In addition to these experimental tools, simulation tools are also important in analyzing complex geometries and materials. Some representative failure mechanisms are discussed in the next section. The recognition that failure analysis is multi-disciplinary in nature is important. Expertise drawn from different disciplines of science and engineering needs to be utilized for successful determination of root cause.

3. FAILURE MECHANISMS

As mentioned earlier, the failure mechanisms in handheld electronic products are different from those commonly found in desktop or telecommunication environments. Broadly, they may be categorized as those caused by a) mechanical loading (including mechanical drop, vibration, bending and twisting loads), b) electrochemical environments that induce corrosion and electromigration, and c) thermal loading.

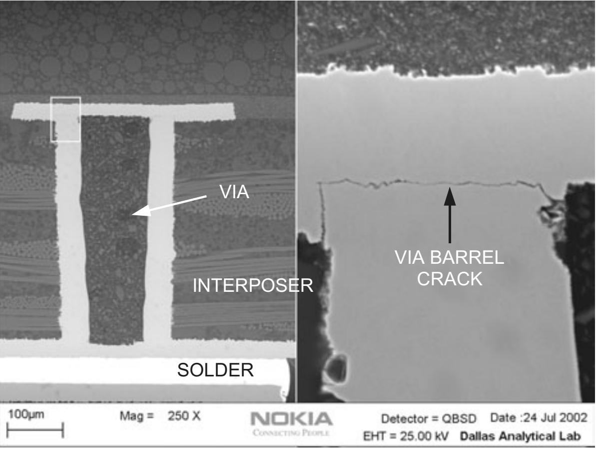


Figure 5. Via barrel cracking due to PWB level mechanical loading causing electrical failure.

In this article, the failures caused by mechanical loading and by exposure to electrochemical environments are discussed.

3.1 Mechanical Loading

It is instructive to review the construction of a generic package mounted on a printed circuit board before discussing the failure mechanisms. The printed circuit board (PCB) or printed wiring board (PWB) in portable electronic products serves not only as a carrier for the different electrical subsystems but also provides mechanical rigidity to the assembly. A typical PWB can have 4 to 12 electrical planes laminated between woven glass fiber reinforced epoxy layers that serve both a dielectric and mechanical support function. Electrical connection between these layers is often achieved through plated-through-hole vias, blind vias or buried vias. The outermost layer of the PWB, sometimes called the build-up layers, is the first interconnection layer between the solder joint and the PWB. Interconnection failures can occur at different levels as shown schematically in Figure 3, and can be classified as follows based on the location of the crack:

1. Die fracture within the package
2. Interposer level failure within the package
3. Solder joint fracture
    a. Crack initiation inside the component and subsequent damage to the solder joint

b. Solder joint fracture

c. Interfacial failure – at the solder/PWB pad interface

4. PWB related failure – trace fracture

5. PWB related failure - micro-via fracture

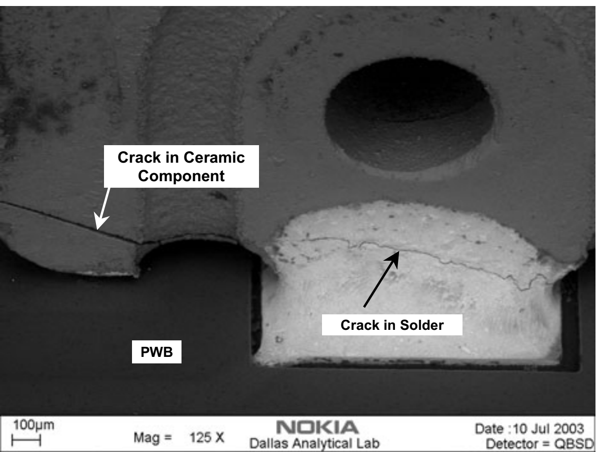


Figure 6. Crack in solder joint and ceramic component after mechanical shock (drop) reliability testing.

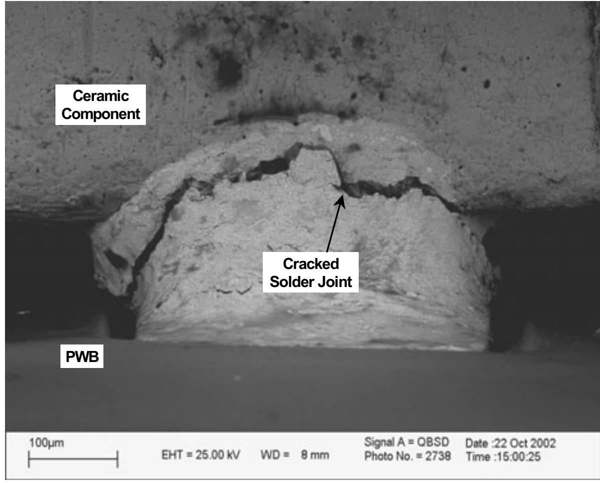


Figure 7. Crack in solder joint after twist testing.

3.1.1 Die fracture within the package

Sometimes, when the die is not supported optimally inside the package, almost all the flexure of the PWB can be transmitted to the relatively brittle semiconductor die inside the package. Cleavage fracture of the die can occur causing electrical failure. An example of this kind of failure is shown in the optical micrograph in Figure 4. The wire bonds on the die can also be seen along with the vertical crack in the die. The cracks at the inactive side of the die (bottom) are attributed to polishing damage during the grinding stage. Such artifacts have previously been observed in samples where excessive normal force was exerted on the sample, and should not be confused with cleavage type of cracking on the active side of the die.

3.1.2 Interposer level package failure

The Cu circuitry inside the interposer can sometimes fail if the process conditions in the fabrication of the interposer are not optimal. The example shown in Figure 5 illustrates the particular case where sub-optimal adhesion between the via-barrel and the via-cap failed upon exposure to mechanical loading at the PWB level.

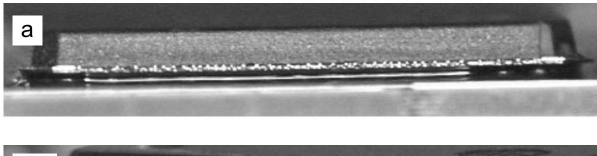


Figure 8. a) A partially underfilled CSP with a corner underfill void, and b) A more severe underfill defect exposing a whole row of solder joints.

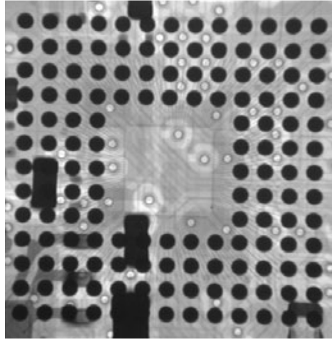


Figure 9. X-ray microscope image of a poorly underfilled CSP incorrectly indicating the lack of underfill defects. X-ray techniques can yield misleading results for certain kinds of defects.

3.1.3 Crack initiation inside component leading to solder joint damage

Ceramic components, due to their weight and lower fracture toughness, are particularly susceptible to failure when the product is dropped. Local stress concentrations on the ceramic component, such as those created by machining, can serve as crack initiation sites and cause premature failure as shown in Figure 6. The crack, that originated at the machining groove, caused an electrical open upon propagation. Apart from the machining on the ceramic component, a second factor contributing to the crack originating in the component is the relatively high strain rate of deformation during drop loading. Since solder deformation characteristics are highly strain rate dependent at room temperature, the solder joint is stiffer and stronger under higher deformation rates, thereby subjecting the ceramic component to proportionately higher stresses. For components operating at radio frequencies, a relatively minor partial crack, as shown in Figure 6, can sometimes cause parametric shift induced failures rather than a hard open.

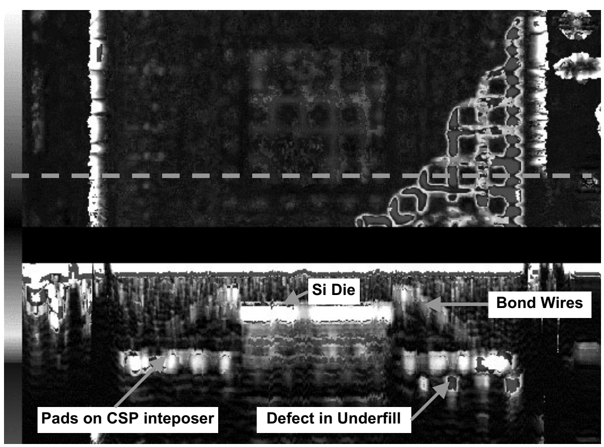


Figure 10. Acoustic image of the same CSP as in previous figure showing voiding in the underfill below the interposer of the CSP. A virtual cross-section (QBAM along the dashed line in the image) in the lower half of the image reveals that the underfill defect is below the interposer.

3.1.4 Solder joint fracture due to PWB level twisting

Bending and twisting are commonly encountered end use environmental hazards for hand-held products [1]. The deformation rates are much lower than observed in mechanical drop. In such cases, the solder joint strength and stiffness are proportionately lower and promote fracture at the solder joint in contrast to locations within the ceramic component. An illustrative example is shown in Figure 7, where the solder joint is completely fractured without the damage extending into the ceramic. The lack of machining damage in the vicinity of the solder joint was probably a secondary factor in limiting damage to the solder joint without cracking the component.

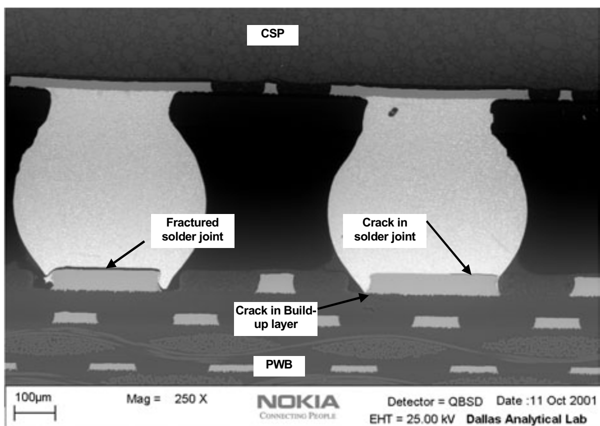


Figure 12. Scanning electron micrograph showing the fractured solder joint and concurrent damage at a neighboring solder joint.

3.1.5 Solder Joint Failure Related to Underfill Process

It is a relatively common practice to provide additional reinforcement to a solder joint to improve its reliability under thermal and mechanical loads. For ball grid array (BGA) and chip scale packages (CSP) soldered onto PWBs, this reinforcement can be achieved by the use of a suitable underfill material in the package to board interspaces. This constrains the assembly against bending and thermal strains. One of the more commonly used procedures for underfilling a CSP soldered onto a board consists of dispensing liquid underfill along one or more edges of the CSP perimeter such that capillary forces the underfill to fill the entire space between the CSP interposer and the PWB. Upon curing, the

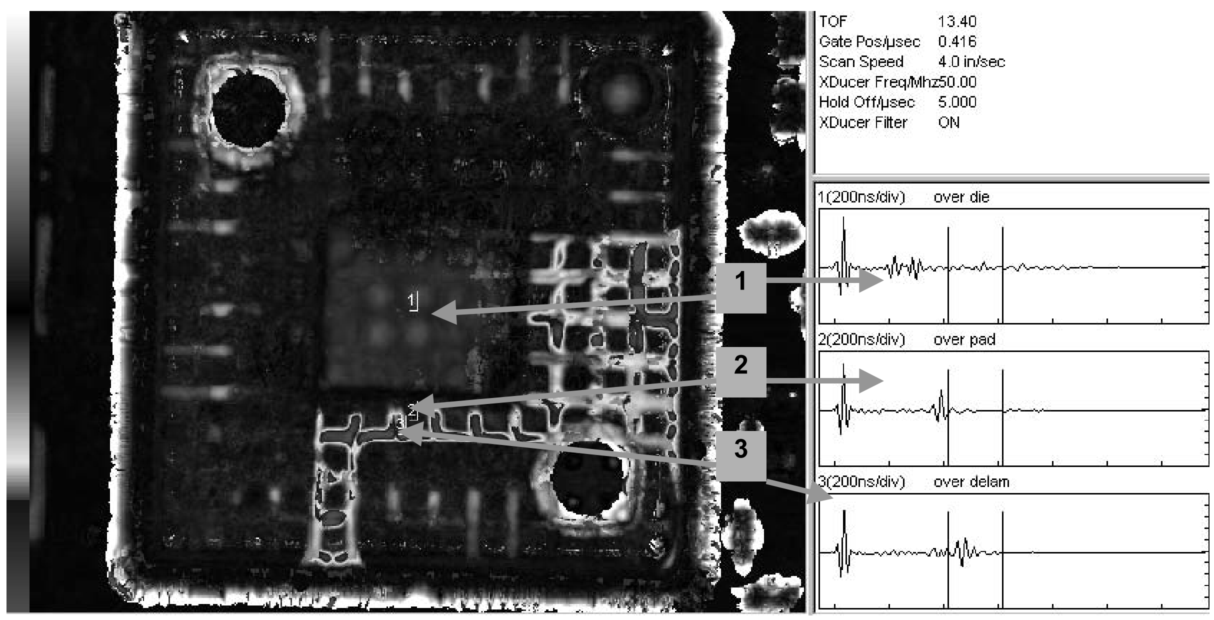


Figure 11. A more detailed acoustic image of a CSP with underfill defect showing the acoustic waveform traces over three locations: 1) the die, 2) the Cu pad on the interposer and 3) over the delamination.

liquid underfill hardens and encapsulates the solder joints completely, thereby providing additional reliability by mitigating the deleterious effect of either thermal or mechanical strains.

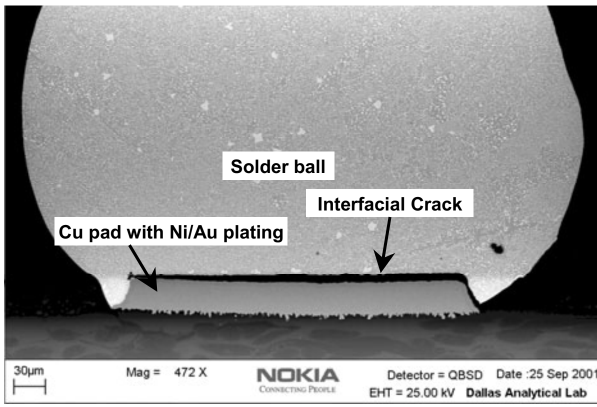


Figure 13. Interfacial fracture resembling brittle cleavage between solder ball and pad.

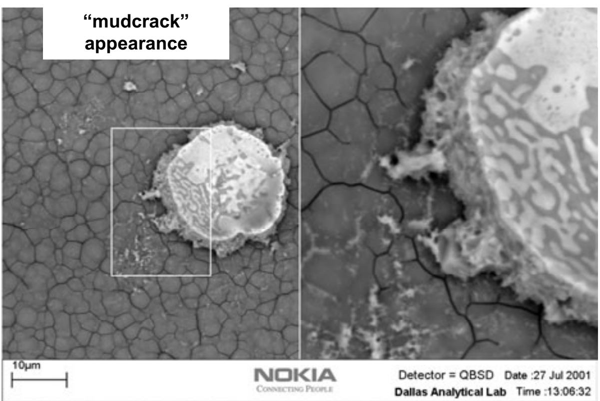


Figure 14. "Mud crack" appearance of Ni fracture surface showing the poor bond quality of solder to Ni/Cu.

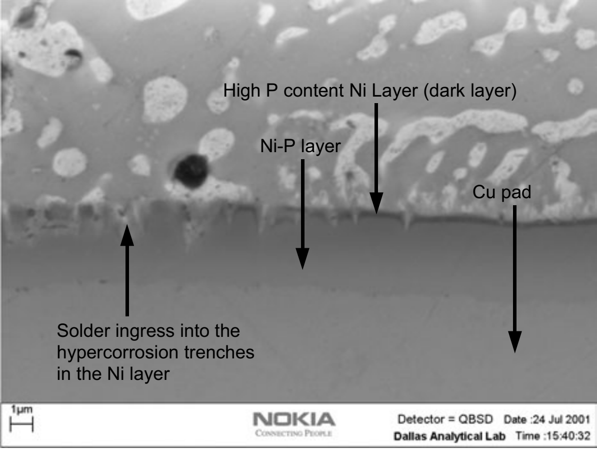
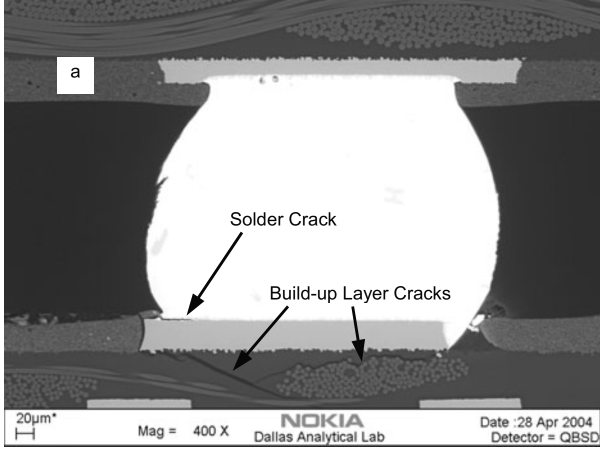


Figure 15. Hypercorrosion of Ni layer observed on a microsectioned sample with black pad defect.

The quality of the underfilling process is dependent on several variables such as temperature of the PWB or liquid underfill, cleanliness of the surfaces, speed of dispensing, etc. It has been shown that when the quality of the underfill is non-optimal and voids are present at the CSP corners, the benefit of the underfill is not realized even if the size of the void exposes only the corner solder joint [2]. An example of a partially underfilled CSP is shown in 8a and an optical micrograph of a more severe underfill defect is shown in Figure 8b.



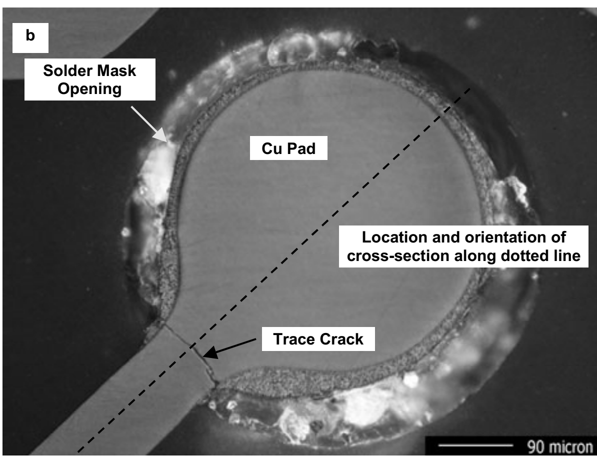
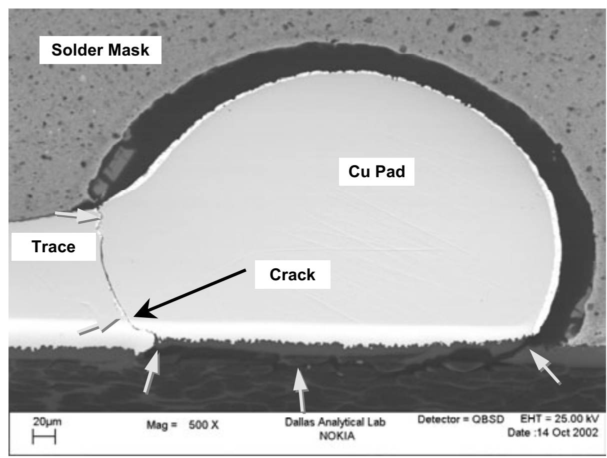


Figure 16. a) Build-up layer cracking in a solder joint with trace and b) Optical micrograph (top view) of a sample suspected to have a broken trace after the solder ball was removed by mechanical polishing. The dotted line represents the location and orientation of a second vertical microsectioning needed to show damage under the pad.

The true extent of an underfill defect cannot be ascertained either by visual or X-ray inspection. For example, Figure 9 shows a representative X-ray microscope picture of a CSP that does not reveal any underfill defect although visual inspection showed a substantial underfill defect at the perimeter. The scanning acoustic microscope, on the other hand, is very sensitive to voids and underfill defects. Difficulties encountered in acoustic inspection of CSP or BGA underfill include the signal to noise ratio due to material attenuation and uncertainty about the specific depth

that the data includes. Both these problems are particularly severe for CSP and BGA underfill, unlike in flip chip underfill inspection. A judicious selection of transducer frequency, F# (ratio of focal length to diameter of the transducer), depth of focus and gating are essential for successful inspection. The acoustic image of the CSP in Figure 8b is shown in Figure 10, and the areas of incomplete underfill can be clearly identified in the top half of the acoustic image. A virtual cross-section along the dashed line is shown in the bottom half of the acoustic image, and the relative depths of the die, the interposer and the void can be seen. In addition, the bond wires extending from the die to the interposer are also visible. It is also useful to present the acoustic waveform along with the image to clarify the nature and location of defects.



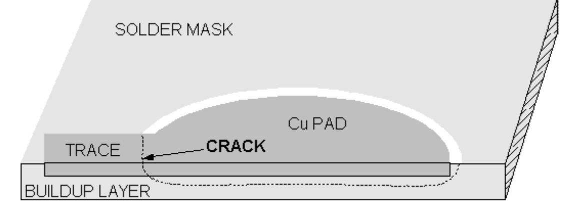
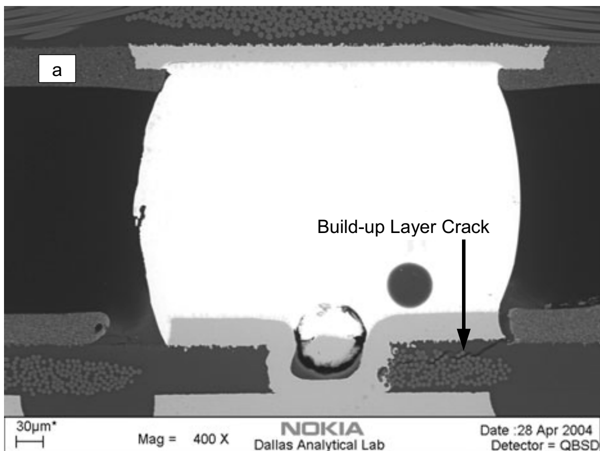


Figure 17. a) Trace fracture accompanied by build-up layer cracking revealed in a double-cross sectioned sample (sample shown different than that depicted in Figure 18), (b) schematic showing the location of the crack in the build-up layer.

An acoustic image of a different, improperly underfilled CSP is shown in Figure 11. The waveforms from three locations are presented alongside the acoustic image for ease of interpretation. The waveform from location 1 and 2 shows how the die and the interposer lie above the depth of the defect shown in location 3. The positive (upward) reflection from the top of the die and the Cu pads on the interposer are in contrast to the negative (downward) pulse from the underfill void. Thus, there is no ambiguity in concluding that the void lies below the interposer, where underfill would normally be expected in an underfilled sample. The lack of support for the solder ball can lead to failure of the interconnect that are now exposed to higher levels of loading. When exposed to adverse environment such as mechanical loading, the solder joints or the dielectric (build-up) layer below the Cu pad on the PWB can develop cracks as shown in the scanning electron micrograph of the polished cross-section in Figure 12.



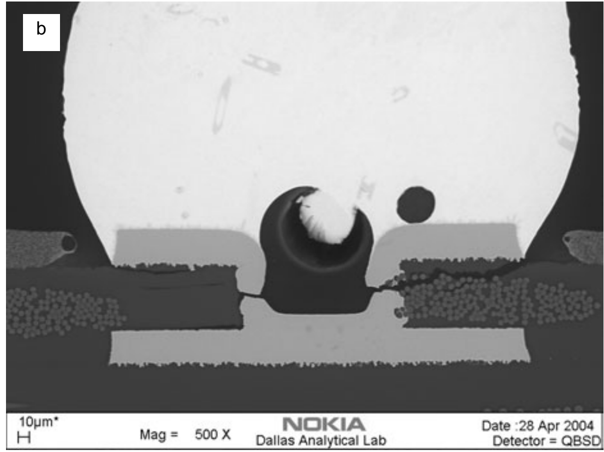


Figure 18. a) Build-up layer cracking in a solder joint with via in pad b) further damage leading to via cracking upon further exposure to mechanical drop related stresses.

3.1.6 PWB quality related fracture at Solder/PWB pad interface.

Electroless nickel/immersion gold (ENIG) plating of the Cu pads on the PWB has gained in popularity as a pad surface finish in recent years. This is because it provides a cost effective means of ensuring coplanarity, which is a crucial requirement in high density, fine pitch packaging. The Ni layer was intended to provide a barrier between the gold and the Cu pad. The very thin gold layer (< 0.5 µm) was intended to protect the Ni surface from oxidation and preserve its solderability until reflow. During reflow, when the solder melts and wets the gold surface, the gold layer dissolves instantly into the solder leaving a clean, solderable Ni surface for metallurgical bond formation. Not so long ago, the interconnection pad sizes were relatively large because fine pitch packages were not widely used. When pad sizes were relatively large, quality variations in the

ENIG plating did not immediately or always result in interconnection failures because the pad size-defect size ratio was substantial. Now, the pad size-defect ratio is smaller due to higher density of packaging. In addition to this, the increased use of less-aggressive organic solvent or water soluble fluxes can result in a pad surface that may not as solderable. These trends in the industry have increased the risks due to ENIG surface finish related failures that are characterized by a brittle fracture at the solder/pad interface along with a dull, dark Ni surface exhibiting "mud crack" type of surface morphology [3].

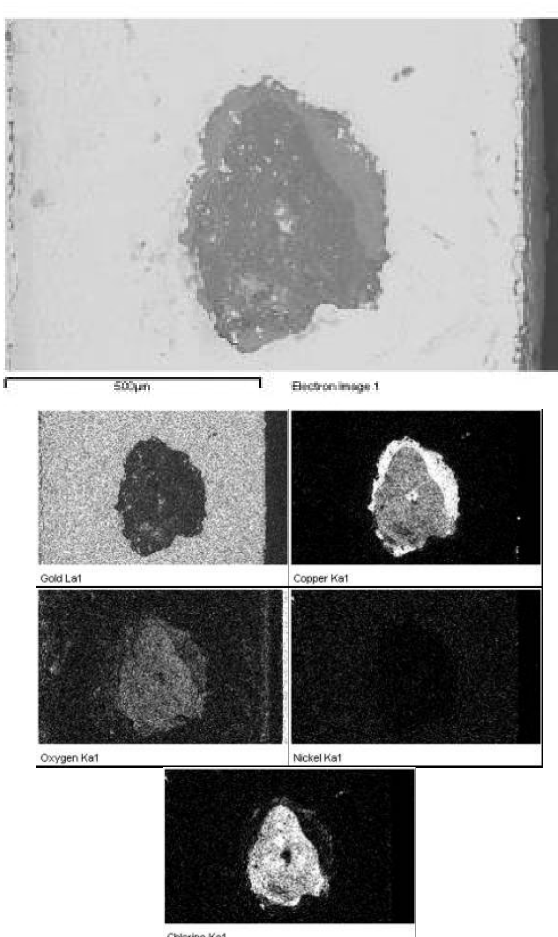


Figure 19. Corrosion of Au plated connector along with EDX elemental maps of Au, Cu, O, Ni, and Cl.

One example of the brittle interfacial crack at the solder-PWB pad interface is shown in Figure 14. The fracture occurs below the intermetallic layer at the solder/ENIG pad, which indicates that the interfacial bond between the Ni-Sn intermetallic layer and the underlying Ni layer is weak. Indeed, the fracture surface on the PWB side is often devoid of any adhering solder as seen in the backscattered micrograph in Figure 14. The only evidence of solder adhering to the surface is seen in locations where a void in the solder offered easier crack propagation than fracture at the solder/Ni pad interface. The characteristic "mud cracking" types of features are also visible on the Ni fracture surface. A high magnification micrograph of the polished cross-section (Figure 15) reveals the high-P Ni layer and a transverse view of the hypercorrosion trenches in the Ni layer where the solder has ingressed.PWB build-up layer cracking leading to Trace fracture

The PWB is usually a multi-layer laminate made up of layers of continuous, woven glass reinforced epoxy and Cu circuitry. The outermost layers sometimes referred to as the build-up or redistribution layers, serve both as the dielectric material and mechanical support for the Cu traces during PWB flexure or extension [4]. Upon subjecting the assembled board to mechanical loading, such as encountered in mechanical drop, damage accumulates in the build-up layer in the form of cracking. Subsequent damage accumulation and electrical failure will depend on the redistribution method employed.

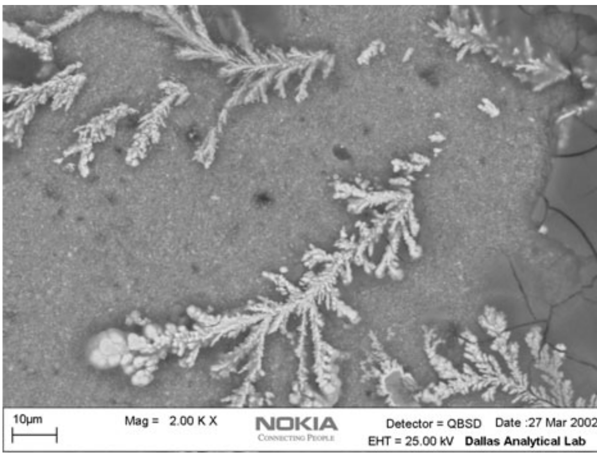


Figure 20. Cu electrochemical migration on a PWB subjected to damp heat exposure.

The progression of damage for the case when redistribution is achieved through traces is described below. Initially, the damage in the build-up layer accumulates until the trace is no longer supported because of extensive cracking under the pad. The damage in the build-up layer is exemplified by the back scatted electron micrograph in Figure 16a. This weakening of the build-up layer forces the trace to shoulder an ever-increasing share of the mechanical loads imposed on the PWB, which eventually causes trace fracture by fatigue processes. An example of the fractured trace is shown in the top-view optical micrograph in Figure 16b. The fracture process can be seen more clearly after a second microsectioning operation along the dotted line in Figure 16b. The backscattered electron micrograph of the double-polished sample is shown in Figure 17a and a schematic explaining the crack under the pad causing the trace fracture is depicted in Figure 17b.

The damage progression is similar in cases where a via-in-pad redistribution method is employed. The damage in the build-up layer, again, accumulates in the form of cracks, as shown in Figure 18a. Once the support afforded by the build-up layer is diminished by the cracking, further

mechanical flexure of the PWB subjects the via to increasingly higher stresses. Eventually, the via fractures as a result of fatigue, leading to an electrical open (Figure 18b).

3.2 Electrochemical environment

For portable and handheld electronic devices, two failure mechanisms related to electrochemical environments are of particular relevance – corrosion and electrochemical migration.

3.2.1 Corrosion

Corrosion is conveniently described in terms of half-cell reactions because these are electrochemical reactions. The electrodes in question could be on the macro- or micro-scales. Macroscopic corrosion cells can occur when dissimilar metals are coupled electrically and exposed to a corrosive environment, while microscopic corrosion cells tend to occur on the scale of grains. In either case, oxidation (and dissolution) occurs at the anode and reduction occurs at the cathode. The medium or electrically conductive environment in which these chemical reactions proceed is usually referred to as the electrolyte. Since all the electrons produced by the anodic reaction are consumed by the cathodic reaction, both anode and cathode reactions must proceed at the same rate for corrosion to occur in a continuous manner.

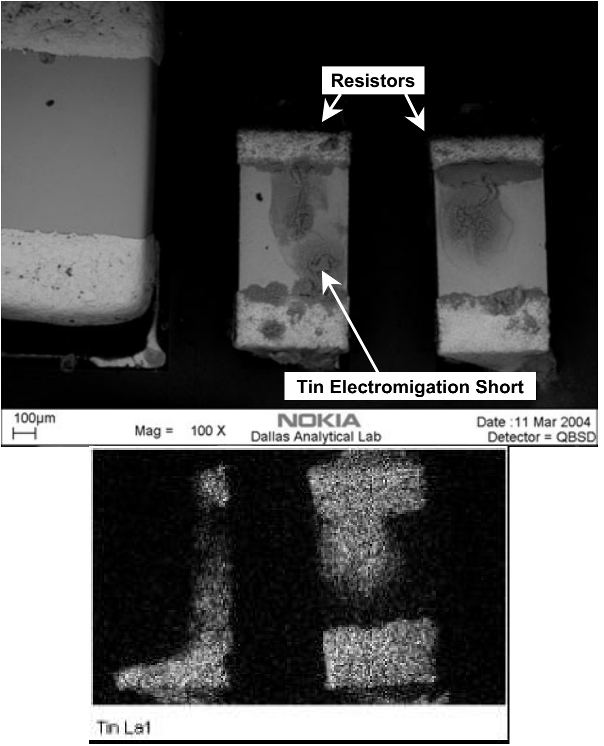


Figure 21. Tin electrochemical migration on a resistor with pure tin termination.

The propensity of a metal to undergo corrosion is often discussed in terms of the standard electrode potentials defined in terms of the standard Hydrogen electrode, which is arbitrarily assigned a value of zero [5]. When two dissimilar metals are coupled, the less noble metal will corrode in relation to the more noble metal. It is also possible to promote corrosion of the more noble metal in a galvanic couple by electrical biasing, which artificially makes the more noble metal the anode. Some forms of corrosion relevant to handheld and portable electronic products are

1. Uniform Corrosion – This form of corrosion is evenly distributed over the surface, and the rate of corrosion is the same over the entire surface. One measure of the severity is the thickness or the average penetration.

2. Pitting and crevice corrosion – This localized form of corrosion appears as pits or holes in the metal if the bulk material remains passive but suffers localized and rapid surface degradation. In particular, chloride ions are notorious for inducing pitting corrosion, and once a pit is formed, the environmental attack is autocatalytic. Crevice corrosion is also a localized form of corrosion closely related to pitting corrosion.

3. Environmentally induced cracking – This form of corrosion occurs under the joint influence of stress (either static or cyclic) and a corrosive environment. A static loading driven cracking is called stress corrosion cracking and a cyclic loading driven cracking is called corrosion fatigue. Residual stresses in package leads from lead bending operations have been observed to cause stress corrosion cracking failures in the presence of moisture [6]. Stress corrosion cracking of package leads was also reported in the presence of solder flux residues in South East Asia [7].

4. Galvanic corrosion – This type of corrosion is driven by the electrode potential differences between two dissimilar metals coupled electrically. The result is an accelerated corrosive attack of the less noble material. Galvanic corrosion tends to be particularly severe if the anodic material's surface is small compared to that of the nobler cathode. Galvanic corrosion includes cases where a nobler metal is coated onto a less noble metal. For instance, when a porous gold plating over a Ni substrate is exposed to a corrosive environment, the non-porous gold coating acts as a large cathode. This sets up a galvanic cell at the exposed substrate that causes intense anodic dissolution. It has been observed that a galvanic corrosion process can enhance pore corrosion when the substrate metal is less noble than the coating, and vice versa [8].

Gold plating of connectors is a common practice designed to protect the underlying Cu and Ni layers from corrosive attack and promote good electrical contact. However, under the action of friction the relatively thin and inert Au coating can be removed locally thereby exposing the Cu and Ni layers underneath. In such cases, fretting corrosion, pitting corrosion and localized galvanic corrosion can occur

simultaneously, especially in the presence of an ionic species such as chlorides. This corrosion product, which is usually nonconductive, can cause electrical failure due to opens or intermittent. An example of gold plated connector corrosion is shown in Figure 19. The EDX elemental map for Au indicates that the coating is intact over the major portion of the area of interest. However, in the central portion of the image, the Au coating appears to have been removed completely, and the underlying Cu is exposed. This Cu surface, identified as a bright area in the Cu elemental map, also shows significant presence of O and Cl. The absence of any areas with high concentrations of Ni indicates that the mating surface of the connector has probably worn through the Ni layer in the area of contact.

3.2.2 Electrochemical Migration

The distinguishing feature of electrochemical migration from corrosion is the formation of dendrites that cause a short between two adjacent conductors. The oxidation of the metal at the anode is common to both processes. Electrochemical migration, which is also known as migrated metal shorts [9], is probably best described as a transport of ions between two conductors in proximity to one another, under applied electrical bias and through an aqueous electrolyte. In general, three conditions are necessary and sufficient for electrochemical migration failures to occur, and they are a) presence of sufficient moisture (sometimes as little as a few monolayers), b) presence of an ionic species to provide a conductive path, and c) presence of an electrical bias to drive the ions between the cathode and the anode.

The first step in the classical model of electromigration consists of metal ion formation by anodic oxidation, which may either be direct electrochemical dissolution or a multi-step electrochemical-chemical process. The second step is the transport of metal ions from the anode, through an electrolyte, toward a cathode. In the final step, at the cathode, the positively charged ions are reduced to a neutral metal atom forming dendrites that grow towards the anode. An electrical short results when the filament touches the anode. Silver [9], Cu, Pb, Sn [10-12], Mo and Zn [13] have all been observed to follow this mechanism of electromigration.

A second mechanism of electromigration has been proposed to explain the migrated metal short formation involving noble metals such as Au, Pd and Pt. Because of the relative chemical inertness of these metals, a halogen contaminate is needed to induce anodic dissolution [14,15].

A third mechanism has been proposed to explain the electrochemical migration of Ni in the presence of a strongly alkaline electrolyte. Nickel electrochemical migration, which exhibits dendritic growth at the anode, can be explained by a process in which the first step is the formation of a cation by anodic corrosion followed by a chemical process resulting in secondary anionic species [16]. This anion complex migrates through the electrolyte under the applied electrical field. Finally, the metal atoms are deposited at the anode in the form of metallic dendrites due to the electrochemical reaction of the cationic species with the H^+, Ni^{2+} or OH^- ions. Similar processes could be operative for Co and Cu electromigration as well, in cases where anodic deposits of the metal are observed.

Copper electromigration

Copper forms complex species such as $CuCl^4$-, $CuCl_2(H_2O)$, $Cu(H_2O)^2$+, etc. in the presence of halide containing species and moisture. An example of Cu electrochemical migration resulting in Cu dendrite formation is shown in Figure 20.

Tin electromigration

Tin electrochemical migration mechanism is similar to that of Cu, but is much more prevalent because Sn constitutes a major portion of several commercial solder compositions such as 62SnPb2Ag, 10SnPb, Sn3.5Ag0.7Cu, etc. In addition, exposed Sn is more widespread on an assembled PWB as compared to Cu. The particular example shown in Figure 21 is from a test vehicle that failed upon exposure to damp heat testing. In this case, the potential difference between the terminals of a resistor with Sn termination resulted in the migration of Sn from the cathode towards the anode. It can be seen from the EDX elemental maps (Sn) that one of the resistors is shorted.

4. SUMMARY

A basic methodology and process flow for successful root cause analysis was provided. A variety of failure modes and mechanisms that exemplify first and second level electronic packaging of portable consumer electronic hardware haven been discussed.

As portable electronic hardware becomes more complex with increased data handling capacity, further miniaturization and higher levels of integration at all levels of packaging will be a natural trend with important implications for root cause analysis. More functions will be integrated into the components and the PWB. The silicon die thickness will be in the range of 40–50 µm. Stacked and/or folded packages will be more prevalent with multiple levels of wire and flip chip bonding. Another emerging trend in packaging is the 3-D integration at the wafer level. New materials that will have better mechanical properties and moisture resistance will have to be developed. More functions will be embedded into the printed wiring board and these may include both active and passive devices, attendant with new embedded interconnection schemes. The printed wiring board technology itself will witness significant changes with thinner and improved materials that are capable of 10–25 µm vias, 10-20 µm lines and spaces, and structures involving several layers of stacked vias. Consequently, hitherto unknown failure mechanisms are likely to be encountered.

As feature sizes diminish, the distinction between first and second level packaging can become nebulous. Failure analysis at the PWB assembly level will be a challenge. With shorter product development cycles there will be an increasing need for automated analytical tools with minimal

operator intervention to achieve rapid and repeatable root cause analysis.

ACKNOWLEDGEMENTS

The authors express gratitude to our colleagues Sesil Mathew, Sambit Saha, Murali Hanabe, and Laura Foss for reviewing this document and permission to include some of their data, and Timothy Fitzgerald for management support.

REFERENCES

1. Daya Perrera, U., 1999, "Evaluation of reliability of μBGA solder joints through twisting and bending," Microelectronics Reliability, 39, pp. 391-399.

2. Canumalla, S., Shetty, S., and Hannan, N., 2002, "Effect of corner underfill voids on chip scale package (CSP) performance under mechanical loading," 28th International Symposium for Testing and Failure Analysis, 3-7 November, 2002, Phoenix (AZ), pp. 361-370.

3. Biunno, N., 1999, "A root cause failure mechanism for solder joint integrity of nickel/immersion gold surface finishes, IPC Printed Circuits Expo, CA, pp. 1-9.

4. Canumalla,S., Yang, H-D., Viswanadham, P., and Reinikainen, T., 2004, "Package to board interconnection shear strength (PBISS): Effect of surface finish, PWB Build-up layer and Chip Scale package structure," IEEE Transactions on Components and Packaging Technologies, v 27, no.1, pp. 182-190.

5. Viswanadham, P., and Singh, P., 1997, Failure Modes and Mechanisms in Electronics Packages, Chapman and Hall, New York.

6. Guttenplan, J.D., 1987, Corrosion in the electronics industry, ASM Metals Handbook (9th ed), v13, ASM International, Metals Park, OH, USA.

7. Raffalovich, A.J., 1971, "Corrosive effects of solder flux on printed circuit boards," IEEE Transactions on Components, Hybrids, and Manufacturing Technology, v 7, no. 4, pp. 155-162.

8. Yasuda, K., Umemura, S., and Aoki, T., 1987, "Degradation mechanisms in tin- and gold-plated connector contacts," IEEE Transactions on Components, Hybrids, and Manufacturing Technology, v 10, no. 3, pp. 456-462.

9. Kohman, G.T., Hermance, H.W., and Downes, G.H, 1955, "Silver migration in electrical insulation," Bell Systems Technology Journal, v34, p. 1115.

10. Benson, R.C., Romanesko, B.M., Weiner, J.E., Nall, B.N., and Charles, H.K., 1988, "Metal electromigration induced by solder flux residue in hybrid microcircuits," IEEE Transactions on Components, Hybrids, and Manufacturing Technology, v11, no. 4, pp. 363-370.

11. Dermarderosian, A., 1978, "The electrochemical migration of metals," Proceedings of the International Society of Hybrid Microelectronics, p. 134.

12. Ripka, G., and Harsanyi, G., 1985, "Electrochemical migration in thick-film conductors and chip attachment resins," Electrocomponents Science and Technology, v 11, p. 281.

13. Kawanabe, T., and Otsuka, K, 1982, "Metal migration in electronic components," Proceedings of Electronic Components Conference, p. 99.

14. Grunthaner, F.G., Griswold, T.W., and Clendening, P.J., 1975, "Migratory gold resistive shorts: Chemical aspects of failure mechanism," Proceedings of the International Reliability Physics Symposium, p. 99.

15. Sbar, N.L., 1976, "Bias humidity performance of encapsulated and unencapsulated Ti-Pd-Au thin film conductors in an environment contaminated with Cl_2," IEEE Transactions on Parts, Hybrids, and Packaging, v 12, p. 176.

16. Harsanyi, G., 1995, "Electrochemical processes resulting in migrated short failures in microcircuits," IEEE Transactions on Components, Packaging and Manufacturing Technology –A, v 18, n0. 3, pp. 602-610.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Failure Analysis Flow for Package Failures

Rajen Dias
Intel Corporation, Chandler, Arizona, USA

Abstract

Two generic package failure analysis flows are described, one for open and high resistance failures and one for shorts and leakage failures. The FA flows detail the logical flow of a series of non destructive tests to fault isolate the failure, followed by a series of destructive tests to understand the nature of the fail mechanism and finally, material analysis tests to characterize the chemical nature of the defects. The flows emphasize the importance of documenting details of the unit, verifying the failure and the need for a detailed visual inspection prior to any other analysis.

Introduction

Package failure analysis (FA) is understanding product failures caused by the package. The electrical or physical failure may be detected during manufacturing processes such as package assembly processes, test, printed circuit board mount processes and system test, during reliability testing or during field use. Diagnostic tests are usually done at the system, board or package level to determine likelihood of the failure being in the system, board, package or device. Pinpointing the global region of failure within the system is not trival. Moreover, increase in integration of components in systems result in very complex interactions between Si devices, packages, enabling solutions such as heat sinks and peripheral components. These interactions sometimes make it difficult to attribute a failure to a specifc cause.

In this paper the focus is on failures associated with the package and which includes interconnections between the package and board, between the package and device or within the package itself. The FA flow described is a generic flow for a typical single device packaged in a plastic or ceramic package. The intent of the FA flow is to step through a logical investigative sequence using a variety of tools and techniques to identify the location of the failure and analyze the nature of the defect.

Definition of a package failure

Failure of a product or test chip can be defined as performance not meeting a specific criteria. Electrical performance such as device functionality or ability to run a set of test vectors is often the most important performance criteria, but other criteria such as speed, noise or electrical/ thermal resistance could also be used. It is important that the failure analyst understand the criteria for failure prior to any analysis. It is also important to understand if the failure is a marginal failure or not, as the failure may recover or may not be a valid failure to begin with.

Product performance failures are first diagnosed by the product or test engineers who run a series of parametric test to determine the nature of the failure. Based on these results and knowledge of the product, the engineer may have sufficient evidence to determine that the failure is package related and sends the failure to a package failure analyst for further analysis. Typically gross functional failures, open or high resistance input/output (I/O) pins or shorts/leakage failures between I/O pins or between I/O pins and power/ground pins may be potential package related issues.

Two generic FA flows, one for opens/high resistance failures and one for shorts/leakage fails are presented below. Many of the FA steps are similar but the tools and techniques used to fault isolate the failure and analyze the nature of the defect can be different.

FA flow for open and high resistance failures

The FA flow for opens and high resistance failures shown in Flow Diagram 1 describes a generic flow that can be used for most package types. Details of the elements in the flow are described below.

Prior to any analysis on the failure, details of the unit, stress or field history and test verification data need to be documented. All markings on the unit that will help in tracability of the device, package or materials origin may be critical in identifying the root cause of the failure. It is important to do this first as critical evidence such as fab and assembly lot marks could be erased during subsequent destructive physical analysis steps.

Electrical failure verification is next. It is important that the analyst verify test results to confirm the value of the resistance of the open failure using standard power meters or parametric analyzers. Any major discrepency in failure validation should be investigated as it may indicate an intermittent failure or a test problem. Possibility of intermittent failure should be investigated by doing the test at a temperature higher and lower than the original test temperature. Identification of additional pins that fail at these temperatures may help in understanding the nature of the fail mechanism.

A good visual inspection of the unit using standard optical microscopy techniques should be done next and any abnormality detected may provide clues to the failure.

Common package defects that may be observed include package warpage, cracks and damage, corrosion, oxidation and discoloration on leads, cracks and breaks on pins or solder joints. If defects are seen and could potentially be related to the failure, the analyst should characterize these defects by further optical and Secondary Eelectron Microscopy (SEM)/ Energy Dispersive Spectroscopy (EDS) analysis.

If external examination does not reveal any obvious defects, a series of non destructive analysis should be done prior to any destructive analysis as the later may cause additional open failures. Typical non destructive test done are Time Domain Reflectometry (TDR), X-ray inspection and acoustic imaging.

The needs to do all three analyses or the order in which they are done is at the discretion of the analyst and maybe be influenced by potential hypotheses of the failure based on the electrical test data or previous failure analysis information.

TDR is used to fault isolate the location of a failure (1). It is often used for complex packaging systems where there are a number of interconnection layers such as sockets, BGA joints, flip chip bumps and wirebonded devices. TDR helps in narrowing down where the failure may be so that further analysis can be concentrated in that region.

X-ray radiographic inspection is used to image high density materials within the package for evidence of damage or defects. It is most often used to identify failures or cracks in wire bonds, flip chip bumps, BGA joints and metallic connections (2, 3)

Another commonly used non destructive technique is acoustic inspection, where ultrasonic waves are used to detect and image the location of defects such as delamination of internal interfaces, cracks and voids(4,5)

When a defect or abnormalities is detected with one or more of the non destructive techniques described above, destructive physical analysis (DPA) techniques can then be used to characterize the defect and determine if the defect could possible be related to the failure. For example, if a wire bond connected to an open I/O appears to have a very small ball bond on the die surface, a reasonable hypothesis is that an abnormal wirebond may be the cause of the open I/O.

If no abnormalities are seen during non destructive tests or if the fault cannot be isolated to a region of the package, then DPA can still be used to exposed internal features of the package from where electrical microprobe testing can be used to fault isolate the failure.

DPA techniques commonly used are decapsulation or package deprocessing and x-sectioning. Decapsulation normally refers to removing the lid on ceramic packages or the plastic molding compound material covering the die and wirebonds in organic packages. Wet chemical etching of the molding material using concentrated acids such as sulfuric or nitric is the most common way of decapsulating molded plastic packages to observe abnormalities such as wirebond failures. Other methods of deprocessing the package include mechanical decapsulation and reactive ion etching (RIE). Mechanical decapsulation is splitting the package open to observe internal features. A common application is when a delamination interface needs to be exposed to analyze for evidence of contamination. Thermal lid or heat sink removal, flip chip die removal are other examples of package deprocessing methods. RIE is used to remove organic material such as solder mask material and dielectrics in organic package substrates to expose the metallic traces and via connections as shown in figure 1.

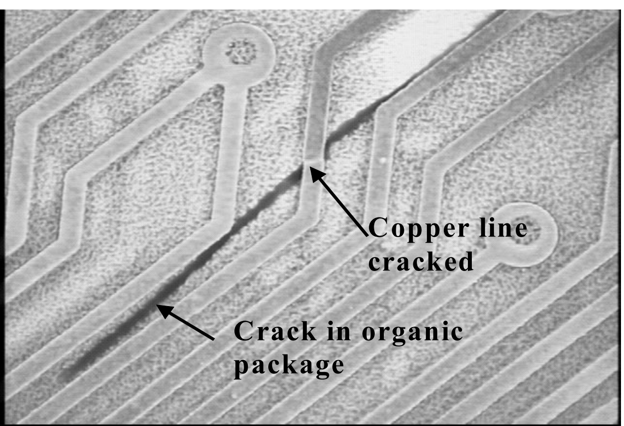


Figure 1. Optical photo of organic package after RIE etch of the top layers shows a package crack severing a copper line.

X-sectioning of the package uses standard metallographic techniques such as diamond saw to cut the package and grinding and polishing techniques to reveal details of the packaging layers. The technique is used often to understand the initiation and propagation of delamination and cracks within the package and to understand geometry and metallurgical reactions of the interconnections. A variation of x-sectioning is grindback or planar grinding where package material parallel to the die surface is sequentially removed.

FA flow for shorts and leakage failures

Many of the steps described in the FA flow for opens and high resistance failures are also applicable for shorts and leakage failures as shown in Flow Diagram 2. Important attributes of each step as they are pertinent to short/leakage fails are described in this section as well as new steps such as thermography and SQUID microscopy (6).

Unit documentation is same as for open failures. Electrical failure verification is next. The analyst should confirm the value of the short or leakage current using standard power meters or parametric analyzers. Any major discrepency in failure validation should be investigated. It is possible for short failures to pass if the unit has been been exposed to mechanical stress such as unit removal from a mother board. Leakage failure could possibly show lower leakage values if moisture level in the unit decreases and this could indicate that the failure is moisture sensitive.

A good visual inspection of the unit using standard optical microscopy techniques should be done next to look for any evidence of foreign material that could cause a short/leakage. Common package defects include contamination, corrosion or discoloration between lands/leads, metal migration or exposed traces and package cracks. If defects are seen and could potentially be related to the failure, the analyst should

characterize these defects by further optical and Secondary Eelectron Microscopy (SEM)/ Energy Dispersive Spectroscopy (EDS) analysis.

If external examination does not reveal any obvious defects, a series of non destructive analysis should be done prior to any destructive analysis as the later may cause the short or leakage to recover. The typical non destructive test done are X-ray inspection, acoustic imaging, infrared thermography and SQUID microscopy.

X-ray radiographic inspection is used to check for potential wire shorts such as wire sweep, metal migration between leads, solder bump bridging etc. Figure 2 shows an x-ray image of a flip chip package that revealed bridging between bumps. Often the package will need to be tilted during inspection to validate if there is a metallic material between the connection of the shorted pins. If defects are found, standard DPA techniques such as decapsulation, RIE and x-sectioning are used to characterize the defect.

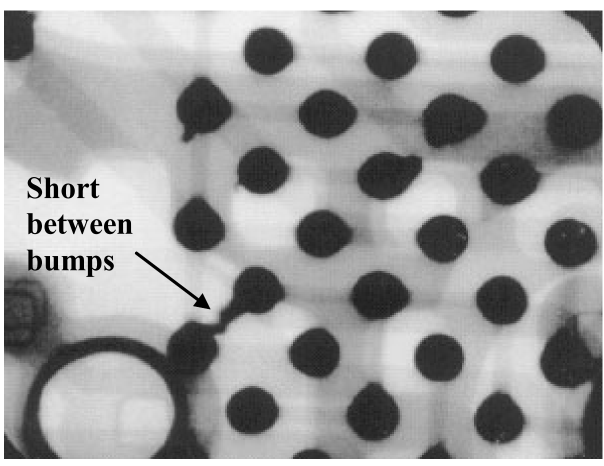


Figure 2. X-ray image of flip chip package showing bridging between bumps that were electrically shorted.

Acoustic inspection can also be used to investigate for delamination, cracks and voids. Detection of these defects could indicate a potential path for contamination or moisture to cause leakage or shorts by metal migration growth. Figure 3 shows an acoustic image of underfill voids that were detected at locations where the bump bridging was detected by x-ray.

If a defect can be located, then DPA is used to characterize the defect. In the above example of a short between bumps, x-sectioning between the bumps to reveal the bridging material and to understand the influence, if any, of the underfill void was done and results shown in figure 4 identified the mechanism as bump extrusion into an underfill void.

If x-ray and acoustic inspections fail to detect any abnormalities, then infrared (IR) thermography inspection should be done to determine the location of the short. A current is forced between the shorted pins and any resistive current is forced between the shorted pins and any resistive

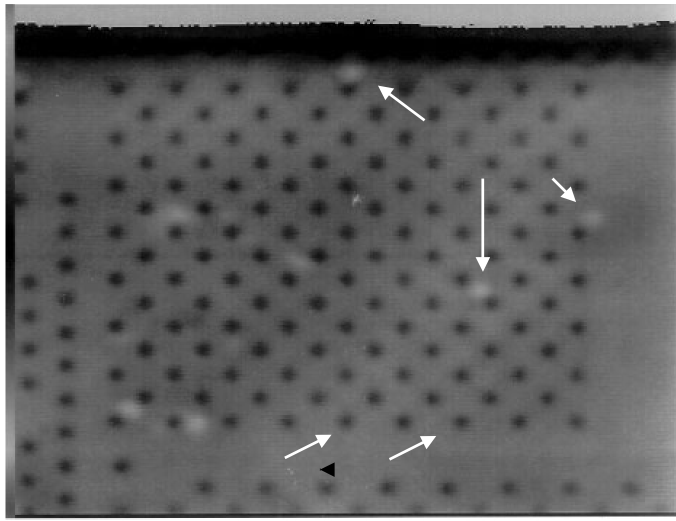


Figure 3. Acoustic image of flip chip package showing underfill voids. Shorted bumps that showed bridging in X-ray (figure 2) also show an underfill void between the bumps

.

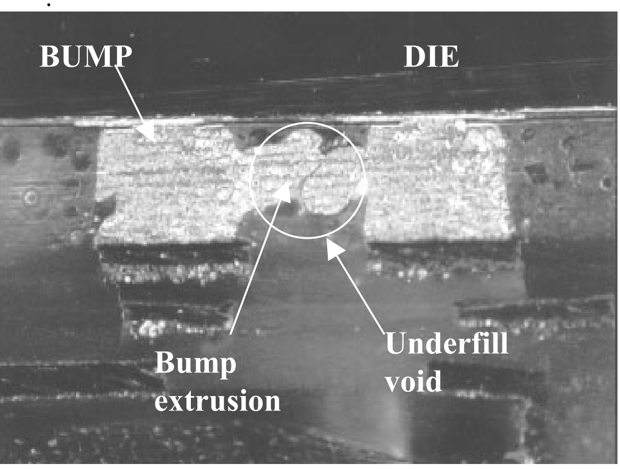


Figure 4. Dark field optical image of x-section of shorted bumps showing that solder extrusion into the underfill void was the mechanism of failure.

heating effect at the short location can be imaged by an IR thermal camera. It is important that the current be increased gradually to prevent the short from being blown open and to obtain a small hot spot so that the centroid can be easily located as the short location. IR thermography is not very effective when the short is low resistance or dead short, because resistive heating cannot occur unless the current used is very high. In these situations, a relative new technique called Scanning SQUID Microscopy (SSM) can be used (6). SSM images the flow of current between the shorted pins by detecting the magnetic field associated with current flow. The the short location is identified by comparing the current image with schematics of the layout of the shorted pins. Very low current in the order of 100 microamps can be used and this ensures that leakage caused by fragile metallic filaments are

not blown open. The technique is useful for determining if the short or leakage is in the die or package.

If the above mentioned techniques do not reveal evidence of the location of the short/leakage failure, then the package may need to be decapsulated to expose the metal trace connections of the shorted pins. Visual inspection after trace exposure may reveal the cause of the short. If no abnormalities are seen, then one of the traces is cut in half and electrically tested to determine which half is shorted. This procedure can be repeated till the region where the short is location is small enough to do a physical x-section or mechanical peeling between the shorted leads. It is important that before and during multiple DPA steps, the short/leakage failure is continually checked to make sure that the failure has not recovered.

Discussion

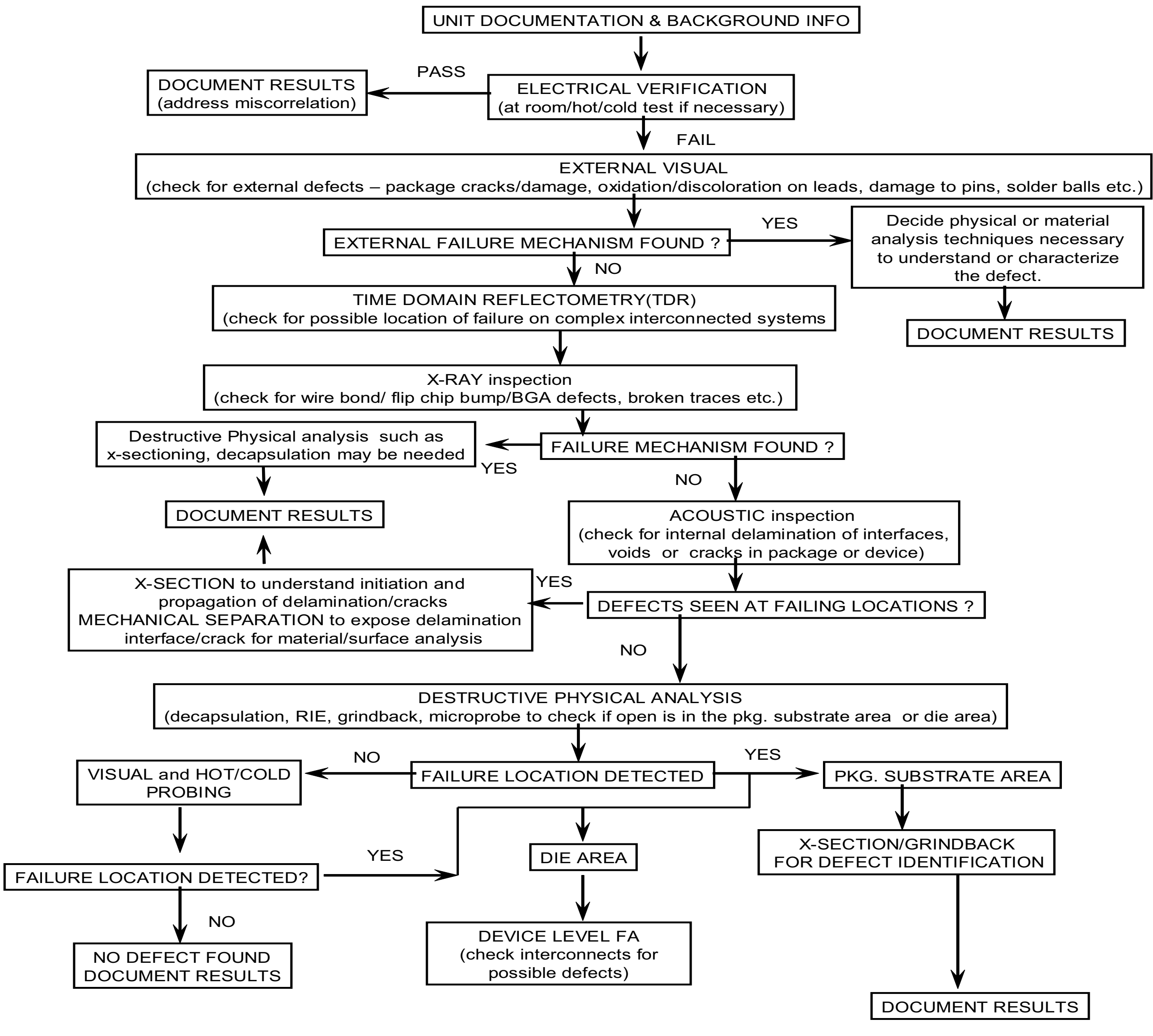
The FA flows described above are generic flows for package fault isolation and failure analysis. The failure analyst should use them as a guide for package analysis. Often, when the failure identification is difficult, reviewing the FA flow may indicate steps that have been skipped that may be worth doing. On the other hand, if the product or package technology is very well understood and its susceptibility to certain failure modes are understood, then the analyst may choose to bypass certain parts of the FA flow to expedite the analysis.

Conclusion

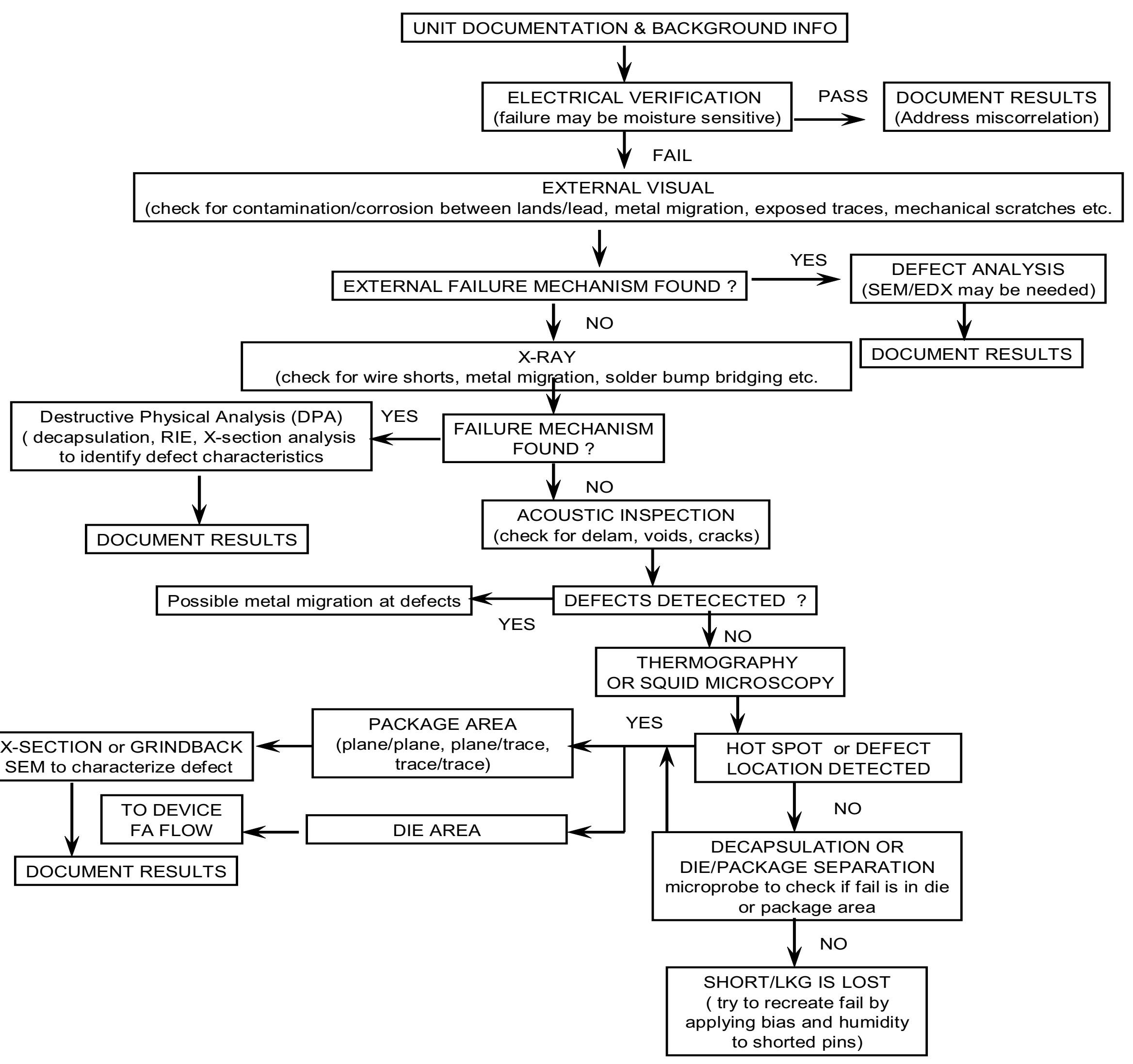
The paper describes generic package FA flows for opens/high resistance and shorts/leakage failures. Documentation of package details and failure verification is followed by a series of non-destructive tests to identify the fail location and then destructive techniques are used to determine the nature of the defect and understand the fail mechanism. Finally, material analysis techniques are used to chemically characterize the defects.

References

1. D. T. Searls, D. Anura, E. Dy, and D. Goyal, "Time Domain Reflectometry as a Device Packaging Level Failure Analysis and Failure Localization Tool" 2000 International Symposium for Testing and Failure Analysis
2. D. Goyal, 2000, X-Ray Tomography for Electronic Packages, Proceedings of the 26th International Symposium for testing and failure analysis, p49.
3. D. Goyal, Zezhong Fu, Jeffrey Thomas, Alan Crowley and Andrew Ramsey, 2003, 3D X-Ray Computed Tomography (CT) for Electronic Packages, Proceedings of the 29th International Symposium for testing and failure analysis, p56,
4. T. Moore, Identification of Package Defects in Plastic Packaged Surface Mount ICs by Scanning Acoustic Microscopy" 1989 International Symposium for Testing and Failure Analysis
5. J. E. Semmens, S.R. Martell, L.W. Kessler, "Evaluation of Interconnects Using Acoustic Microscopy for Failure Analysis and Process Control Applications", 4th International Conference & exhibition on Multichip modules, p 279-285 (1995)
6. R. Dias, L. Skoglund, Z. Wang, and D. Smith "Integration of SQUID microscopy into FA Flow," 2002 International Symposium for Testing and Failure Analysis.



Flow Diagram 1



Flow Diagram 2

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Chip-Scale Packages and Their Failure Analysis Challenges

Susan Xia Li
Spansion, Inc, Sunnyvale, California, USA

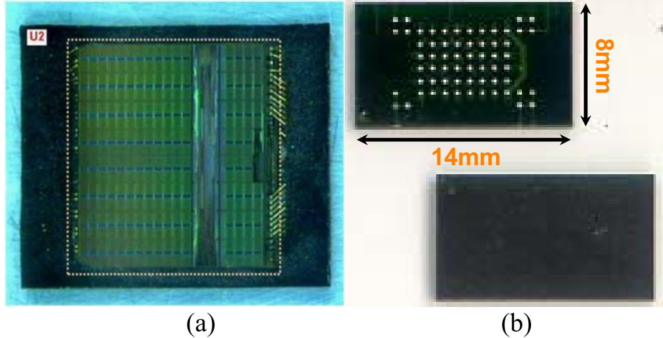
Abstract

Chip Scale Package (CSP) is ideal for the applications of Cellular and Portable devices that require better use of real estate on the PC boards. It has advantages of low package profile, easy routing and superior reliability. However, due to their small form factor, it is difficult to handle this type of package for both package level and die level failure analysis. In this paper, a brief overview of definitions for CSPs and their applications are included. The challenges for performing failure analysis on CSPs, particularly for Multi-Chip Packages (MCP), at package level and die level are discussed. In order to successfully perform device electrical testing and failure diagnostic on CSPs, special requirements have to be addressed on precision decapsulation for FBGA packages, and additional attention has to be paid to top die removal for MCPs. Two case studies are presented at the end of this article to demonstrate the procedures for performing failure analysis on this type of device.

Introduction

There is a trend in the electronic industry to miniaturize. From tower PCs to laptops to Pocket PCs, from giant cell phones to pager size handsets, the demand for smaller feature-rich electronic devices will continue for many years. In response, the demand for Chip-Scale Packages (CSPs) has grown tremendously. Their main advantage is the small form factor that provides a better use of real estate on the PC board in many applications such as cell phones, home entertainment equipment, automotive engine controllers and networking equipment. All have adopted CSPs into their systems.

A Chip-Scale Package is, by definition, a package about 1 to 1.2 x the perimeter (1.5 x the area) of the die (Figure 1). Within this definition, CSPs have many variations. There are more than 20 different types of CSPs in the market today, but all of them can be grouped into 4 main categories based on their technologies and features (Table 1).



(a) (b)

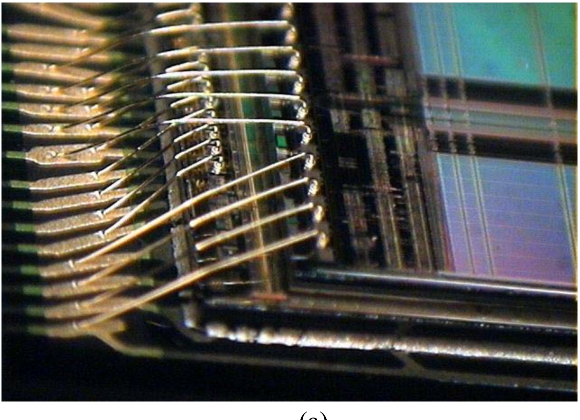
Figure 1: A typical FBGA package shows the die size similar to the package size (a) and the dimension of the package is small compared to conventional BGA packages (b)

To further increase Silicon density within small CSPs, the so-called Stacked Chip-Scale Packages (SCSPs) or Multi-Chip Packages (MCPs) have emerged into the market and are becoming one of the most rapidly growing sectors for CSPs. Packaging for MCPs begins by stacking two or three dice on top of a BGA substrate with an insulating strip between them. Leads are bonded to the substrate, and molding around the stack completes the device. Since a MCP is a package that may even be smaller than the enclosed die area (counting area from the stacked dice), our definition of a CSP no longer holds. The MCP example in Figure 2 shows paired Flash memory and SRAM.

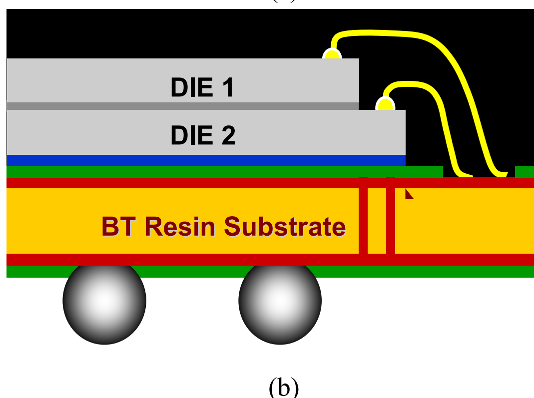
Another MCP example is memory to support logic. Package-on-Package (PoP) stacking is often used to make the assembly process high yield, low cost and more flexible for the integration of memory to logic devices. Figure 3 shows an example of PoP stacking. These stacked packages are so thin (1.4mm) that wafers must be thinned to 150-200um prior to wafer saw for die separation and placement in order to fit into the MCPs.

Table 1: Four Main Categories for Chip Scale Packages

TECHNOLOGY	REMARKS
Flexible Substrate Micro BGA	It uses a flex circuit interconnect as a substrate by which the chip is connect to the circuit board. There are two ways of interconnecting the chip to the substrate: flip chip bonding and TAB style lead bonding.
FBGA-PI	As for wire bond type, the chip is protected using conventional molding technology
Rigid Substrate Organic Or Ceramic FBGA-BT	This type of CSPs has a ceramic or an organic laminated rigid substrate. Either conventional wire bonding or flip chip can be used to connect the die to the substrate.
JACS-Pak	The wire bond version can be considered as a downsized version of the standard PBGA even though assembly techniques could be different.
Lead Frame Type LOC/SON	The most common CSPs of this type are the LOCs (Lead on Chip) where the lead frame extends over the top of the chip.
Wafer Scale Ultra CSP	This type of CSPs has the chips that are processed in wafer form before singulation. A re-distribution layer is created in the wafer scribe to pin out the bond pads to a standard ball grid array footprint.



(a)



(b)

Figure 2: A two die stacked MCP package with SRAM on top of Flash configuration: (a) top view of the MCP package (b) cross-sectional view of the MCP package

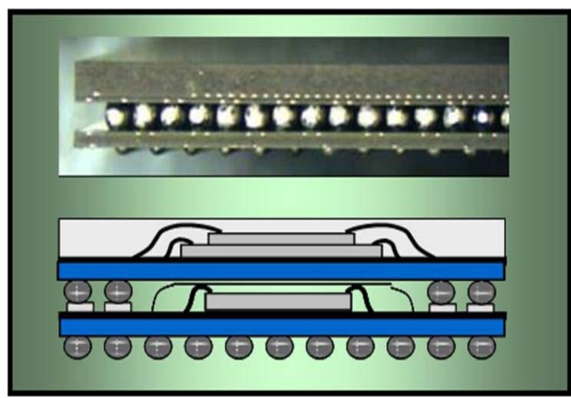


Figure3: PoP stacking provides high yield, low cost and more flexible assembly process for the integration of memory to logic devices

Failure Analysis Challenges

The many advantages that CSPs bring to the IC applications also bring new challenges for device testing and failure analysis at both the package and die levels. Since they are small, device handling for CSP operations is difficult. The following discussion mainly focuses on one type of CSPs, the

FBGA (Fine-Pitch Ball Grid Array) package, which illustrates the issues that we have encountered during failure analysis on CSPs.

A typical wire-bonded CSP, such as an FBGA package with a BT (Bismaleimide Triazine) resin substrate, uses a standard die attach material and is gold wire bonded to external I/O lead frames. The entire package has a straight edge, since the dicing process makes it so. Its package height is normally less than 1.2mm, with 0.8mm ball pitch, and 0.3mm solder ball size, which maintains a nominal stand-off of 0.25mm (Figure 4).

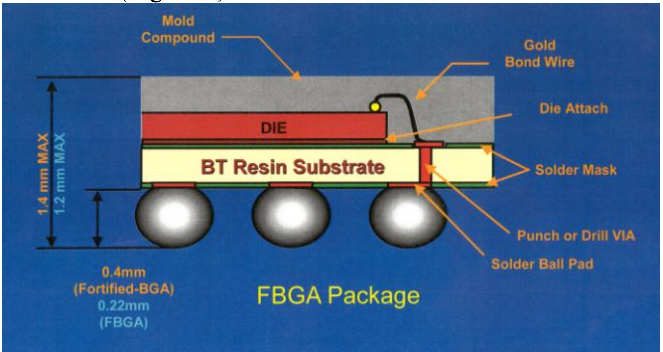
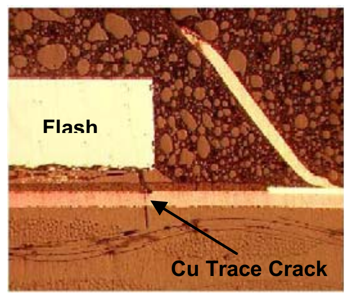
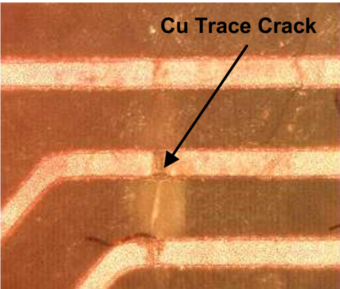


Figure 4: Single Die FBGA Package Build-up Structures

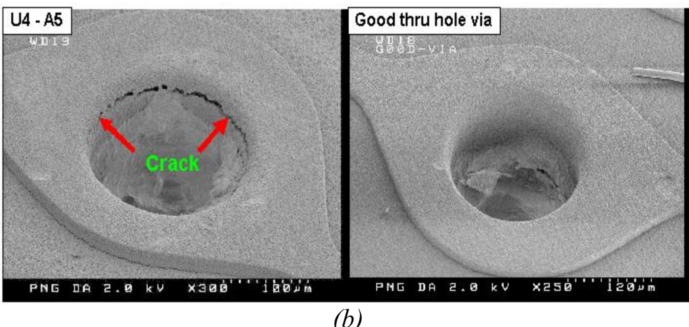
Package Related Failures

Like any other BGA packages, the typical failures at the package level for FBGA include open and short circuits after reliability stress testing. Most of common open circuit failures are related to package and/or die cracks, Cu trace cracks or through-hole via cracks (Figures 5a and 5b). The failures may also involve wire bond cracks and solder ball cracks (Figure 5c). Most common short circuit failures involve Cu trace shorts due to contamination, wire bond touching silicon die edge, or adjacent wire bonds touching each other. Traditional failure analysis techniques still apply to FBGA packages except more attention is required due to their smaller feature sizes. Non-destructive analysis, such as X-ray micrographs, Scanning Acoustic Microscopy (SAM), or even mechanical probing should always be done first to isolate possible failure site(s). Destructive analysis, such as, mechanical cross sectioning or dry and wet chemical etching for exposing the failure site(s) at the package level, must then be performed to collect physical evidence.

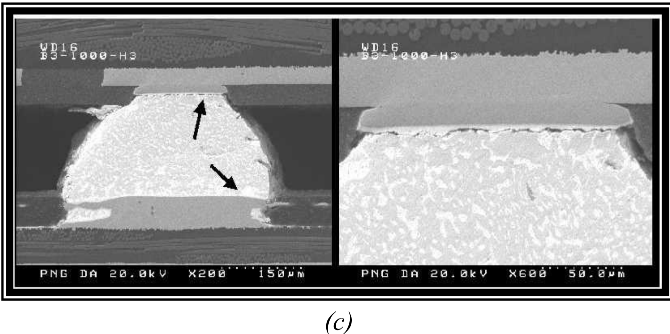




(a)



(b)

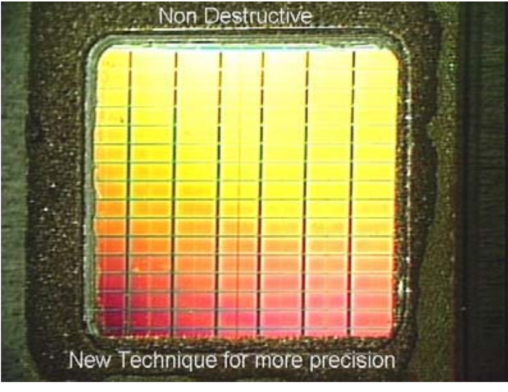


(c)

Figure 5: Typical package level failures on FBGA packages:
- *(a) Cu trace cracks in the package*
- *(b) Through-hold via cracks after temp cycling*
- *(c) Solder ball cracks at the interface of the FBGA package to the system PCB board.*

Die Related Failures

The real challenges for failure analysis of CSPs come from analysis of die related failures. Since a CSP device has a package size about the same as the die size, it is difficult to decapsulate the package and to expose the die for circuit diagnosis and analysis without destroying device connectivity to the package. A traditional plastic package usually has a package much bigger than the die, so decapsulation can be done easily using an automatic package-etching tool with a proper rubber gasket since there is a lot of room for error tolerance. For a CSP, precision decapsulation is required since there is no room for error. The etching window on the gasket needs to precisely correspond to the die size, requiring that the gasket be more rigid and have no deterioration, even under long exposures to H_2SO_4 acid (>60 seconds) at high temperature (>220C). Any over-etching could damage the package PCB substrate, break Cu traces, and even lift wire bonds (Figures 6a and 6b). Without package integrity after decapsulation, no electrical testing and subsequent diagnostic work can be done on a die related functional failure. Furthermore, the CSP devices have very thin dice that can be easily cracked by applying extra force on them during the decapsulation process.



(a)

(b)

Figure 6: Comparison of good (a) and bad (b) decapsulation of FBGA package

Compared to a CSP that has a single die within the package, a MCP has additional challenges due to its stacked die configuration, particularly for failure analysis of the bottom die or dice. Since the stacked dice share power supply pins and most address pins, there are fewer connections to the package with a MCP than if the enclosed dice were in separate packages. Normally, a Chip Enable (CE) function activates each die to run the diagnostic testing, so that the functional failures can be isolated to certain suspect die. However, for a high IccSSB (Icc Super Standby) current failure, multiple dice often share the same power supply pins (Figure 7), so it may not be easy to determine the source of the IccSSB leakage. To find out which die is responsible for the failure, the bond wires for power supply pins must be removed from one die at a time, then the leakage re-checked until the source is found.

The analysis can proceed once the failing die is identified. If the failing die is on top, further analysis on the device is similar to a single die CSP. However, if the failing die is on the bottom, it is most likely partially or entirely covered by the top die. Circuit analysis and fault isolation cannot proceed until the top die is removed and the failing die is exposed. The most challenging part for MCP failure analysis is removing the top die or dice while preserving electrical connectivity of the bottom die to the package. Some of the solutions for this challenge will be discussed in the following sections.

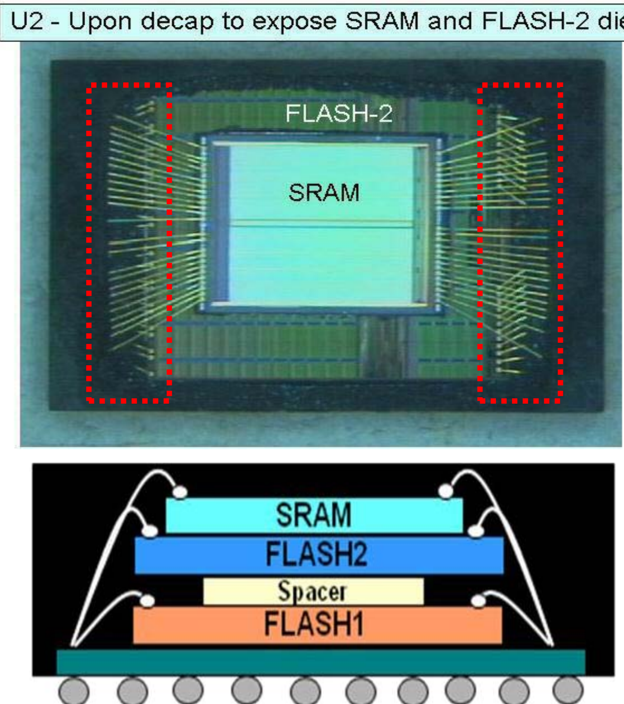


Figure 7: A MCP package has stacked dice sharing most power supply and data pins

Corresponding Solutions

There are no easy solutions to address all issues encountered for failure analysis of CSPs. However, many new techniques and special tools have been developed around CSPs to overcome the difficulties.

Precision Decapsulation

Precision decapsulation is required for CSPs, particularly for most FBGA packages. The key for successful decapsulation is to use well-defined gaskets. Rubber gaskets are currently used for most automatic package etching systems. These gaskets are made of acid-resistant viton rubbers, but they easily deteriorate with long exposure to H_2SO_4 at high temperatures. As a result, a rubber gasket can only be used once or twice before high precision decapsulation cannot be achieved. Furthermore, the debris from a deteriorated rubber gasket often get into the etch head of a package etching tool, preventing proper acid flow and causing severe system problems.

One solution to this problem is the use of a different type of gasket, such as those made from Teflon materials [1]. As we know, Teflon materials have acid resistance and heat endurance. They also can have well-defined cuts that can form etching windows for precision gasket. However, they

may not have as much flexibility as rubber materials to provide an airtight environment for the automatic package etching tools. The combination of a Teflon gasket and a rubber gasket has proven to be the ideal solution for precision decapsulation (Figure 8). In addition, using a less aggressive acid, such as, a mixture of HNO_3 and H_2SO_4 acid also reduces the damage on the FBGA packages but requires a longer etch time.

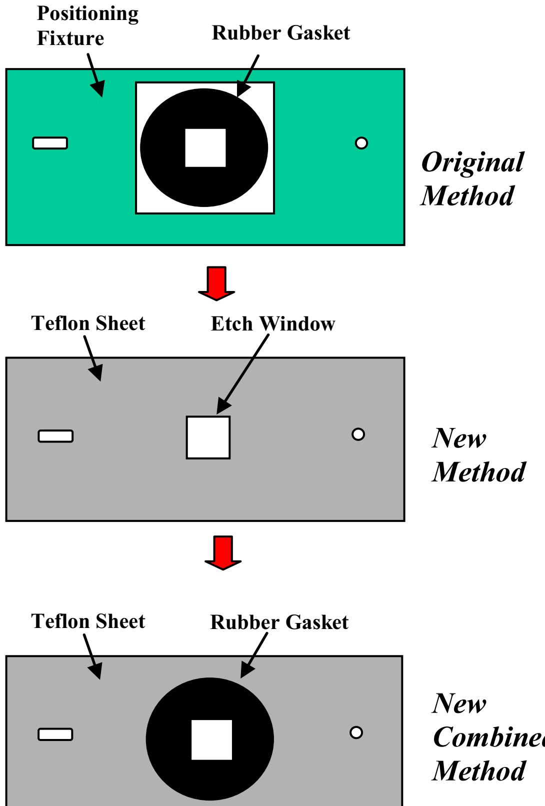
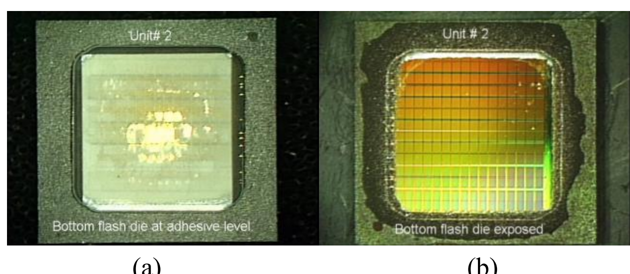


Figure 8: A combination of Teflon gasket and Rubber gasket provides a better etching result for precision decapsulation.

Top Die Removal

If the package in a MCP has two stacked dice and the failing die is identified to be the bottom one, then top die removal becomes necessary for further analysis. There are several approaches for top die removal. One uses a chemical etch after polishing off the active region of the top die. KOH or TMAH are good selective etchants that etch bulk silicon with a high etch rate. The etch stops at the adhesive layer before reaching to the bottom die. One concern for this approach is that the adhesive layer may not completely stop the chemical from etching the bottom die if the adhesive layer does not cover the bottom die entirely. Another concern is that the entire package is subjected to a chemical environment at an elevated temperature (~60-80C), and the solder mask from the FBGA package substrate is often etched off. Therefore, the Cu traces in the package could be easily damaged during subsequent handling. To avoid a chemical etch for top die removal, another approach uses mechanical milling [2]. A micro drill with computerized control can mill off the plastic encapsulation material and the top die or dice from a pre-defined area (Figure 9). However, the mechanical force applied to the die is a concern during the milling. The MCP dice are very thin, so the milling process has to be well controlled to ensure that the force applied to the device will not crack the die or the package.



(a) (b)

Figure 9: The top die was removed using a micro drill from the MCP package with two same die stacking

(a) After molding compound and top die removal (stopped at the adhesive layer)

(b) After bottom die exposure (etched off the adhesive layer)

The chemical and mechanical approaches for MCPs with more than three stacked dice may reach their limitation. An alternative approach should be considered. One approach loosens the package with a wet chemical etch to get the dice out for repackaging. A mixture of nitric acid and water can free the plastic FBGA packages and leave the silicon dice in good condition after the etching. However, prolonged etch with this mixture will also remove the Al metal within the bond pad areas, resulting in wire-bonding problem afterwards. Another approach mechanically polishes off undesirable top die/dice as well as the solder balls on the backside of the package, and then repackages the remaining piece with exposed failing die into another open top package for analysis. However, repackaged device with longer wire loop could have noise issues during electrical testing so that this approach may not apply to the devices sensitive to noise. Experimenting and practicing on the suggested techniques are the keys to success on MCP analysis.

For more advanced process technology, copper material is used for circuit interconnection. A traditional chemical etch process becomes incompatible with this new metallization. An increased number of stacked dice in MCP packages also requires each die to be thinner (individual die thickness to be 50um for a 6 stacked-die MCP). Mechanical milling for top die/dice removal for MCP analysis becomes impossible. Laser ablation using next generation multi-wavelength ND:YAG laser technology is being developed to handle decapsulation of advanced MCP and copper metallization devices [3]. This new technique avoids chemical reaction to the copper materials and eliminates mechanical force applied to the packages during decapsulation. However, controlling a laser to mill off materials with very different properties is

straightforward in principle, but more difficult in practice. The optimal process settings for bulk decapsulation of the package may be significantly different from those required at the device surface. Currently users are still experiencing laser damage on the die surface if the process parameters are not properly set.

Backside Accessibility

For some CSPs, such as FBGA package, it becomes difficult to access the device from the backside due to the solder ball attachment from the backside. With multi-level metallization on the IC devices, emission analysis is not always successful from the front side. The upper level power buses often times cover the emission site at the lower level to make emission detection impossible. In addition, backside emission analysis is more capable to detect the emission from low level leakages. In order to perform backside analysis on FBGA packages, a special sample preparation is required. One example of sample preparation for backside emission analysis is illustrated here:

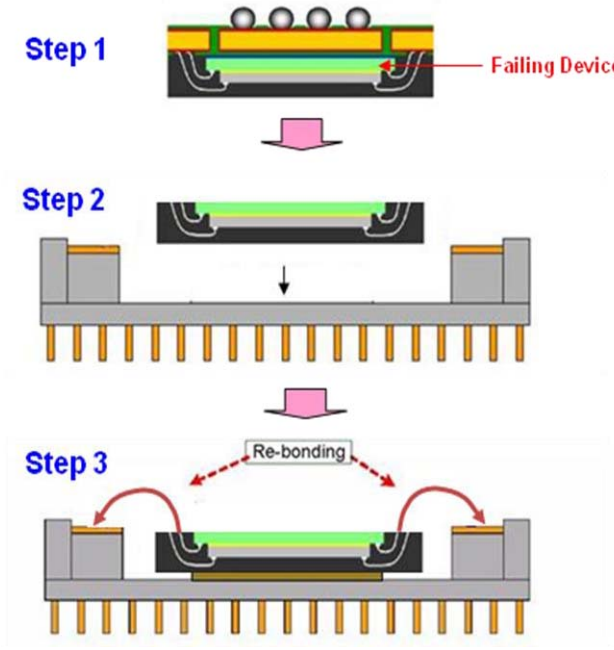


Figure 10: the basic steps of sample preparation for backside emission analysis on FBGA packages

As illustrated in Figure 10, the steps for sample preparation are: 1) polish the package to expose the backside of the failing device for analysis; 2) attach the polished the device onto a ceramic PGA package; 3) use a manual wire bonder to connect the polished device to the PGA package. After this sample preparation, it is ready to perform emission analysis on the repackaged device.

Case Study I

One Flash device failed after endurance cycling. Electrical testing results showed a cluster of sectors failed for embedded programming and erase operations. Bitmapping of the failing bits showed a hairline crack-like pattern across some of the failing sectors (Figure 11). The cell Vt tests also showed higher leakage values at the failing bits.

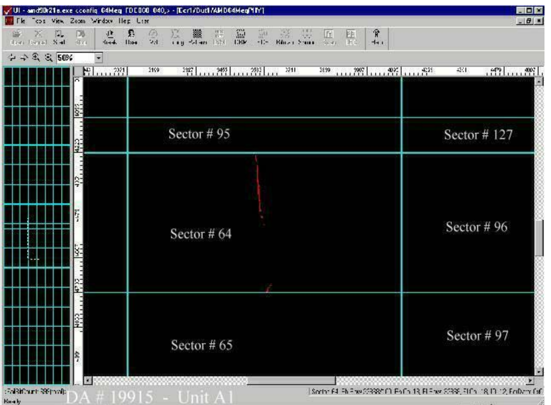


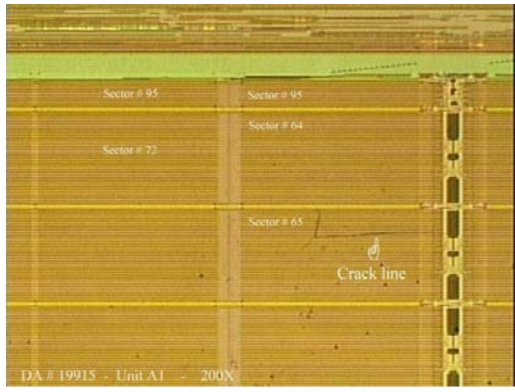
Figure 11: Bitmapping results using a memory tester showed Hairline crack-like pattern of the bits (red colored) that could not be programmed and erased at the failing sector(s)

With the speculation of a hairline crack on the failing device being present, careful SAM scan was performed on the unit, and C-scan (x/y plane) and B-scan (x/z cross-sectional) results indeed showed the possible die crack at the failing sectors (Figure 12). With this information, the FBGA package was re-examined again, and a faint crack line could be seen on the package surface (Figure 13).

Figure 12: SAM C-scan (left) and B-scan (right) results showed a possible die crack at the failing sectors location (circled)

Figure 13: Re-examination on the FBGA package surface showed a faint crack line at the suspected area (circled)

The unit was then carefully decapsulated to expose the die surface for inspection. Optical inspection at the failing sector locations showed hairline cracks that match the patterns shown on the bitmap and SAM scan (Figure 14). Further investigation on the root cause of the failure concluded that the device had been mechanically damaged during the read point testing.



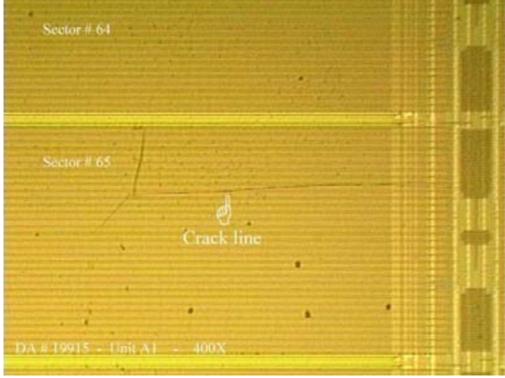


Figure 14: Optical inspection at the failing sectors showed hairline cracks that match the pattern shown on the bitmap and SAM scan.

Case Study II

One MCP device (one SRAM + two FLASH) failed for the top flash erase (FLASH1). Further electrical testing showed that one of the sectors on the failing device had pre-programming problem at a single column (Figure 15). The column leakage test showed high leakage at the affected column. The row leakage test also showed a single row with high row leakage. The intersect cell of the leaky row and column had an abnormal Ids curve (Figure 16). Based on this electrical results, the focus was at the intersect cell of the affected row and column for physical defect location.

Since this is a MCP device with one SRAM on top of two Flash devices, and the top Flash (FLASH1) is the one identified with the failure, mechanical polishing using the milling tool was done to remove the top SRAM and the Si spacer to expose the top Flash die (Figure 17).

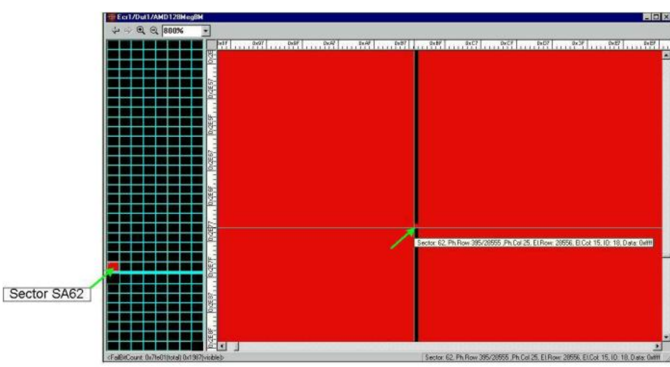
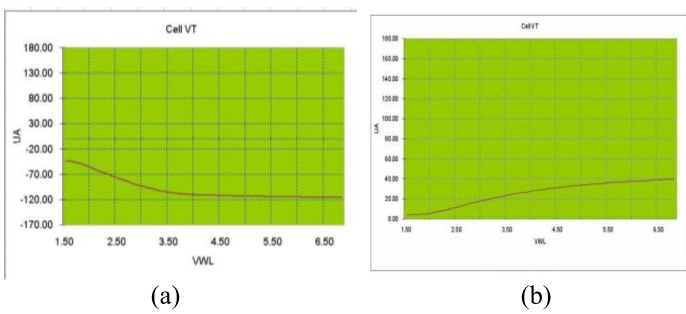


Figure 15: Bitmapping results using a memory tester showed a leaky column (unable to be programmed) and a leaky row at the failing sector. The arrow pointed to the intersect cell which was suspected to be the cause for the failing row and column.

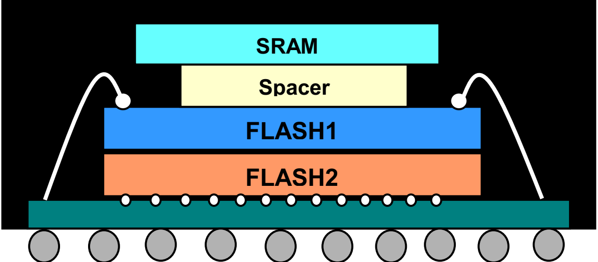


(a) (b)

Figure 16: Cell Vt analysis on the intersect cell showed abnormal Ids curve (a) compared to a good cell (b)

After exposing the top Flash die (FLASH1), emission microscope analysis was performed on the device; however, no emission site was detected. Since the intersect cell of the failing row and column was the suspected failure site,

deprocessing was done with dry etching and mechanical parallel polishing down to the ILD0 level (after metal 1 removal). Passive voltage contrast analysis showed the illuminated drain contact for the failing column and poly2 wordline contact for the failing row, indicating that the affected row and column had a leaky path to the substrate through the intersect cell (Figure 17). FIB cross-section through the intersect cell showed a CoSi residue bridging the poly2 wordline (the failing row) to the drain contact connected to the failing column (Figure 18). The information was then fed-back to the fab for process improvement.



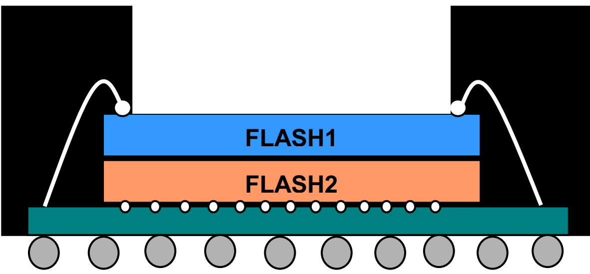
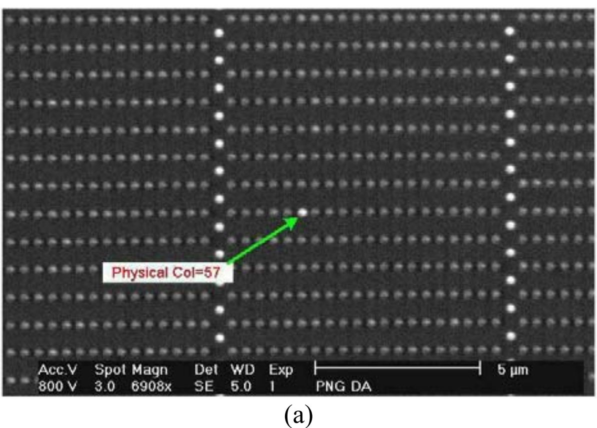
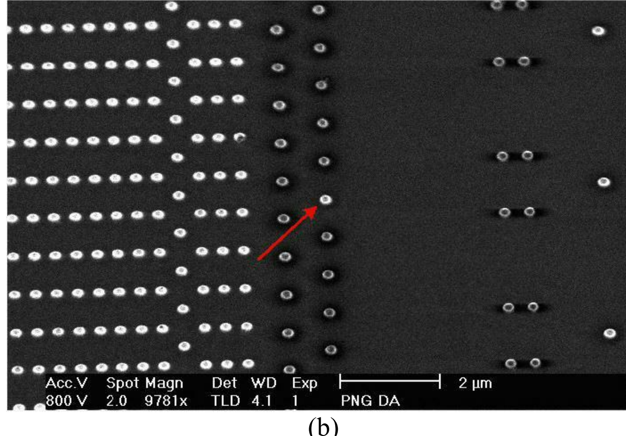


Figure 17: The top SRAM die and the Si spacer was removed using a mechanical milling tool to expose the top Flash die for further analysis (the milling window was setup to such that the wire bonds for FLASH1 were intact). Note that SRAM and FLASH 2 were bonded 90-degree rotation from the FLASH 1 device.

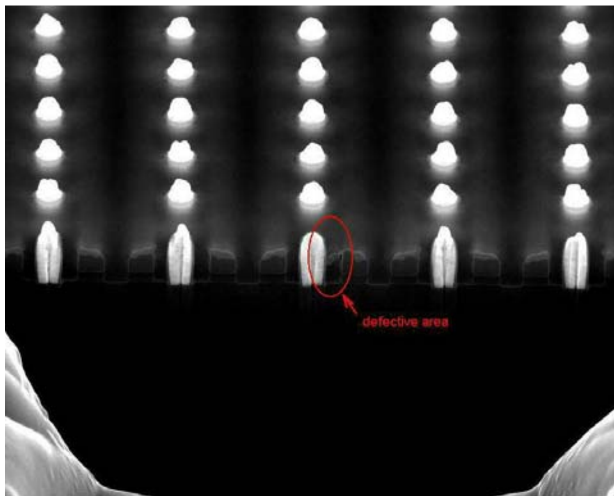


(a)



(b)

Figure 18: Passive Voltage Contrast Analysis showed illuminated drain contact of the failing column (a) and the poly2 wordline contact of the failing row (b)



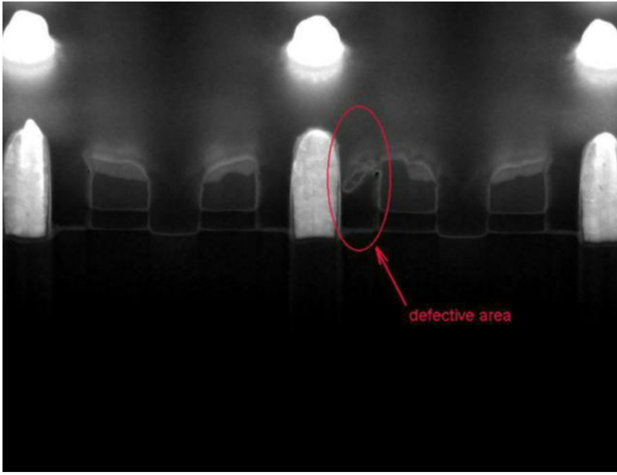


Figure 19: FIB cross-section through the intersect cell showed a CoSi residual bridging the failing wordline and the drain contact for the failing column

Summary

CSPs are getting increasingly more attention due to their advantages of small form factor, cost effectiveness and PCB optimization. They are widely utilized for hand-held devices, automotive engine controller and networking equipment. Their small size creates new challenges for testing and failure analysis. Special requirements, such as precision decapsulation for FBGA packages, top die removal for MCP packages, must be addressed in order to successfully perform device testing and diagnostic analysis. In this article, some new challenges and corresponding solutions were discussed to increase the awareness of particular issues associated with CSPs. Two case studies are used to demonstrate how a typical CSP failure analysis is done.

References

[1] Xia Li, Joseph Vu, Mohammad Massoodi and Jose Hulog, "Automatic decapsulation system utilizing an acid resistant, high heat endurance and flexible sheet coupled to a rubber gasket and a method of us", US Patent Number 6409878 (2002)

[2] Ultratec Application Note, "Enabling Backside Analysis, an overview of Selected Area Preparation", 2002

[3] Control Systemation, Inc, Application Note by Jon W. Heyl, "Laser Based Failure Analysis, including Decapsulation, Cross-sectioning and Materials Characterization", 2004

Acknowledgements

The author would like to acknowledge following people for their contributions to this publication:

- Joseph Vu and Shamsual Mohamed for their device anlaysis data
- Jose Hulog, Mohammad Massodi and Joseph Vu for co-developping the precision decapsulation technique
- Mario Vargas for providing precision decapsulation images
- Gene Daszko and Andy Gray for reviewing this paper and providing valuable input

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Wafer Level Failure Analysis Process Flow

J. H. Lee, Y.S. Huang, D.H. Su
Taiwan Semiconductor Manufacturing Company, Hsin-Chu, Taiwan

Abstract

Failure analysis (FA) plays a very important role in the integrated circuit industry, in particular, wafer level FA constitutes a large portion of the daily work of FA engineers in a fab. In this section, we present a commonly used process flow for performing good, thorough wafer level FA.

Introduction

In a wafer fab, wafer level failure analysis is performed in various situations: on test structures during development and process qualification, after different stages of visual inspection, and after wafer level electrical testing. The main steps for all these FA processes are very similar. The basic concept is to identify possible failure mechanisms based on failure mode and testing results, verify the failure mechanism with manual or engineering mode wafer level testing, perform fault isolation to identify the physical location of the failure, and finally perform physical analysis to find the physical cause of the failure. In a wafer fab, the subsequent actions are also very important. The failure needs to be linked to process steps so that changes can be made to fix the problem. Finally, it is important to track the implementation of the solution to ensure that no side effects result from these actions and that the problem is really resolved.

Overview of the Yield Enhancement Loop

For the purpose of this general introduction, we will assume that the failure analyst is engaged in a yield enhancement activity. Thus, a generalized yield enhancement flow is shown in Fig. 1.

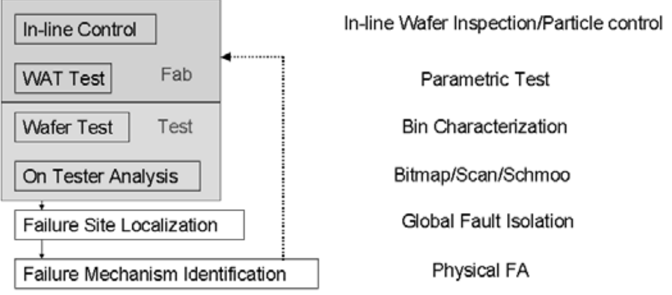


Figure 1: The yield enhancement loop. The column on the right contains brief descriptions of the activities or data acquired in each step.

In Fig. 1 the first three blocks on the left are normal steps encountered in a wafer processing fab. In-line control refers to inspections performed in differents steps of the process using non-invasive means, e.g. after etch inspection (AEI). Wafer acceptance test (WAT) is a series of parametric tests performed at the end of wafer processing to measure individual electrical parameters such as contact or via resistance, threshold voltage etc. Wafer test refers to what is commonly known as wafer sort also known as chip probing (CP). The main output of wafer sort are: yield, wafer maps and bin summaries. Samples for failure analysis can be generated from any one of these steps. We will discuss failure analysis of problems identified at wafer sort since these involve more complex procedures, analysis of failures detected at the earlier stages follows in a pretty straight forward manner.

Continuing with the flow shown in Fig. 1, the first step after sort is on-tester analysis. In this stage, the FA engineer works with a tester in manual or engineering mode to verify the failure mode detected during wafer sort and tries to narrow down possible failure mechanisms and isolate the physical location electrically. Once on-tester analysis is completed, the samples are sent for more detailed fault localization. The latter is the stage in which techniques such as emission microscopy (EMMI), liquid crystal, optical beam induced resistance change (OBIRCH) are employed. After the location of the fault is identified, physical FA takes place. Here is where delayering, cross-sectioning, scanning electron microscopy (SEM) examination, transmission electron microscopy (TEM) analysis and other characterization techniques are used. The aim of this last stage is to identify a physical cause for the failure and feed that information back to the fab so that the problem can be corrected. We will examine each of these steps in more detail later in this article.

Wafer Level FA Process Flow

A general wafer level FA process flow is shown in Fig. 2. In fabs such as the ones at TSMC, several different products are manufactured in the same fab, as a first approximation we divide the products into memory-based and logic-based products. Of course, in many of the cases memory and logic are combined into the same product, but similarly the overall approach for FA can also be combined. The upper half of the process flow was discussed in the previous section, what was described as physical FA earlier is now expanded to show the individual steps. What follows is a more detailed discussion

of the phases following failure part identifications, starting with failure verification.

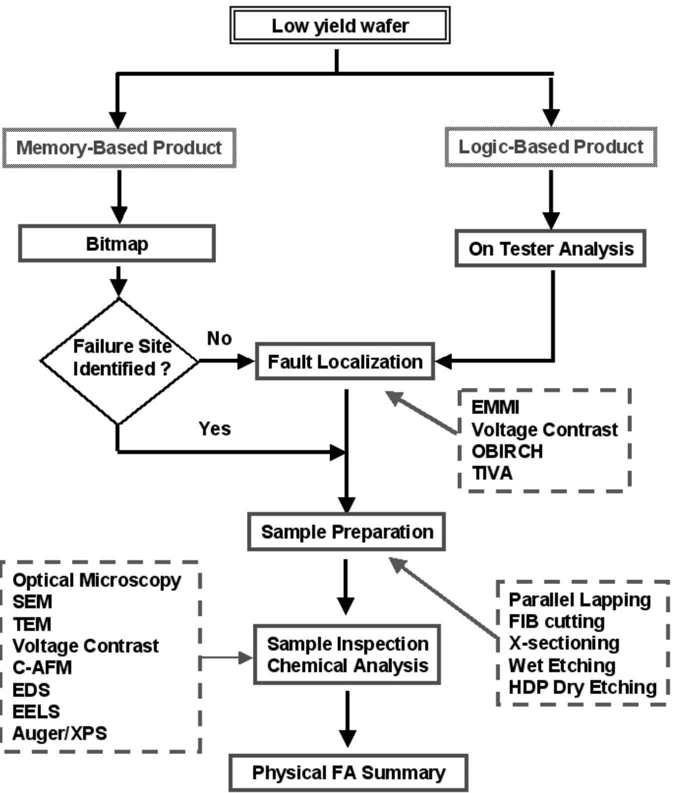


Figure 2: A typical wafer level failure analysis process flow.

Failure Mode Verification and Electrical Analysis

The aim of this step is to confirm the failure mode detected in wafer sort. In the case of purely memory based products, using different memory testing programs/patterns allows one to get a bitmap of the bits that failed some of the tests. Depending on how these bits failed, the information contained in the bitmap may be sufficient to identify the physical location of the failure. If that information is not sufficient, additional fault isolation is necessary.

The main tools used in this stage are probe stations and testers; bench testing is the principal activity at this phase. In order to be successful in this juncture it is important to have very accurate design information about the part being analyzed. In particular, accurate pin assignments, test patterns, Netlist and layout information are critical.

Initially the FA engineer tries to eliminate possible problems caused by the testing program or the testers during wafer sort. During initial product introduction and ramp-up it is very common to have frequent updates of the testing program, so use of an outdated testing program can happen fairly easily. Thus, judgement needs to be made as to the validity of the failure based on correlation of different sets of data such as probe card and program versions. If different testers were used during sorting, it is common to check if a low yielding part correlates with a certain tester and verify that the problem is not with the tester. On the physical side, it is customary to examine probe marks of low yielding wafers to eliminate possible overkills caused by improper probing, such as insufficient contact force or inaccurate placement of the probe.

In addition to verifying the failure, bench testing can be very useful in narrowing the possible cause of the failure by uncovering additional electrical data from the failing dice. Thus, generating a schmoo plot and checking the waveform of the failing die and comparing them to those of a good die can yield valuable information about any marginality issues.

Finally, it is very important to work with a good process integration engineer in this phase. Important hints about the problem can be uncovered by meticulous statistical examination of process and parametric test data. For example, comparison of WAT vs. yield and bin correlation can provide information about parameter sensitivity. Results from skew lots may establish whether there is a correlation between yield and speed. Process history and equipment correlation can sometimes reveal problems with a certain piece of equipment during a particular period of time. It is very common to also check yield or failure rates as a function of wafer sequence in a batch. Detailed examination of wafer maps also reveal crucial information about a problem.

Fault Localization

After a failure has been confirmed the next step is to determine the physical location of the failure. Since a large portion of this desk reference is dedicated to different fault isolation techniques, we will only discuss this subject in general terms. The most common approach is to start with global, static fault isolation, i.e., the failing part is powered up and/or brought to standby/quiescent state and then different signals such as emission of light, are detected to identify the location of an anomaly. Typical examples are uses of EMMI, and liquid crystal to detect “hot spots”. Thermal beam induced techniques such as OBIRCH [1], thermally induced voltage alteration (TIVA) [2] , and Seebek effect imaging (SEI) [2] are also commonly operated in this mode.

However, biasing in standby or quiescent mode, may not turn on all the different operation modes of the failing part. A more sophisticated approach is to use dynamic fault isolation. In this case a tester is docked to the fault isolation tool such as an EMMI system and different modes are triggered exciting different circuits to identify the physical failing site. One advantage of dynamic fault isolation is that the failing site can be more easily related to the electrical failure mode.

All the above mentioned approaches work effectively with so-called hard failures, i.e., failures that are not very sensitive to testing conditions such as the applied voltage. However, a

different kind of functional failures, soft failures, are notably difficult to isolate. In recent years, a new class of techniques combining dynamic testing with thermal beam induced techniques were conceived to deal with these types of failures. Techniques such as resistive interconnect localization (RIL) [3], soft defect localization (SDL) [4] and stimulus induced fault testing (SIFT) [5] all belong to this class and work by stimulating the failing part externally with localized heat and detecting changes in the pass/fail criteria in a looping test vector.

In some instances it is necessary to use even more detailed and elaborate failure site isolation approaches. For example, sometimes the global localization techniques point to a general area or circuit, but is not able to pin-point the failure to one specific location. In this scenario, a possible approach is to examine the part layer-by-layer. This results in an iterative approach in which, a layer is removed by a combination of parallel lapping, wet etching and dry etching; and then voltage contrast, e-beam probing or microprobing are used to determine the location of the failure. If the results are not satisfactory, another iteration of delayering and probing takes place until a satisfactory result is found or the entire part is deprocessed down to silicon.

As complicated and labor-intensive as these procedures may be, the multiple levels of metal layers present in the modern integrated circuit make front side examination harder with each new generation. Thus, backside variants of the techniques just described are also used.

Physical Failure Analysis

Up to now all the discussion has centered on what is commonly known as electrical FA. Once the failure site has been determined the task of the FA engineer is to identify the physical cause of the problem. Physical FA constitutes a discipline in itself. Not only does one need to be able to "see" smaller and smaller features, one needs to be able to prepare samples that upon examination will reveal the cause of the problem. For simple problems, a quick examination under an optical microscope may be sufficient However, with each new technology node, not only do the features that one needs to examine and perform chemical analyses on become smaller, new materials are also introduced to make physical and chemical analyses even more challenging. Materials such as the infamous low-k dielectric materials are not only mechanically fragile, but are also very vulnerable to exposure to electron and ion beams. To make matters worse, there are the so-called non-visible defects that defy any physical means of characterizing them. Advances in many techniques for physical analyses have kept up with many of the needs of the industry so far, however the challenges seem to appear as quickly as advances are made. Physical FA is the same whether you are dealing with wafer level or product level FA, furthermore other chapters in this desk reference are dedicated to discussions of these individual techniques, thus the subject will be left to be dealt with in those chapters.

Summary

Wafer level FA is an integral part of the operation of a wafer fabrication facility. It is unique in that the FA engineer can have a great deal of interaction with process engineers, and a lot of information related to wafer processing is available to assist him in finding the root cause. As in any type of FA work, each step of this process flow requires that the engineer exercise good professional judgement in order to succesfully identify the root cause of a problem. For the wafer level FA engineer one of the rewards of the work is to be able to see the fruit of his/her hard work in the next batch of wafers.

Acknowledgements

The authors would like to acknowledge fruitful discussions with Mr. Y.C. Lin and Mr. Y.T. Lin.

References

1. K. Nikawa and S. Inoue, *New Capabilities of OBIRCH Method for Fault Localization and Defect Detection, Proc. Of Sixth Asian Test Symposium,* 219-219 (1997).
2. E.I. Cole Jr., P. Tangyunyong, D.A. Benson, and D.L. Barton, *TIVA and SEI Developments for Enhanced Front and Backside Interconnection Failure Analysis, European Symposium On Reliability of Electron Devices (ESREF),* 991-996 (1999).
3. E.I. Cole Jr., P. Tangyunyong, C.F. Hawkins, M.R. Bruce, V.J. Bruce, R.M. Ring, and W.L. Chong, *Resistive Interconnection Localization, International Symposium for Testing and Failure Analysis (ISTFA),* 43-50 (2001).
4. M.R. Bruce, V.J. Bruce, D.H. Eppes, J. Wilcox, E.I. Cole Jr., P. Tangyunyong, C.F. Hawkins, *Soft Defect Localization (SDL) on ICs, International Symposium for Testing and Failure Analysis (ISTFA),* 21-27 (2002).
5. J. Colvin, *Functional Failure Analysis by Induced Stimulus, , International Symposium for Testing and Failure Analysis (ISTFA),* 623-630 (2002).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Failure Analysis of Microelectromechanical Systems (MEMS)

Jeremy A. Walraven, *Sandia National Laboratories, Albuquerque, NM, USA,*
Bradley A. Waterson, *Analog Devices, Cambridge, MA, USA,*
Ingrid De Wolf, *IMEC, Leuven, Belgium*

Microelectromechanical systems, MEMS, that sense, act, and "think" are rapidly becoming integral components in today's system and subsystem technologies. Estimates of the global market for MEMS are $20 billion in the year 2002 [1]. Texas Instrument's Digital Micromirror Device (DMD™) (an array of micromirrors) used for Digital Light Projection (DLP™) systems [2], Analog Device's accelerometers for airbag deployment systems [3], Hewlett Packard's inkjet print heads [4], and Lucent Technology's Lambda router [5] are only a few examples of MEMS technologies currently fabricated and applied in commercial products.

MEMS, available in many commercial products, can be fabricated using a variety of manufacturing techniques. These MEMS manufacturing techniques include (devices fabricated on the wafer surface), bulk micromachining (devices fabricated through the bulk and on the surface), LIGA (*LIthographie Galvanoformung Abformung,* lithography, electroforming, and injection molding, electrodepositing of metal into a plastic mold), and IMEMS (integration of MEMS with ICs) and polymer MEMS. Many commercial products available on the market today are fabricated using bulk, surface, or integrated MEMS fabrication techniques. Within these categories of MEMS fabrication, several materials systems and processes have been developed for several application specific requirements. Some examples of materials systems used in MEMS fabrication include three-dimensional polymer MEMS structures, thin film ceramics such as silicon carbide and diamond like carbon, thin film semiconductors such as polysilicon, thin film and bulk metals such as aluminum, gold, nickel iron, etc. New processes using materials such as diamond like carbon (DLC), silicon carbide, and polymers are being developed for very application specific purposes.

Bulk Micromachining

One MEMS fabrication technology known as bulk micromachining is the fabrication of MEMS devices via bulk materials. Here, bulk substrates are preferentially etched either isotropically or anisotropically to produce holes or channels. The means of producing these channels or holes consist of either wet or dry etching the monocrystalline substrate. In wet etching, etchants to remove material from the substrate (i.e. <100> silicon will use potassium hydroxide (KOH) attack the backside of the wafer) will create a hole through the wafer in the desired shape. The etch rate of this anisotropic approach is dependant upon crystal orientation, <111> planes etch at a higher rate than <100> planes in silicon with KOH [6], however KOH in silicon will also leave 54° sidewall geometries.. The wet etching process can be stopped by two common methods. The first is a timed etch stop. Here, the etch rate of the material is known and understood. The etch time is calculated by knowing the wafer thickness and/or how deep one wishes to etch into the sample. This technique is well understood, but has limited accuracy. The second technique is through the use of a chemical or electrochemical etch stop [7]. Here, a material is deposited on the surface of the substrate. The deposited material is more chemically inert to the etchant and is not etched or etched mildly when in contact with the solution. Etch stops are used when it is desired that a hole not be etched through the entire wafer. This technique is used to fabricate membranes for pressure sensors, inertial sensors, microfluidic channels, or other MEMS devices. Some drawbacks to using wet etch techniques include reduced accuracy in etching to the correct structure and the large openings on the backside of the substrate.

As requirements become more stringent for hole etch placement, processes were needed that required high aspect ratios structures, high etch selectivity, controlled sidewall slopes, and IC compatible technology. This form of etching often referred to as Deep Reactive Ion Etching (DRIE) is produced using high-energy ion Bombardment. These high-energy ions react with the material via ion bombardment induced chemical reactions [8]. The sidewalls stay relatively protected against reaction with radicals by absorbed atoms [9]. A modified version of this process referred to as the Bosch etch [10] is performed using a similar process as DRIE. In the Bosch etch, an iterative etch/deposition cycle is performed where a thin polymer is deposited over a lithographically defined area. The ion bombardment removes the polymer (in select areas perpendicular to the ion beam. When the trench is formed, the same polymer is deposited in the trench (trench surface and sidewalls). The next etch cycle is performed where the ion bombardment removes the polymer from the bottom of the trench more quickly than it does from the sidewalls resulting in aspect ratios of ~ 30:1. Before the sidewall polymer is completely removed, the polymer deposition process is performed. Modifications and optimization steps [11] have been added to enhance the Bosch process. Both bulk micromachining processes along with a device example are shown in Fig. 1.

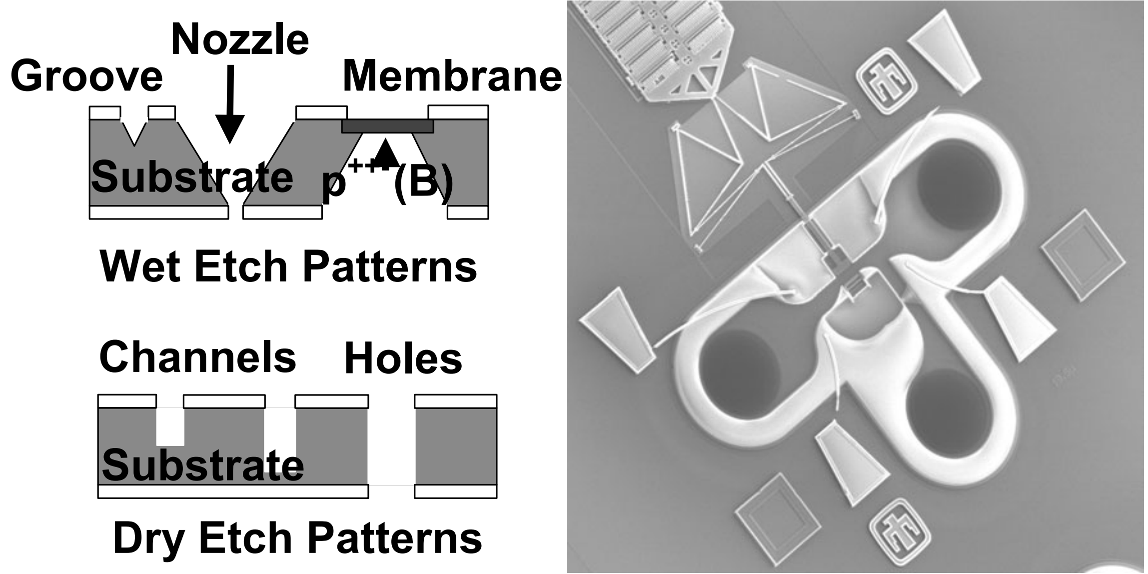


Fig. 1. *a) bulk micromachining process illustrates patterning through the bulk material to fabricate a MEMS device. b) bulk micromachined device with surface micromachined top layers form a fluid transport manipulator capable of processing various materials in a fluidic medium.*

Surface Micromachining

Surface micromachined MEMS are a branch of MEMS technology where the MEMS components are fabricated on the surface of the substrate. Unlike their bulk micromachined counterparts, surface micromachined MEMS use thin films of structural material to form the device. Many materials systems have been used to create surface micromachined structures. IC compatible materials and processes using polysilicon, aluminum, gold, diamond like carbon, amorphous carbon, and silicon carbide are just a few. The choice of materials systems is commensurate with device application and ease of process integration. Surface micromachined structures have even been added to bulk micromachined MEMS to improve and enhance their performance and capabilities.

The basic premise of surface micromachining is selectively etching sacrificial layers of material from the structural layers, where all the layers are deposited and patterned on a substrate. After patterning and deposition of the sacrificial; and mechanical layers, the remaining sacrificial material is removed via a release process, freeing up the mechanical structures. For this processes, control of the built in residual stress in the mechanical structures is paramount. Too much residual stress results in deformed structural layers evident after the release process. Residual stress will occur regardless of what type of structural material is used and will require annealing prior to release for stress relief. Other thin film mechanical properties paramount to successful MEMS fabrication include the average stress of the film, stress gradient across the film, Young's Modulus, Poisson's Ratio, and others [12].

Surface micromachining using polysilicon is performed using silicon dioxide as the sacrificial material and polysilicon as the structural material. Two well-known polysilicon processes include the MUMPS (Multi-Users MEMS Process) and SUMMiT™ (Sandia Ultra-planar Multi-level MEMS Technology) process. As an example, in the SUMMiT V™ process, five layers of structural material are used to fabricate complex MEMS structures. Poly 1 and poly 2 form a laminate layer. This layer, poly 3, and poly 4 are used for mechanical motion. The bottom layer of poly is used for electrical contact via electrostatic actuation or grounding. The poly 0 layer is separated from the substrate by thermal oxide and silicon nitride. In the unreleased state, sacrificial layers of silicon dioxide (oxide) separate the structural poly layers. The final step of the fabrication process is the removal of the sacrificial oxide layers using a hydrofluoric (HF) acid bath. The acid bath (release process) etches away the sacrificial oxide (sacox), releasing the poly structure thereby allowing it to move. Fig. 2 shows MEMS components fabricated using the SUMMiT™ process.

Fabricating surface micromachined MEMS on GaAs wafers can be performed using polyimide and metal as the sacrificial and structural materials respectively. GaAs offers higher bandwidth for Radio Frequency (RF) applications. In this technology, MEMS must be compatible with current microwave integrated circuit (MMIC) processing technologies for maximum integration levels.

To fabricate mechanical structures, polyimide was deposited and patterned on the surface of the GaAs substrate. Silicon oxynitride and photoresist were deposited and patterned over the polyimide. A reactive ion etch (RIE) was used to remove regions of the SiON and polyimide to the substrate. Gold was deposited via electron beam evaporation to form the conductive lines. The remaining gold was removed using a lift-off process. A third layer of polyimide was deposited and cured. Photoresist was deposited and patterned over the polyimide. Gold was then deposited to form the contact region. After lift off, SiON was deposited to

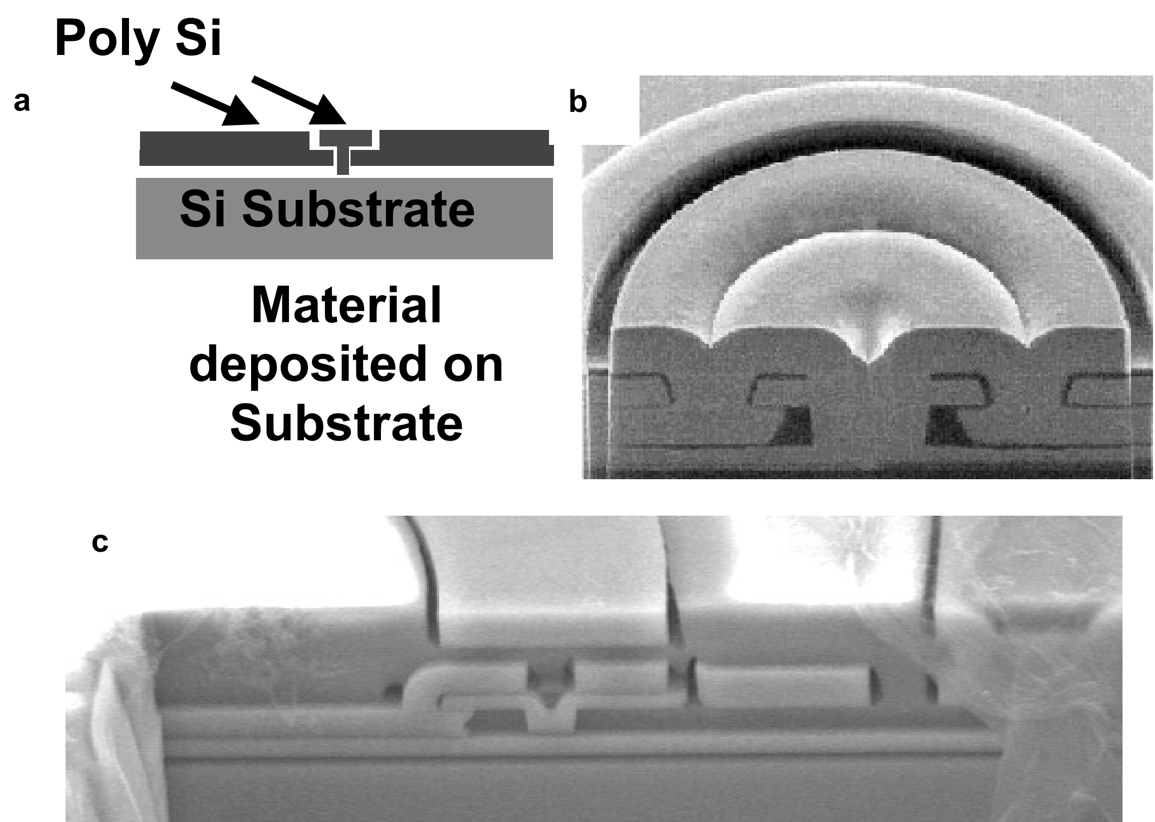


Fig. 2. a) surface micromaching process illustrating alternating layers of sacrificial and structural materials. b) cross section of a released MEMS component showing the structural material (polysilicon) and void where sacrificial material was present. c) cross section through an unreleased device showing the structural and sacrificial material (note this device was prepared using a dilute BOE etch to delineate the features).

form the fixed-fixed structural component. Photoresist is deposited and patterned to allow deposition of Au or Al on top of the SiON to form the top electrode. Al deposited over selected regions of the surface forms the top electrodes [13]. The polyimide acts as the sacrificial material, containing the fixed-fixed beam until ready for release. The RIE barrel etch was used to free the structural components and is preferred over a wet chemical process to avoid stiction and adhesion related failures during the release process. An example is shown in Fig. 3.

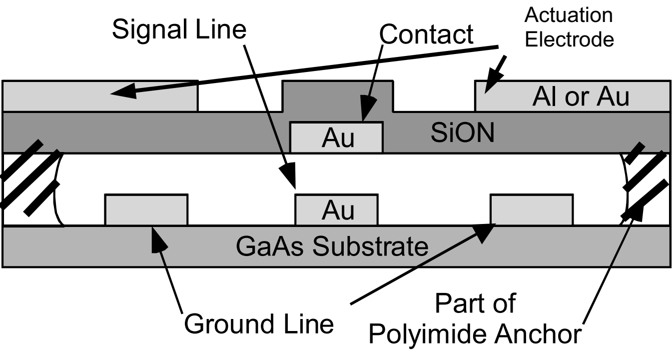


Fig. 3. Cross section diagram of a metal series contact switch revealing the contact and signal line metallization and the top and bottom electrodes.

Diamond or diamond-like MEMS materials can be used when applications such as surface rubbing, bearing, or impact. Diamond like materials offer excellent tribological properties such as low stiction, chemical inertness, and high elastic moduli [14] making them excellent candidates for devices where failure due to wear is of concern. In amorphous carbon, the material is deposited at room temperature with a pure carbon beam with the carbon ions energized near 100 eV [15]. After the film is deposited, the residual stress is relieved using a 600°C anneal. SiO2 is used as the sacrificial material between amorphous carbon layers. A timed wet etch is needed to form the base to hold the MEMS components in place. An example of an amorphous carbon surface micromachined MEMS is shown in Fig. 4.

Polycrystalline or nanocrystalline diamond have several properties that make it a material better suited for friction bearing MEMS applications. In the crystalline state, diamond has the widest electromagnetic radiation transparency of materials ranging from UV to far IR [15]. The major difficulties associated with diamond MEMS include process integration, and ease of manufacturing. For the polycrystalline MEMS (grains in the range of 1 μm), typical fabrication processes include hot flame chemical vapor deposition (CVD) and microwave plasma CVD. For nanocrystalline diamond, microwave plasma CVD with 98%

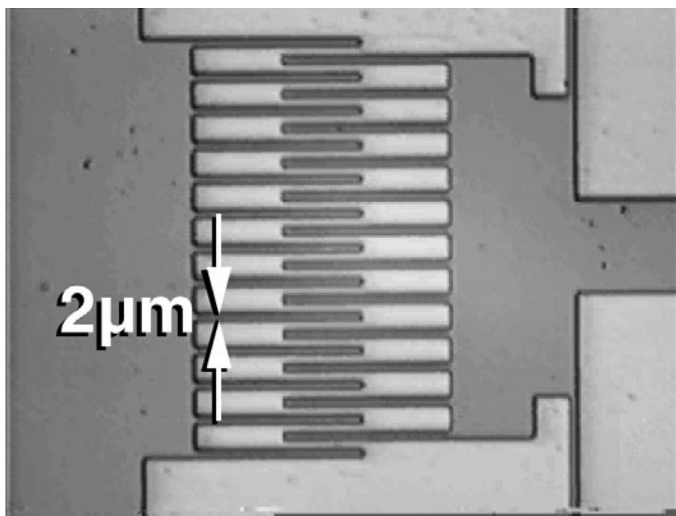


Fig. 4. Electrostatic actuator fabricated using amorphous carbon.

Ar, ~1% H_2, and the remainder methane is used to produce these films with ~ 3-5 nm grain sizes. Diamond can also be deposited on existing polysilicon MEMS structures to improve performance and materials properties without micromachining a new device.

Like diamond, silicon carbide is a desirable material for MEMS applications. SiC is known for having a good combination of electrical, mechanical, and chemical properties and would be especially useful in areas where environments are too harsh for polysilicon [16]. Similar to diamond, SiC can be deposited iin thin films to make up a surface micromachine, or it can be deposited on a pre-existing device to improve materials and device properties.

Initially, problems with using SiC as a MEMS material dealt with depositing it on a substrate (Si or polySi). New techniques have been developed to deposit SiC films via CVD. One new technique developed at UC Berkeley [17] uses a single precursor molecule 1,3-disilabutane (DSB) deposited at temperatures between 800 – 1000^oC. The precursor is in liquid form aith a vapor pressure of ~ 20 mTorr. This technique is performed at pressures between 10^{-4} and 10^{-5} Torr [17]. Conventional SiC processing techniques involve considerably expensive gas handling and delivery systems, along with the use of dangerous gases (SiH4 and acetylene). SiC films are generally deposited on single crystal or polycrystalline substrates. SiC has been deposited on SiO_2 substrates, however, results have shown [18] that the SiC films tend to be more porous than their Si substrate counterparts. For silicon based substrates, a KoH bath can remove the silicon, freeing the SiC device. Other techniques such as a molding process developed at Case Western University uses a mold fabricated out of polysilicon or SiO_2. SiC is deposited into the mold, and KoH or HF is used to remove the mold material leaving the SiC structure intact [18]. An example of a SiC coated MEMS structure is shown in Fig. 5. Applications for this technique as well as diamond include harsh environments with combustible fuels, fluids, friction bearing surfaces, and ultra-high temperatures (>1100^oC) to name a few.

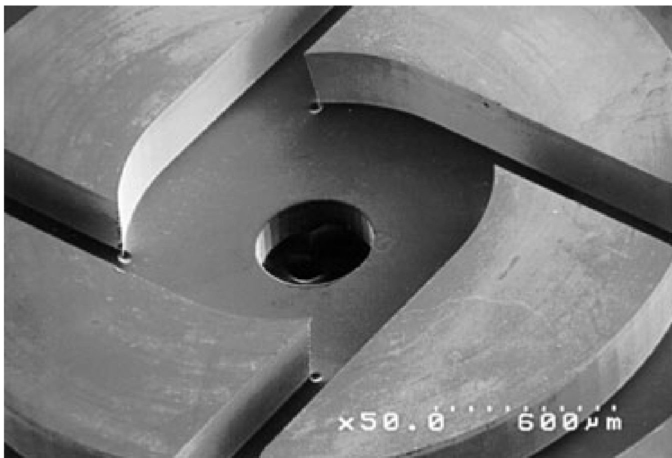


Fig. 5. SiC coated polysilicon based atomizer designed to swirl fluid down the hole. After several runs, no damage is evident on the surface unlike it's polysilicon counterparts. Image courtesy of Case Western University [18].

LIGA

LIGA (Lithographie, Galvanoformung, Abformung) is a German acronym for lithography, electroplating, and molding. All three processes are used in the fabrication of metallic based mechanical structures. The LIGA process is capable of creating defined features with 1-10 μm minimum lateral feature sizes, and thicknesses ranging from hundreds of microns to a few millimeters thick. In this process, x-ray synchrotron radiation acts as the lithographic light source. The high energy x-rays from the synchrotron impinge on a patterned mask in proximity to an x-ray sensitive photoresist (typically polymethyl methacrylate PMMA) [19]. The masks used for LIGA are non-standard due to the ex-ray synchrotron beam line used, lateral feature sizes, height, and allowable tolerance. The masks themselves can be fabricated out of thin, x-ray transparent supported by a thicker substrate that is etched away, or a substrate mask. This mask is an entire wafer that has been processed with the desired LIGA pattern in it. The PMMA is typically attached to a substrate that is used as the base for electroplating. The areas of PMMA that are exposed to x-ray radiation are selectively dissolved in a developer and rinsed away. The remaining non-conductive PMMA mold is used to create the mechanical components comprising the LIGA MEMS structure. Other methods used to create molds similar to LIGA include UV radiation [20], DRIE [21], and laser ablation [22].

The PMMA mold is filled with metal by elecrodeposition from a conductive base. Typical metals used in the electroplating process consist of gold, copper, nickel, and nickel iron. In LIGA processing, the current needed for electroplating is much less than that required for large scale electroplating typically used for protective coatings [19]. After electroplating, the wafer must be planarized to remove surface non-uniformities from the plating process. The

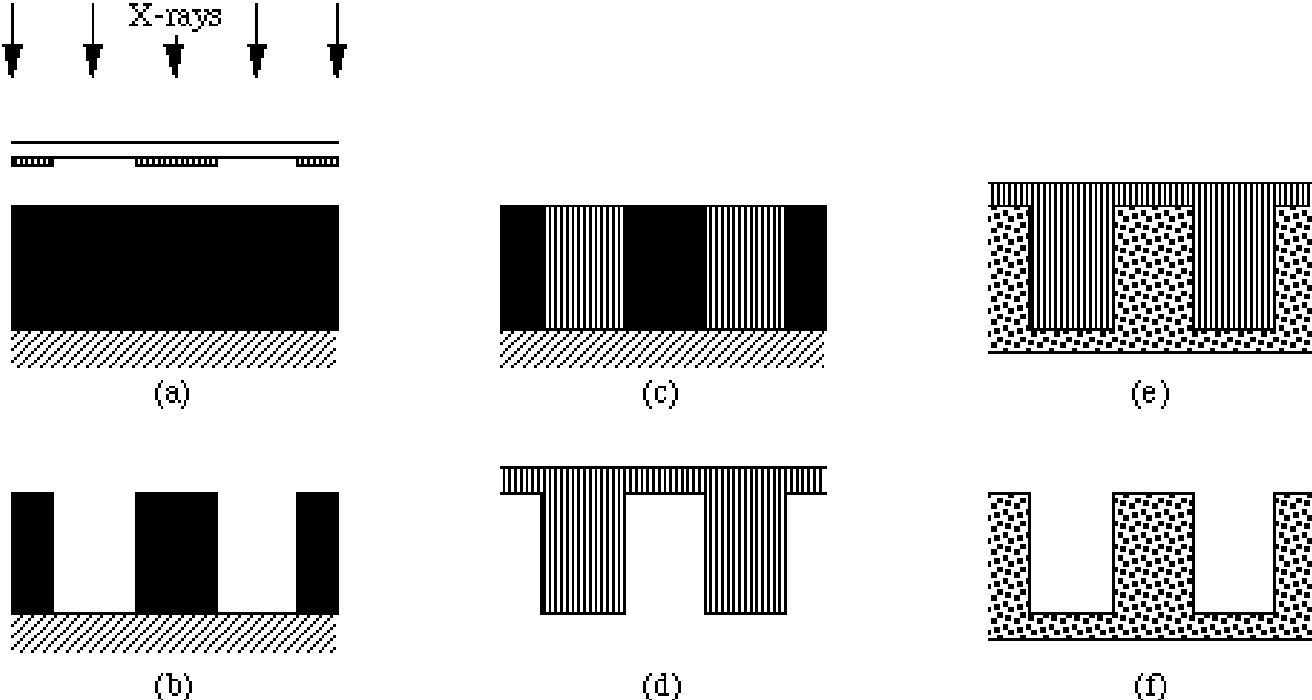


Fig. 6a-f. LIGA fabrication process. a) X-rays impinging on a mask, b) removal of the x-ray exposed PMMA, c) electrodeposition of metal into the PMMA mold, d) metal structure with the PMMA removed, e) metal structure in the PMMA mold, f) PMMA mold. Photo courtesy of Daniel Banks found at: http://www.dbanks.demon.co.uk/ueng/liga.html [23].

mechanical structures produced can be the final product, however it is common to produce a metal mold. This mold can then be filled with a suitable material, such as a plastic, to produce the finished product in that material. The LIGA process is shown in Figs. 6a-f. This process of using x-ray synchrotron radiation lithography and electroplating and replication are the cornerstones for MEMS LIGA based MEMS fabrication. Fig. 7 shows a LIGA based actuator made out of iron nickel.

Fig. 7. LIGA fabricated iron nickel actuator.

Integrated MEMS

This technology base of MEMS fabrication is unique in that it is essentially a hybrid or mix of MEMS based mechanical structures designed to produce motion or a response with typical integrated circuitry. The integration of MEMS technology with IC's on a single chip offers considerable advantages compared to two separate units. By leveraging current IC processing technology, a fast, reliable and repeatable manufacturing process can be established. Arrayed systems can be manufactured on the same wafer or (depending on device footprint) on the same die.

The processes designed to manufacture these components varies greatly depending upon materials systems used. Many manufactures of IMEMS technology use either a polysilicon MEMS based surface micromachining process where the MEMS are fabricated prior to the IC portion, or thin film surface micromachining process where the metallization can be processed prior to or post IC processing. In the case of polysilicon MEMS structures, the polysilicon mechanical structures must be fabricated first to ensure the high temperature anneals needed to relieve the stress are performed. This anneal would be very difficult to perform in a post IC processed structure due to the low melting point and damage that would be caused to the thin film metallizations.

In the polysilicon integrated MEMS process, the polysilicon is processed as it would be in typical surface micromachining. Some manufactures may have their device fabricated in a trench where others may have it directly on the surface. In either case, the MEMS are fabricated and the high temperature anneals are performed to relieve the stress from the poly. The polysilicon MEMS region is capped with oxide. The circuitry is processed around the device with the proper

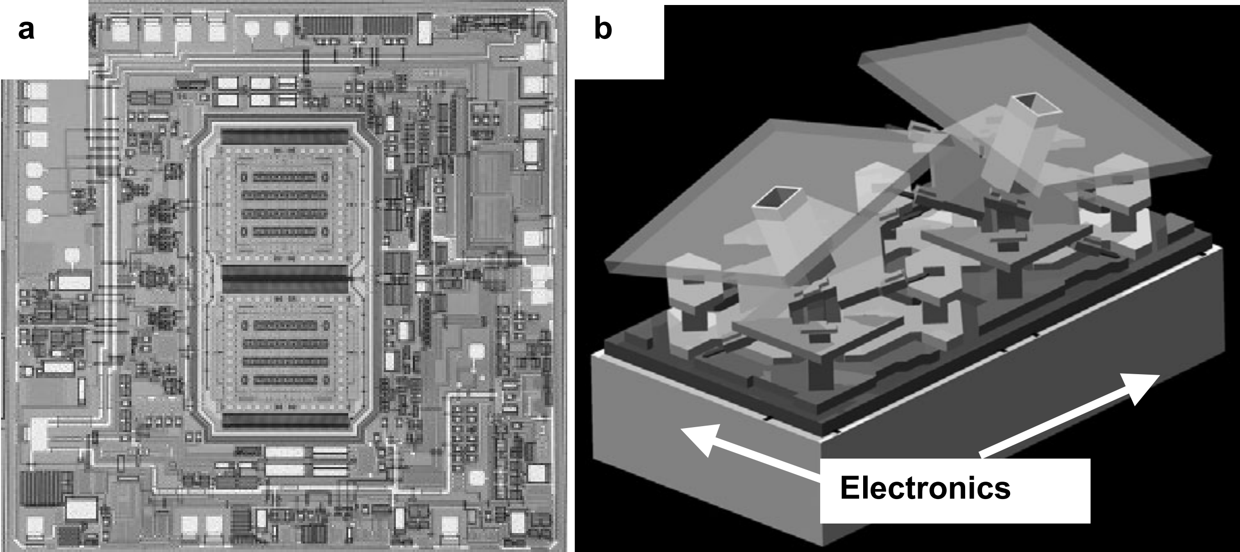


Fig. 8a and b. a) polysilicon based IMEMS gyroscope fabricated at Analog Devices with the MEMS structure in the middle surrounded by the IC circuitry. Photo courtesy of Analog Devices. b) Graphic representation of a Digital Mirror Device DMD™ fabricated directly above the controlling electronics. Image courtesy of Texas Instruments.

interconnections made from the MEMS to the IC. After the IC portion is fabricated, it is protected from the release needed to free the MEMS elements while maintaining the IC integrity.

In the metallized MEMS, the same metal that is used in IC processing (Aluminum) can be used to make mechanical structures. Some advantages for using metallized MEMS over polysilicon MEMS include highly reflective properties for optical applications, process integration with existing CMOS processes, the ability to have on chip circuitry either to control the MEMS motion, or sense the movement of the MEMS structure. One example of an integrated MEMS component is CMOS-MEMS designed and fabricated at Carnegie Mellon University [24]. Their CMOS-MEMS process produces microstructural sidewalls by stacking the drain/source contact cut and metal via cuts in the CMOS and removing the metallization layers above the cuts [25]. This technique allows use of both the metal and dielectric materials to be used as mechanical structures. Commercial metal MEMS devices such as the Digital Mirror Device (DMD™), are designed and fabricated with the mirror component directly above the IC. In this device, the IC controls the operation of the DMD™. Examples of integrated MEMS technology are shown in Fig. 8a and b. Note the circuitry fabricated around the MEMS based gyroscope produced by Analog Devices, while the Texas Instruments DMD™ has the circuitry (not shown) fabricated directly underneath the device.

Polymer MEMS

Polymer MEMS is a relatively new class of MEMS fabrication and materials technology where polymers are used for micromotion and sensing based applications. These components offer distinct advantages over their poly, diamond, SiC, and metal based MEMS. These include low Young's modulus for bending applications, high elongation, moldability and ease of fabrication, and most importantly, bio-compatibility [26]. Polymer MEMS can also be integrated with polymer electronics to create low cost, flexible, biocompatible, and even disposable units.

Until recently, polymer capability did not extend much beyond thin film membranes for microfluidic applications. Recent research and development have led to MEMS based polymers fabricated in true 3-dimensional architecture. The fabrication of polymer MEMS consists of a structural polymer and a sacrificial one. The structural polymer is usually UV curable. Polymers ingredients such as urethane acrylate, epoxy acrylate and acryloxysilane all have low viscosity and allow them to be processed without the use of solvents [27]. Polymer MEMS have been used to fabricate sensing and actuating components and polymer strain gauges [28].

One technique used for fabricating polymer based MEMS is microstereo-lithography (MSL) [29]. Using MSL, polymer MEMS can be fabricated by focusing a laser beam with a spot size of ~ 1 μm directed onto the resin surface to begin polymerization [27]. Through repeated scanning of the light source or different structural layers, a true 3D polymer MEMS component can be fabricated.

MEMS Classification

Many types of MEMS are currently employed for a variety of applications. To properly categorize these devices and those under development, a general taxonomy of MEMS has been developed. This taxonomy divides MEMS into four classes. They are: **Class 1** - devices with no moving parts. This class consists of accelerometers, pressure sensors, ink jet print heads, and strain gauges. **Class 2** - MEMS that contain moving parts without rubbing or impacting surfaces. This class of devices includes gyros, comb drives, resonators, and filters. **Class 3 –** devices that contain moving parts with impacting surfaces. These devices consist of relays and valve pumps. **Class 4** - devices that contain moving parts with

impacting and rubbing surfaces. This class of MEMS consists of shutters, scanners, and gear trains [30].

To assess MEMS, experiments intended to test their functionality, reliability, and materials properties are paramount. For a failure analyst, experiments yielding knowledge and insight into failure modes will supply critical information needed to improve the reliability of the device. These improvements may be performed at the design, fabrication, or test phase to obtain the desired level of reliability. The quest for the MEMS failure analyst is to develop and utilize tools and techniques to determine the failure modes from design, fabrication or testing.

The first step in diagnosing any failure mechanism is the performance of continuity tests. By assessing device functionality one can determine if the failure mechanism is either electrical or mechanical in origin. After identifying the failure mechanism, the failure analysis engineer is left with the task of locating the failure site and determining the root cause of failure quickly, accurately, and cost effectively. Once the failure site and root cause have been identified, the results are fed back to reliability engineers, design engineers, package engineers, and/or process engineers. The results are evaluated and actions are taken to improve device reliability. Failure analysis determines if actions are required at the design, fabrication, or test phase.

Many types of MEMS are currently available or are under research and development. These MEMS can be incorporated into any one of six different technologies. These MEMS technologies include RF MEMS for switching signals, optical MEMS for displays and switching optical signals, microfluidics for fluid transport and ejection, biological MEMS (biomems) for fluid mixing and injection, sensors for detecting changes in force or environment, and actuators to produce motion. This chapter will discuss the various failure mechanisms found in different MEMS technologies and FA tools and techniques used to diagnose the failure mechanisms.

Failure Modes/Mechanisms

The failure modes of MEMS in response to a failure mechanism are difficult to predict because MEMS vary greatly in design, use, and technology/materials base. The failure analyst should retain accurate failure mode and mechanism data and attempt to build a reliable electrical failure signature database for a given MEMS design and mechanism interaction.

Particles/Obstruction

MEMS rely on the movement of mechanical elements to perform their designed operation. Any obstructions to that movement will render the device inoperable. Particles are considered obstructions in the case of MEMS, unlike typical IC failures where they are embedded in the circuitry. Particles usually rest on the surface of the sensor/MEMS element. In most cases the particle is either wedged down between the sensor elements or resting on the top. The case illustrated in Fig. 9 is extremely severe and the particle has physically lifted the moveable beams out of plane. Fig. 10 is an example of particulate material accumulated on and obstructing a micro-engine gear set. The particulate material, in this case, was the result of wear in the gear hubs.

From a failure analyst's perspective, particles are relatively easy to diagnose once the device is decapsulated and the element is exposed. A simple optical inspection with enough magnification to resolve the micromachined features should be enough to correctly diagnose particle failures. Exposing the MEMS element for optical inspection, however, is one of the more challenging tasks when working with particle failures. Surface contamination from improper decapsulation/delid will cloud FA results. The failure analyst should also reduce the amount of time between decapsulation and optical inspection, in order to minimize any surface contamination from the laboratory environment.

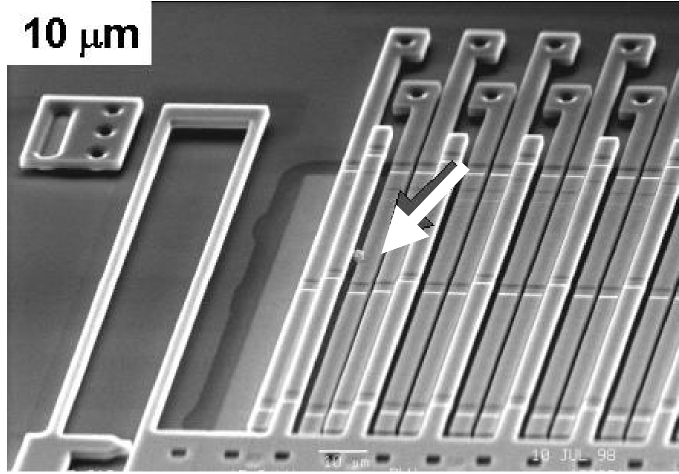


Fig. 9. A particle wedging a MEMS sensor element out of plane.

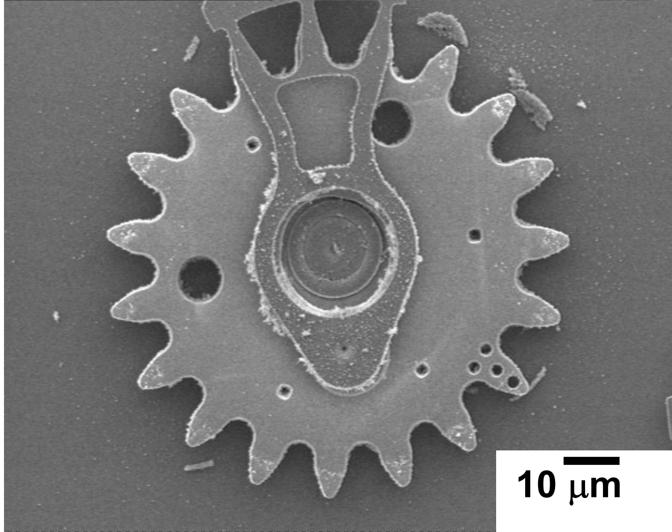
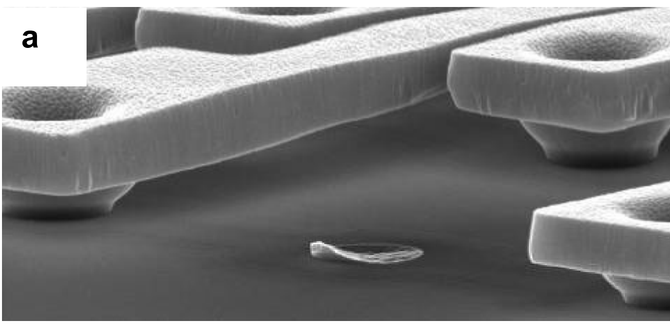


Fig. 10. Wear derived particulate material impeding micromachine gear train operation.

Fracture

Fracture is a consideration with MEMS as well as standard IC's. While fracture is generally limited to die cracking for standard IC's, portions of a MEMS device, when subjected to high mechanical stresses, may actually free themselves from the surface of the die. Figs. 11a and b show both fracture surfaces from a failed cantilever MEMS beam. This beam was most likely exposed to a concentrated point load as the surrounding structures are intact.

Diagnosis of the failure mechanism is easily accomplished with an optical microscope. The fractured piece will either be missing or redeposited onto another portion of the die. Characterization of the fracture surfaces using SEM will yield clues, such as the crack initiation site and loading area, and may assist the analyst in determining the root cause of the stress event.



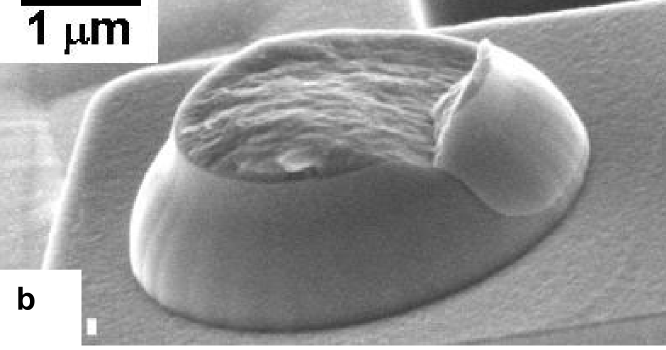
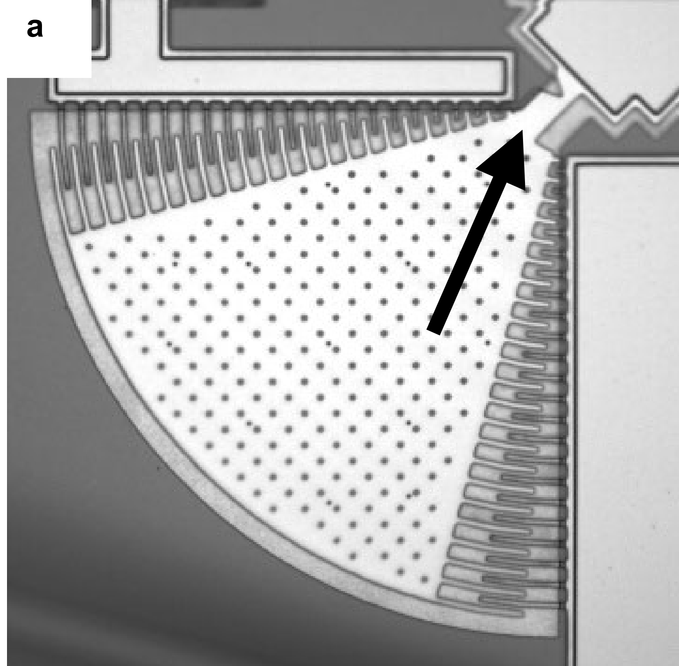


Fig. 11. SEM images showing the fracture surface of a failed anchor. **a)** Substrate region revealing a fractured component, and **b)** cantilever beam segment showing the other half.

Fatigue

Unlike most of the mechanisms discussed in this chapter, which are universal to surface and/or bulk micromachined devices, fatigue is limited mostly to active elements such as comb drives or resonators. Sensors and other passive structures, by design, are not exposed to enough mechanical stress cycles to induce fatigue.

Fatigue begins with a crack initiating at an area of high stress concentration and slowly growing through the material until failure occurs. Figs. 12a and b show a test specimen used to characterize the fatigue life of polysilicon. The notch,



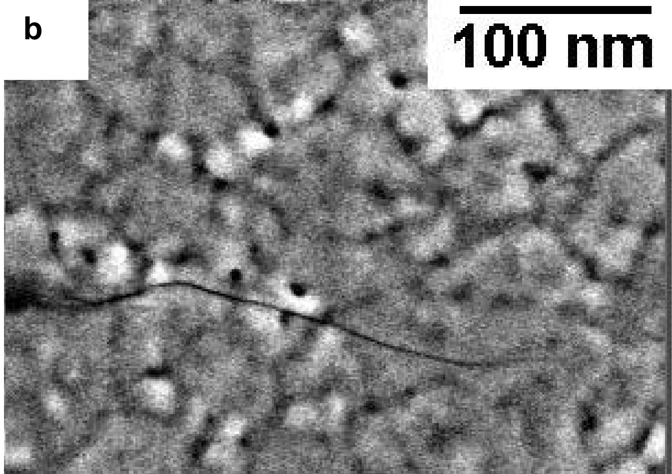


Fig. 12. **a)** SEM image of a resonator designed to test polysilicon for its fatigue life. **b)** A fatigue initiation crack in a specimen similar to the one above.

highlighted by the arrow in Fig 4a, was fabricated into the specimen to create a stress concentration site. In a fatigue event, a microcrack (Fig. 12b) will initiate from the point of stress concentration and will propagate through the material with subsequent loading cycles until the stress is great enough to cause fracture. Fatigue of MEMS has almost exclusively been limited to specific structures designed to model polysilicon MEMS fatigue behavior. Currently available MEMS were designed to minimize conditions that induce fatigue.

If fatigue is suspected as a mechanism, hermeticity testing should be performed prior to element exposure as moisture can play a large role in fatigue of polysilicon. Once the MEMS device has been exposed, a catastrophic fatigue failure will be visible with optical inspection. If the element has not

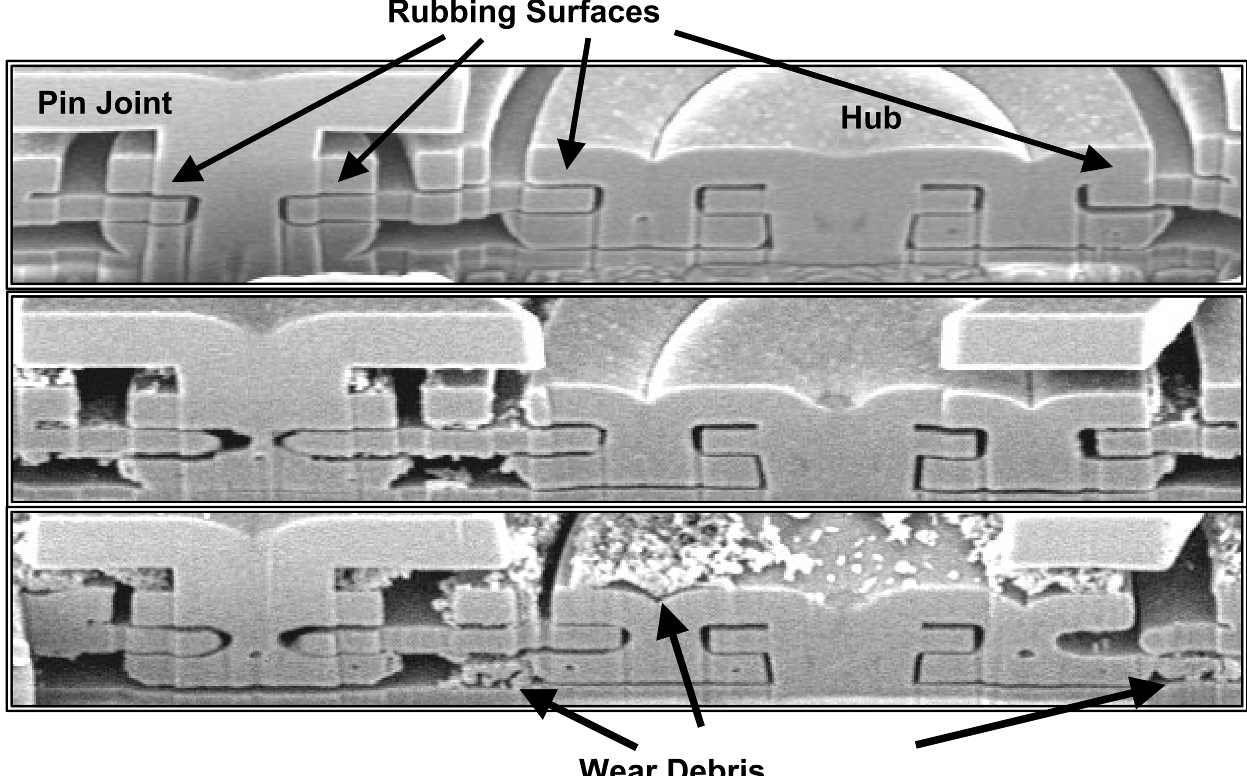


Fig. 13. SEM images of the control, the 39% RH sample and the 1.8% RH sample illustrate the amount of wear debris created in each experiment. Arrows indicate the rubbing surfaces. In both stressed samples, the pin joint has been worn down from its fabricated 3 μm diameter.

yet failed catastrophically, high magnification SEM inspection will probably be required in order to see the crack.

Wear

The effects of wear on MEMS with rubbing surfaces (class 4) can significantly impact the functionality and reliability of a device. Wear is detrimental to the operation of a device and can occur in various ways. Two mechanisms of wear found in MEMS include conventional wear, or wear of contacting surfaces, and third body wear. Conventional wear occurs when two surfaces come into contact, rub against one another and lead to failure by adhesion (sticking together of contacting surfaces). Third body wear occurs when a foreign particle or piece of material is introduced between two contacting surfaces. This piece of material can be in the form of a broken fabricated component, foreign particles, wear debris, etc.

It has been shown that humidity is a strong factor in the wear of rubbing surfaces in polysilicon micromachines [31, 32]. It has been demonstrate that very low humidity can result in large amounts of wear debris being produced while micromachines operated under the same accelerated conditions at high humidity do not produce as much wear debris. As illustrated in Fig. 13, FIB cross sections of microgears worn to failure at 1.8% RH and 39% RH reveal the effects of humidity and third body wear on a class 4 device. Here, the third body material present and enhancing the wear process is debris originating from the rubbing surfaces. A control microgear was cross-sectioned to illustrate the as-fabricated dimensions prior to reliability testing. Fig. 14 shows a microgear that failed due to adhesion of rubbing surfaces. The sidewalls of the pin joint have been worn in a non-symmetrical pattern. The sidewall of the gear was essentially "driven" into the pin joint during operation. This resulted in a locked microgear, which caused the microengine to fail by seizing up. The microengines used in this experiment were operated under accelerated conditions of 1740 Hz or 104,400 RPM at environments of 1.8% RH up to 34% RH.

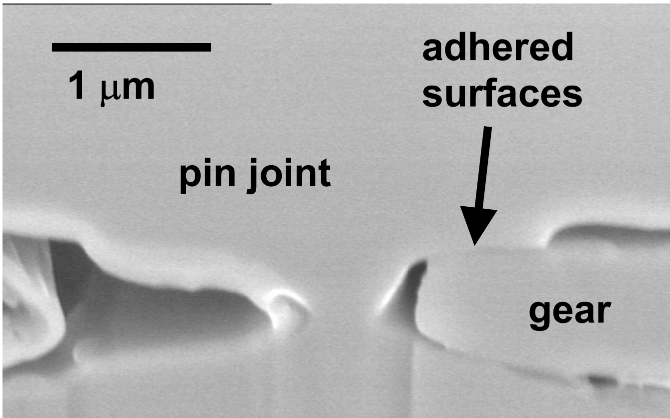


Fig. 14. SEM magnified view of the pin joint region. A FIB cross section shows the area where the two surfaces adhered causing the microengine to seize. This microengine was tested at 10% RH at 25°C.

The mechanism for wear in these polysilicon devices was oxidation along rubbing polysilicon surfaces. EDX analysis and electron beam diffraction of the wear debris revealed silicon and oxygen. Electron beam diffraction analysis of the worn material revealed no crystallinity. This finding demonstrates that polysilicon is not broken off the device and oxidized during third body wear. Wear debris produced from such an event would have a crystalline core where this wear debris does not. This evidence supports our hypothesis that amorphous oxide is removed by rubbing of polysilicon surfaces.

The formation of wear debris can damage other devices in close proximity by the generation of particles. Worn material can be ejected from the device during operation. These wear particles can interact with other MEMS rendering them inoperable.

Stiction

Many classes of MEMS are designed such that the moving elements should not come into contact with each other under normal operating conditions. If a device experiences an overstress condition, the micromachined surfaces may deflect far enough to contact each other or adjacent MEMS surfaces. When contact occurs between MEMS surfaces, adhesive forces can keep them stuck together. The adhesion force depends on the surface states and contact area of the MEMS structures. This adhesion phenomenon is called stiction.

Stiction occurs when surfaces come into contact and restoring forces designed into the MEMS are not great enough to overcome the surface adhesion. Contact of the surfaces can result from mechanical shock or from electrostatic capture due to an electrical overstress event. Fig. 15 depicts a laterally stuck beam system on a MEMS accelerometer where the moveable portions of the sensor are stuck to the adjacent fixed portions. Fig. 16 is an SEM micrograph showing a vertically stuck device with the moveable silicon displaced vertically and contacting the ground plane of the device.

The failure analyst can easily diagnose a stiction event with high magnification optical or SEM inspection. The main concern, however, if stiction is suspected, is recovery of the reject unit due to handling, analysis, or shipping. While the stiction event was most likely the result of an overstress condition, a subsequent shock or overstress during failure analysis can provide the restoring force (externally) needed to overcome the surface adhesion and promote release of the stuck MEMS structure. Once a stiction failure has recovered, there is generally little physical evidence remaining to indicate that stiction was present in the system.

Stiction was a great concern in the infancy of MEMS. Improvements in MEMS design, such as the incorporation of contact area reducing structures (dimples) and the development of anti-stiction coatings, have given way to some extremely stiction-robust MEMS devices today. The combination of these design improvements and increased customer awareness to MEMS handling have virtually eliminated stiction in end-use applications.

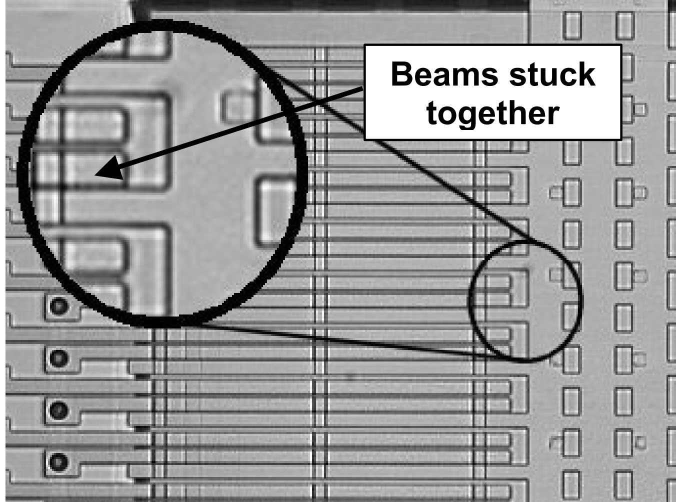


Fig. 15. Optical micrograph of laterally stuck MEMS accelerometer.

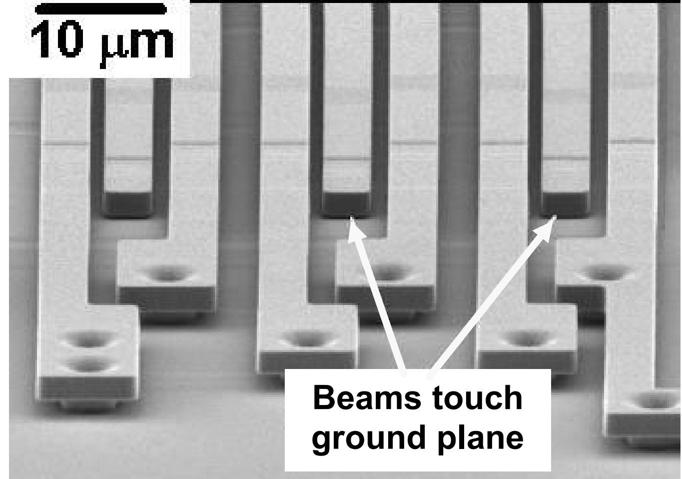


Fig. 16. SEM micrograph of vertically stuck MEMS accelerometer.

Electrostatic Discharge

EOS/ESD events can severely damage ICs and other semiconductor and electronic devices [9]. The severity of this damage mechanism is evident by the extensive development of ESD handling procedures and protective circuitry. This failure mode was not previously addressed on any class of MEMS (bulk, LIGA, or surface micromachine) due to the proprietary nature of the device or fabrication process, or the close resemblance of ESD failure mechanisms to stiction-adhesion of critical functional components rendering the device inoperable. Some MEMS are fabricated directly over or around integrated circuits. MEMS such as the accelerometer fabricated by Analog Devices [3], and the Digital Mirror Device 'DMD' fabricated by Texas Instruments [2] use on-chip integrated circuits for detection of motion, stimulating mirror motion, and ESD protection via ESD protective circuitry.

Most MEMS operate using some kind of electrostatic force or actuation. This operation typically requires voltages greater than those used in conventional CMOS circuitry to create strong electric fields. The actuation of these devices using electrostatic forces is found to be susceptible to ESD events. ESD testing has revealed damage along various devices after testing [33, 34]. As shown in Fig. 17, a torsional ratcheting actuator (TRA) was tested. The results revealed a design flaw in the TRA. The failure mechanism found on the TRA was the same for each ESD test; compliant regions of the device fractured when the ESD pulse was given. As a result, the fractured components (conductive polysilicon particles) present in the active regions bridged the biased component to the grounded component creating a short. (Introduction of large broken components could also become lodged in the device.) As shown in Figs. 18a, b, and c, the failure mechanism for HBM, MM, and CDM ESD tests is a short caused by particle contamination. The location of the failed compliant component is shown in Fig. 3d. No ESD failure voltage information could be obtained on this device from these ESD tests due to a design error.

Another example of ESD susceptibility of MEMS is in an electrostatically actuated microengine. The failure mechanism found in this device was contact of a poly 1 comb finger to the ground plane. Closer examination of the failure sites revealed damaged and melted silicon along the front bottom edge of the comb fingers. Damage was also observed on the ground plane where contact occurred. The spot weld damage shown in Figs. 19a-c occurred on all ESD tests except CDM. We

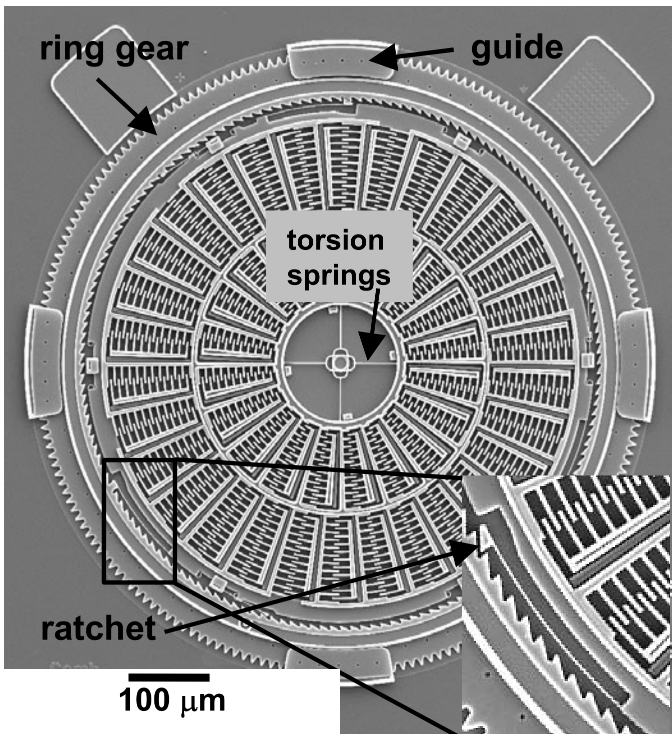


Fig. 17. SEM image of a torsional ratcheting actuator (TRA). The inset shows an enlarged view of the ratchet gear and curved comb fingers used for electrostatic actuation.

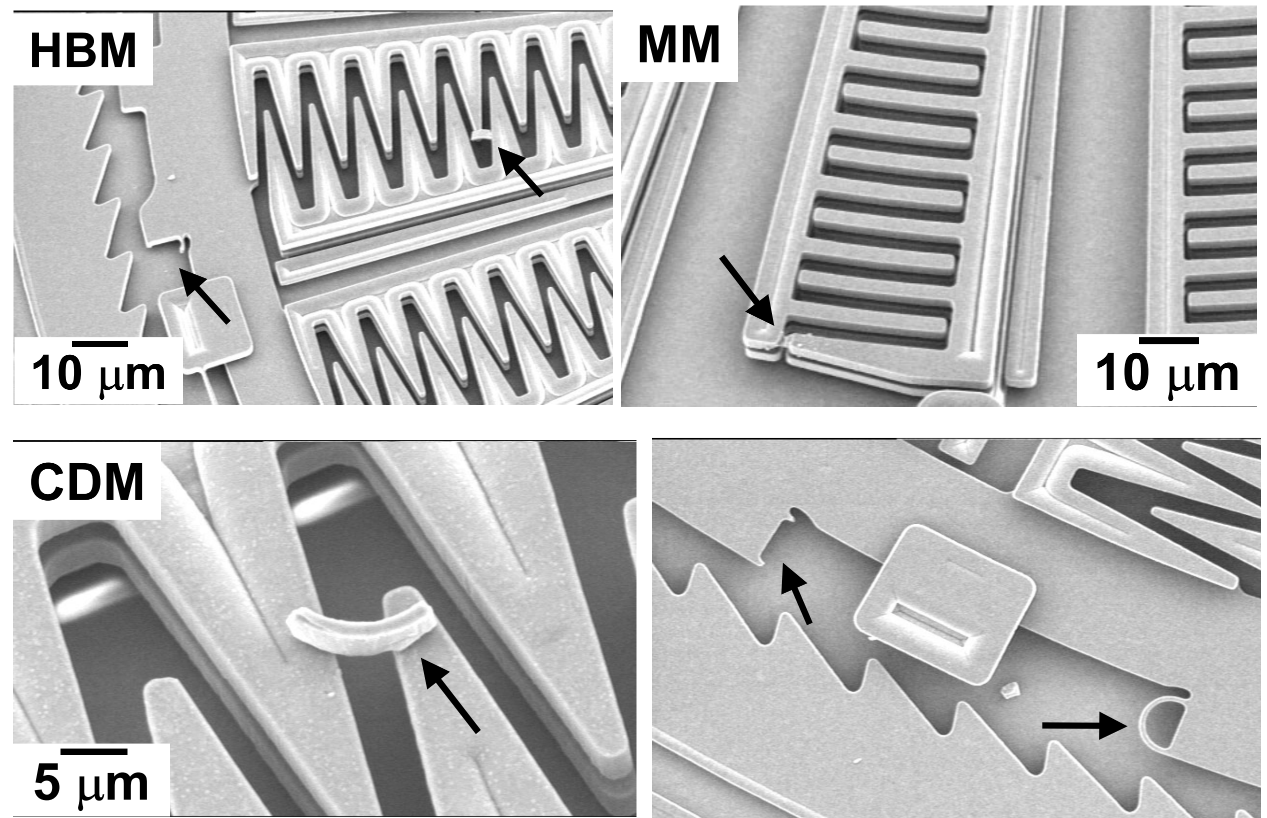


Fig. 18. SEM images of failed TRAs after **a)** HBM, **b)** MM, and **c)** CDM ESD testing. **d)** Flexible structures shown after ESD testing. Note the broken element (left side) compared to its counterpart (right).

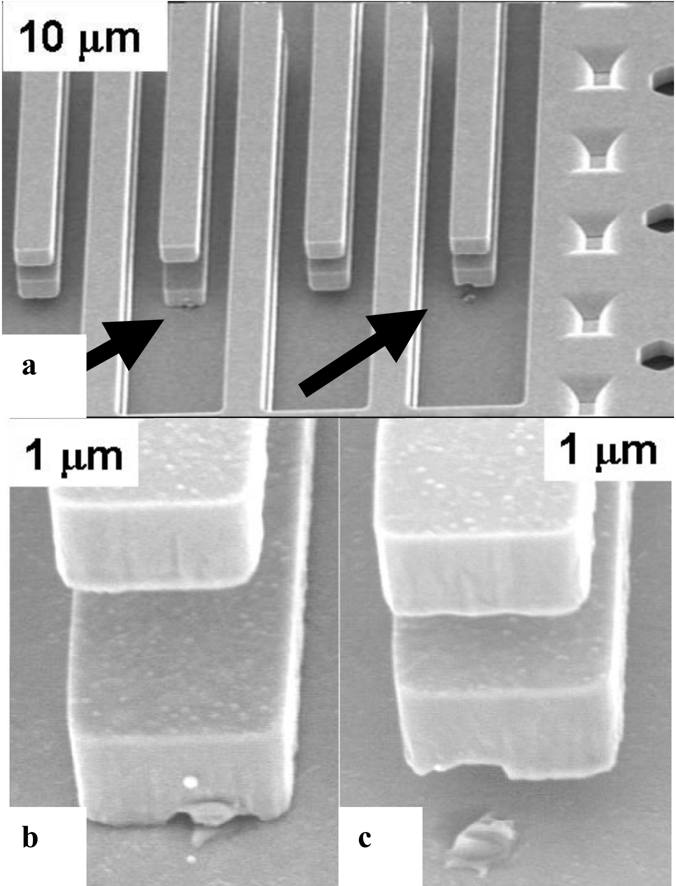


Fig 19. a) SEM image of a failed comb drive. Failure analysis revealed **b)** a MEMS element spot welded to the ground plane, and **c)** a MEMS element that contacted the ground plane and popped back.

believe CDM testing did not affect the structure due to the response time of the comb fingers to the ESD pulse [34].

ESD testing has proven to be valuable in assessing the susceptibility and robustness of any device. ESD tests have shown weaknesses in design and manufacturing which have led to improvements in design tolerance (spacing) and component rigidity. Both these improvements enhanced the device robustness to ESD.

Oxide Charging

Charging of dielectric material is especially problematic in RF-MEMS capacitive switches and tunable capacitors [35, 36]. Also MEMS with large regions of dielectric material present under the active MEMS components may be susceptible to failure due to dielectric charging. One manner in which dielectric charging leads to failure in for example a RF-MEMS capacitive switch is through a change in the pull-in voltage. This is caused by the large electric fields (1 to 3 MV/cm) across the dielectric, due to the high actuation voltages (20 to 50 V). As the charge builds up, a screening voltage is formed which results in a decrease of the voltage required to pull the metal of the switch down to the dielectric. Since the voltage required to keep the switch down is much smaller than the actuation voltage, the trapped charges may provide enough attraction to keep the switch down. The switch then fails due to stiction (Fig. 20). The effects can be seen in-situ (optically) and measured electrically. SEM can be used to investigate the stiction site [37]. Mostly this charging effect is not permanent and the devices can function again after some time.

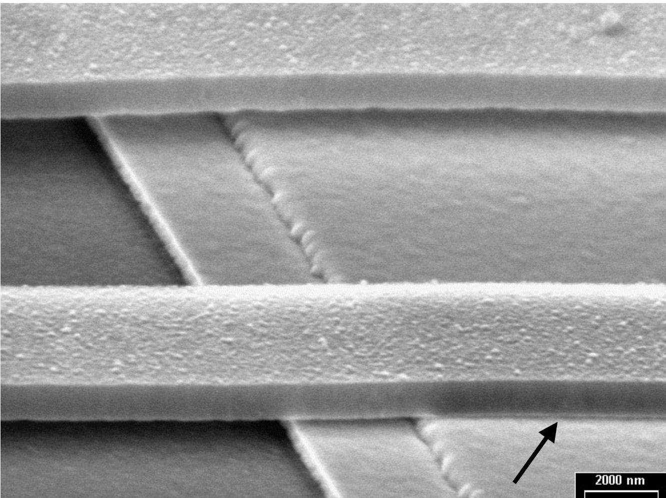


Fig. 20. SEM micrograph of a stuck (arrow) part of an RF MEMS bridge (from [37]).

Another failure mechanism attributed to dielectric charging is dielectric breakdown. It causes permanent damage and is irreversible. In a typical switch (or other MEMS structure), a voltage is applied activating the mechanical structure in one form or another. When a significant amount of charge accumulates in the dielectric material, the voltage applied to activate the device is not sufficient. More voltage is required which in turn supplies more charge to the dielectric material. After sufficient increases in voltage, the dielectric eventually fails (typically due to breakdown) and leaves the device permanently and irreversibly damaged.

Tools and Techniques

FA tools and techniques developed and commonly used in the IC industry can be used to analyze failure mechanisms in MEMS. But the failure analyst will find that some tools and techniques will not be readily applicable to MEMS and require some adapting to properly assess the failure mechanism. This section will discuss some of the tools and techniques used to help diagnose MEMS failure mechanisms.

Scanning Electron Microscopy (SEM)

SEM is a very useful tool for diagnosing mechanical or electrical failure mechanisms of MEMS. The large depth of focus, high depth of field, wide range of magnifications, and minimal sample preparation make this a technique of choice for MEMS inspection and analysis. One drawback to using the SEM for MEMS failure analysis is the inability to see under multiple layers of mechanical structures. Having the

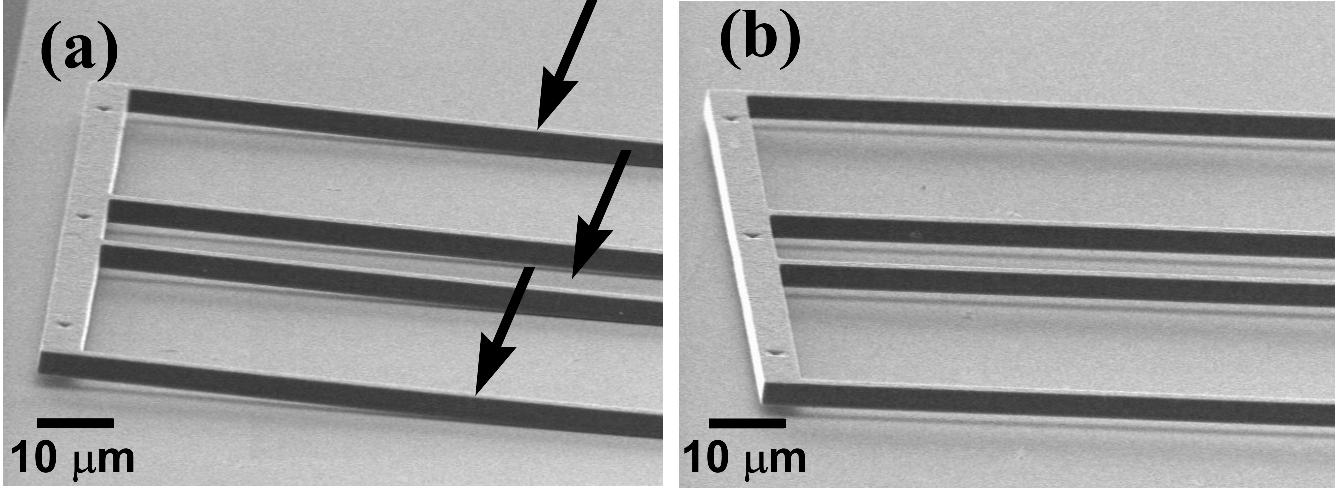


Fig. 21. SEM image of spring elements. **a)** Defective spring (arrows reveal stuck sites). **b)** Spring element suspended properly above the substrate.

ability to tilt and rotate samples to large angles improves the ability to diagnose failure mechanisms but requires more care to avoid contact with the pole piece or secondary electron detector during examination. SEMS equipped with electrical feed through capabilities allow actuation of a device during characterization. This can be particularly valuable for vacuum or low pressure dynamic analysis.

SEM analysis of failed MEMS has revealed stuck springs, fractured and broken elements, and abnormal displacement. Although these failure modes can be characterized with other analytical techniques, SEM analysis can identify and localize the failure site very quickly with little or no sample preparation. Figs. 21a and b reveal the location of springs adhered to the ground plane compared to functional (floating) springs. Fig. 22 identifies a stuck comb finger resulting in a functional failure. SEM can also be used for determining the electrical continuity of MEMS in static and operating conditions. Electrical continuity can be analyzed using Voltage Contrast (VC) or Resistive Contrast Imaging (RCI) [38]. To determine electrical continuity, proper electrical connections and equipment are needed to drive the device. Other than equipment needed to operate the devices, no special sample preparation is required.

For conventional SEM imaging, very little sample preparation is required. Methods to improve image quality (i.e. sample coating) cannot be performed on MEMS due to their motion-based mode of operation. Applying a conductive coating will bind up structures that will prevent motion or short the device. Wet etch delineation of MEMS such as optical switches, mirrors, gears, or hinges, with reflective or wear resistant coatings risks damaging the device. Thin film delineation using a wet process will damage these devices by fracturing components causing the broken components to float off during etch or rinse. Another potential failure mechanism using wet etch delineation is stiction (or adhesion) to the ground plane. Etch delineation should only be conducted on structural layers firmly adhered to the substrate. Test

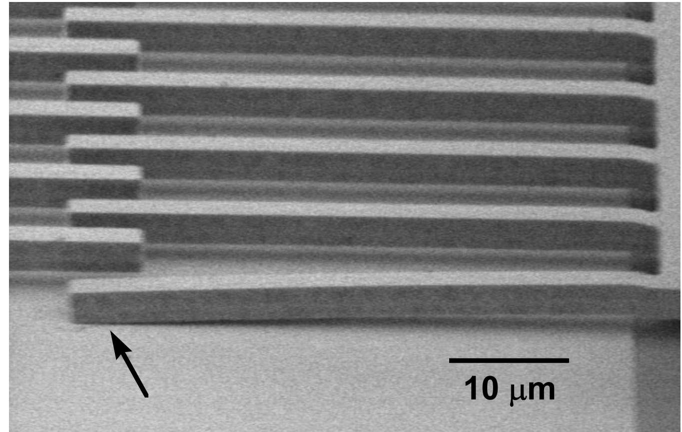


Fig. 22. SEM micrograph of a comb finger stuck to the substrate.

structures or solid, multi-layered structures anchored to the substrate should be used for thin film delineation.

Transmission Electron Microscopy (TEM)

TEM is a very valuable analytical tool capable of obtaining information on crystallographic orientation, materials interfaces, thin films, contact sites and more. Thin film applications for reflectance or wear resistance require knowledge of the interaction between the film and “bulk” material. TEM analysis will accurately reveal the film thickness, crystallographic behavior, and continuity along the interface. As shown in Fig. 23, a cross-section of the pin joint and hub portion of a microgear reveals the crystalline properties of the deposited polysilicon.

Sample preparation of released free-floating MEMS for plan-view or cross section is difficult to perform. The multiple levels of material that constitute the device make it difficult to thin the top or bottom of the device without fracturing or destroying the region of interest. Plan view TEM samples can be made using the replication method. This method uses a poly-acetate film to make a replicate of the

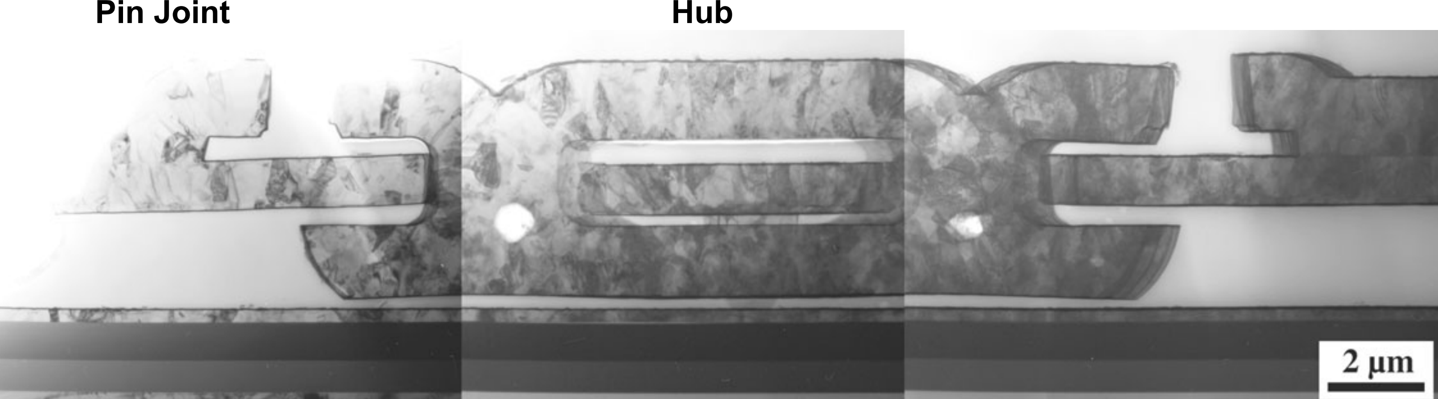


Fig. 23. TEM cross-section of a released tungsten coated micromachine.

surface of interest. A thin section of poly-acetate film softened with a drop of acetone is placed on the structure and then removed. This process usually removes the MEMS from the die or package obtaining the material of interest. The parts are then coated with a thin carbon film and transferred to a TEM specimen grid. The remaining polyacetate film is dissolved with acetone, allowing the device to rest on the carbon film supported by the TEM specimen grid. The sample is transported to the TEM in a TEM specimen loader and characterized. This technique is typically used for "grab and transfer" component analysis for devices that were tested or failed where plan view TEM analysis can provide insight into the failure mechanism. Removing the device from the die or package, however, should not compromise the failure mechanism.

The most difficult method of TEM sample preparation for MEMS involves producing cross sections on released movable devices. The difficulty lies in immobilizing the movable parts. This can be done by depositing a thin film on the movable components connecting them to rigid structures or by embedding the device in an adhesive. Use of a focused ion beam (FIB) will aid in the cross section sample preparation process. Further analysis of a wear resistant film deposited on the microgear of Fig. 23 is shown in Fig. 24. Here, we can examine the interaction between the film and the polysilicon. Information is obtained regarding the crystal structure of the film and at the interface.

Focused Ion Beam (FIB)

FIB systems are extremely valuable tools used to verify design layout and diagnose failure mechanisms in MEMS [39, 40, 41]. FIB systems use a focused beam of Ga^+ ions (typically 25 – 50 keV) for precise material removal (by physical sputtering). The FIB provides the best method for producing clean cross sections of the area of interest in MEMS structures. Cross sections can be made of both large and small structures with submicron accuracy.

In MEMS, cross sections of mechanical components can be made without disturbing other mechanical components. Fig. 25 shows a FIB cross section along a worn microgear revealing the hub and pin joint regions operated to failure

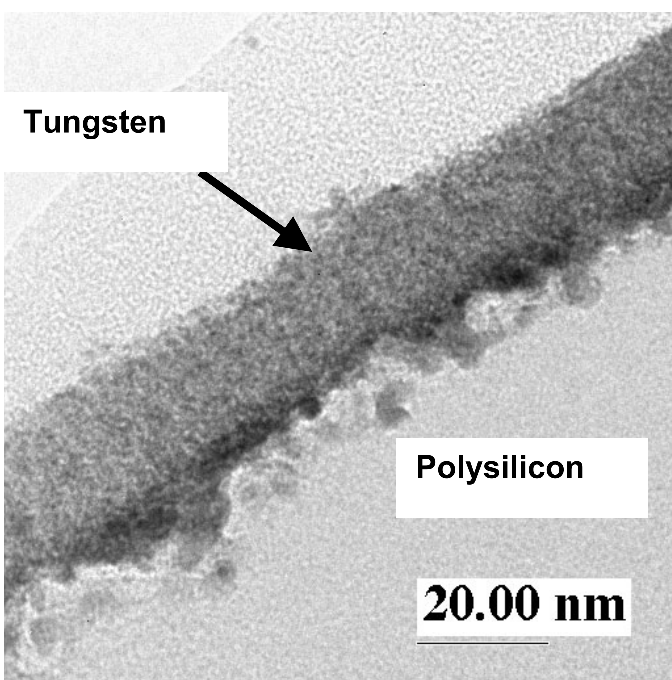


Fig. 24. Tungsten wear resistant film deposited on a polysilicon micromachine.

under accelerated conditions. Wear debris is observed along the top surface of the gear. Although this debris can be observed using an optical microscope or SEM/FIB analysis reveals the effects of wear on underlying rubbing surfaces. This analysis could not be achieved using conventional optical and/or SEM. Conventional cross-sectional analysis (potting, polishing, etc.) of released MEMS is extremely difficult to achieve without compromising the failure mechanism or completely destroying the MEMS in the process.

Another advantage to using the FIB for MEMS failure analysis is the preferential cutting used to free up stuck (adhered) portions of a device. As illustrated in Fig. 26, a FIB cut is made along the guide region of an electrostatic actuator adhered to a moveable shuttle used to drive a microgear. Wear material is observed along the moveable shuttle in the region below the guide of Fig. 18a. After cutting through the guide, the portion adhered to the shuttle and the shuttle itself move back to its as-fabricated position. This indicates the

failure mechanism is adhesion of the guide to the moveable shuttle and the failure site is the guide and upper part of the shuttle. The microengine was successfully operated after the incision was made.

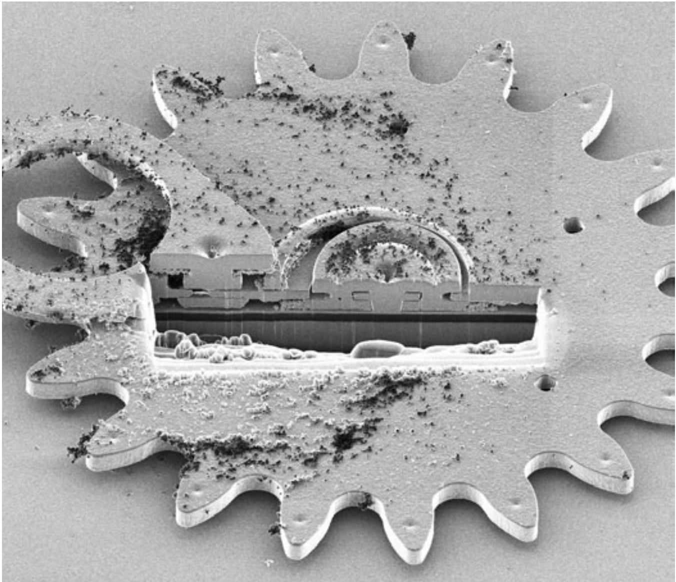


Fig. 25. A FIB cross-sectioned microgear after accelerated lifetime testing. The exposed cross-section reveals the hub and pin joint regions.

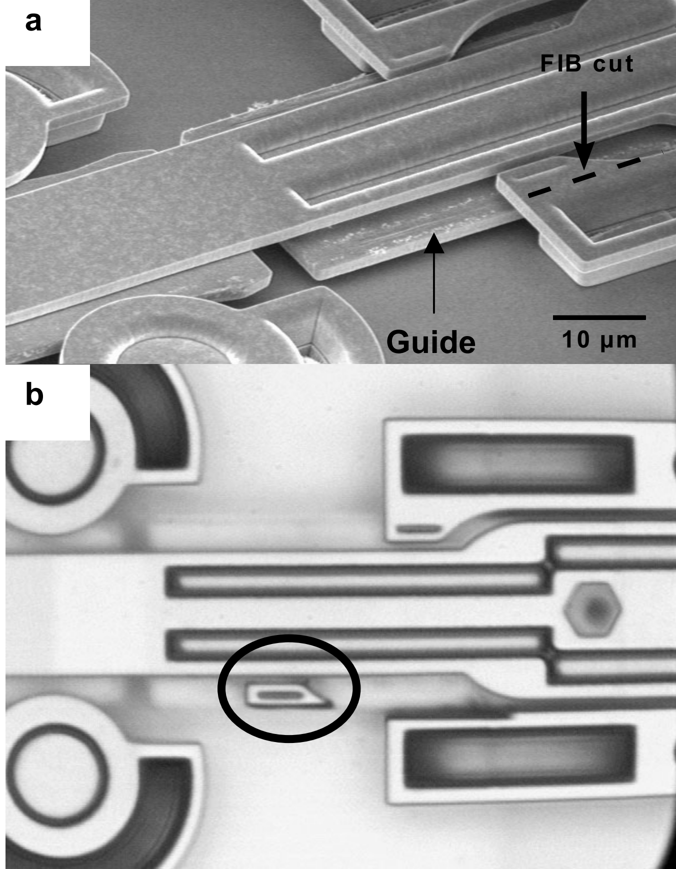
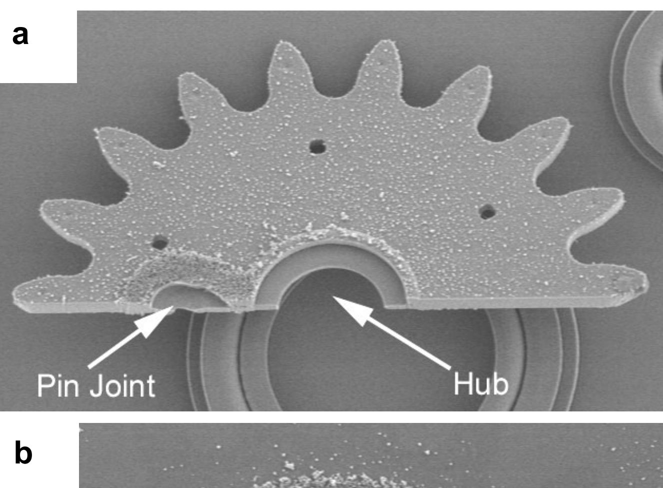


Fig. 26. a) Location of FIB cut on the pinned link microengine guide of the up-down shuttle in a failed binary counter. **b)** Optical image showing the FIB cut freed the shuttle from the guide.

The FIB has been used to free up and in some cases, excise specific components of the device to enable the analysis of otherwise inaccessible areas [39, 40, 41]. These components are then free to rotate, flip, or manipulate to access exposed surfaces of the device that were not readily accessible. As shown in Fig. 27, a microgear was cut in half. One half of the microgear was excised and turned over. Analysis of the backside revealed accumulation along regions where rubbing surfaces are prevalent. Analysis of the substrate directly underneath the gear reveals tracks of debris being “swept” to certain areas during operation. These debris tracks would not be observed without the removal of the gear.

Other capabilities of the FIB have proven useful for MEMS failure analysis. Capabilities such as the deposition of conductive and insulating materials from ion beam-assisted chemical vapor deposition have proven valuable for device repair/modification. One example where these features have been used was in a failed RF switch [42]. The ends of the cantilever beams supporting the switch were open so the switch could not contact the pad below it. As shown in Figs. 28a and b, the non-functional cantilever beam has a separation leading to an open between it and the electrode (anchor). Deposition of tungsten (using an organometallic gas) bridged the electrode to the cantilever beam allowing the beam to flex and function.



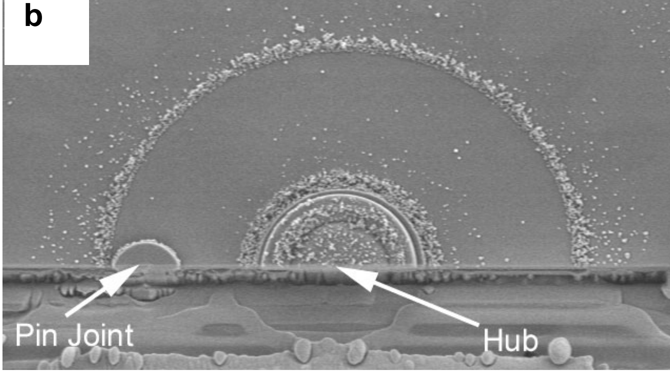


Fig. 27. a) Structural analysis of the backside of a worn gear. **b)** Analysis of poly 0 directly underneath the gear. These areas were exposed by FIB cutting the gear in half and removing one half.

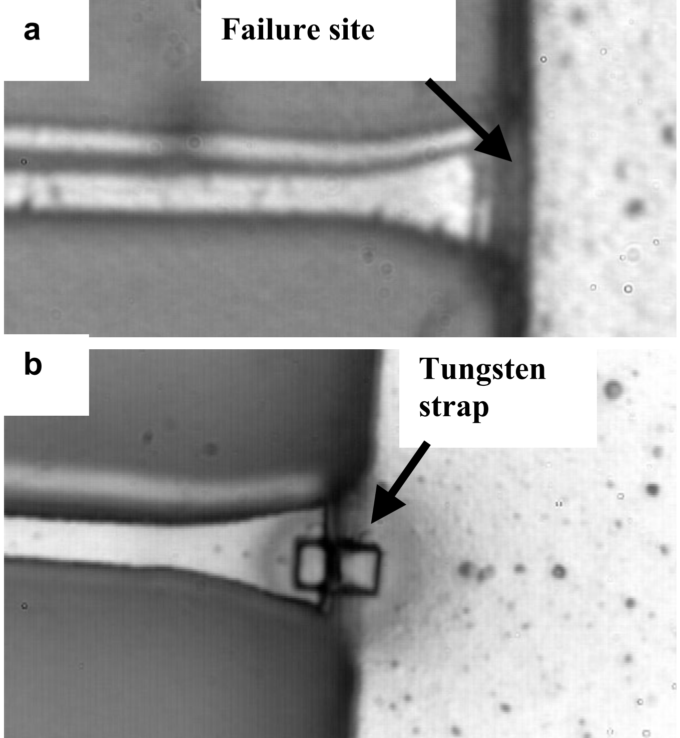


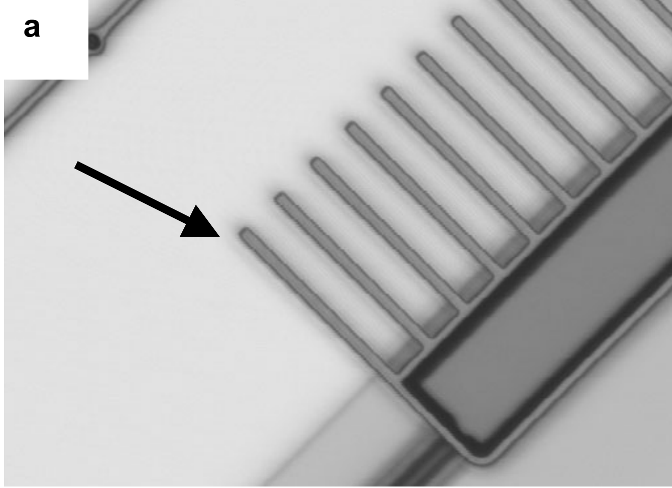
Fig. 28. a) Optical image of the failed beam of an RF switch. **b)** Optical image of the same beam with tungsten deposited over the failure site. This device functioned after the tungsten deposition.

Thermally-Induced Voltage Alteration (TIVA)

The application of TIVA to MEMS [43] presents some special opportunities for MEMS failure analysis. MEMS have been shown to fail electrically due to stiction, particle contamination, or EOS/ESD [35]. Most MEMS structures are thermally isolated from the substrate. These structures can be easily heated with much lower laser power compared to the laser power used when TIVA is applied to ICs [3]. In several cases, the thermal isolation does not require any external stimulus to the MEMS to identify the short.

Thermally isolated structures get hotter than devices with interconnects. in the unreleased state (still encased in sacrificial material) as well as after release and testing. Some MEMS require care to avoid overheating and possible damage to the MEMS structure. Figs. 29a and b reveal a comb finger in contact to the ground plane on an electrically failed electrostatic actuator. This actuator was electrically tested and diagnosed as electrically shorted. In one device, there are over 400 comb fingers. Any comb finger could cause an electrical short in this device. Examination of the comb fingers using an SEM alone to diagnose the failure mechanism proved tedious and time consuming.

An added benefit of TIVA analysis for MEMS is that most MEMS do not have active IC's in the immediate region i.e., no Si diffusions. With these particular MEMS structures there is no concern about photocurrent (electron-hole pair



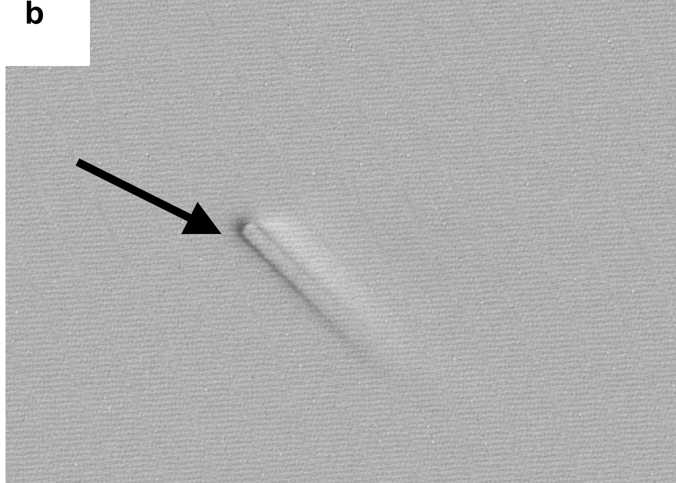


Fig. 29. a) Reflected light and **b)** TIVA image using a 5 mW, 540 nm laser. A bottom level comb finger from a stack of three shown shorting to the ground plane has been identified.

recombination) effects swamping out the TIVA signals. Therefore, shorter-wavelength lasers can be used for TIVA analysis and offer improved spatial resolution. The improved spatial resolution can also be seen in the reflected light images used for registration.

When ICs are present in the area of the device (IMEMS), conventional TIVA analysis can be used to diagnose the failure mechanism in both of the components at the same time. This technique has been very valuable in isolating the failure site of electrically shorted MEMS quickly, accurately, and cost-effectively. This technique has also been shown to work on several types of devices ranging from RF MEMS, optical micromirrors, microfluidic pumps, and electrostatic actuators. As shown in Fig. 30, TIVA is performed on a microfluidic pump encased in sacrificial silicon dioxide. Further analysis of the TIVA site of Fig. 30b reveals a polysilicon stringer at the poly 0 level deposited directly over the silicon nitride (Fig. 30c).

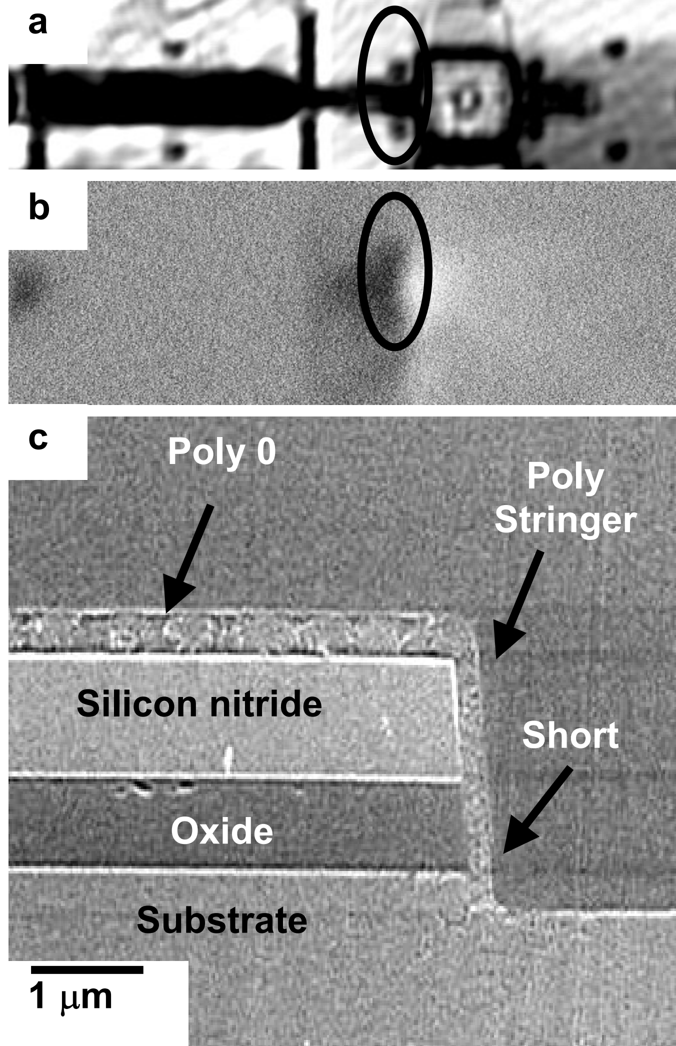


Fig. 30. Unreleased drop ejector that failed by electrical shorting. **a)** reflected light image showing the unreleased ejector, **b)** TIVA image revealing the location of the short, **c)** polished cross-section identifying the failure mechanism as a poly 0 stringer (arrow) connected to substrate.

IR Confocal Laser Microscopy

Processing of MEMS often contains a release step where sacrificial material has to be removed to free up the device. This process is not always straightforward to control. As a result, the release is not complete or residues of the sacrificial layer remain present. Both effects can lead directly or indirectly to failures.

IR confocal laser microscopy is a well-established technique to non-destructively inspect semiconductor devices through the backside. It can also be used to study and optimize release and packaging in MEMS devices through the front side. Fig. 31a shows a laser-scanning image (100x magnification) of a defective arm of a MEMS component [44] taken through the upper Si layer. The bright area indicates a region where the release was not complete. Very good contrast can be obtained by using the confocal component of the system and focusing on the correct interface. The failure was verified by removing the top layer with a sticking tape and inspecting the region of interest. The SEM picture (Fig. 31b) confirms that there was indeed a micrometer-sized hillock of sacrificial material present under the MEMS layer.

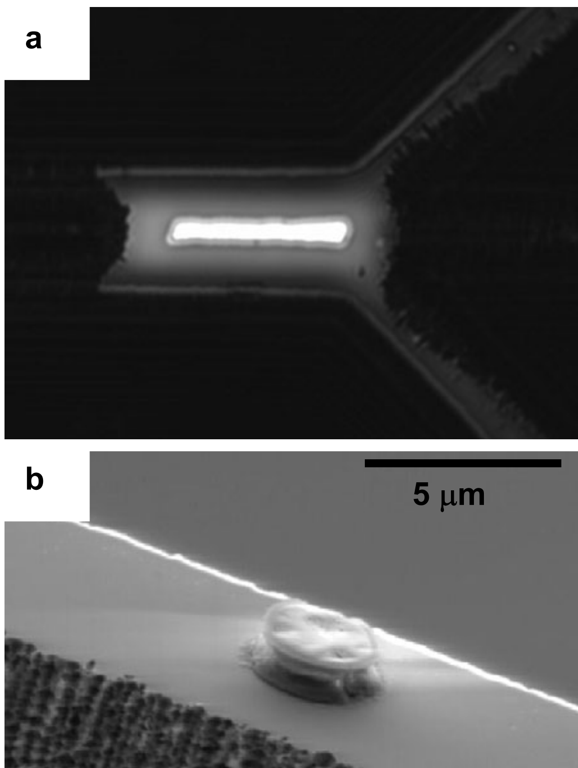


Fig. 31. a) Laser scanning image obtained at 100x through the top Si layer (< 100 μm) of a MEMS. The bright image indicate remaining SiO_2, **b)** SEM image obtained at the same spot after removal of the top Si film, showing the micro-size SiO_2 hillock [44].

If metal layers are present at the front side, the technique can also be applied from the backside of a polished Si substrate (Fig. 32). The advantages of backside analysis are straightforward. Backside analysis of a packaged device does not require delidding. The delidding process can introduce particles, foreign material, and/or a shock to the MEMS environment. This can result in failure mechanisms attributed to sample preparation that can impede diagnoses of field or test failures.

3-D Motion Analysis

Observation of the 3D-motion of MEMS, the measurement of their resonance frequencies and the observation of possible cross-talk is certainly of interest for reliability testing and failure analysis. A variety of systems exist that can be used to monitor the motion of MEMS. These include: high-speed photomicrography and optical 3-D imaging methods based on laser interferometry, Mirau

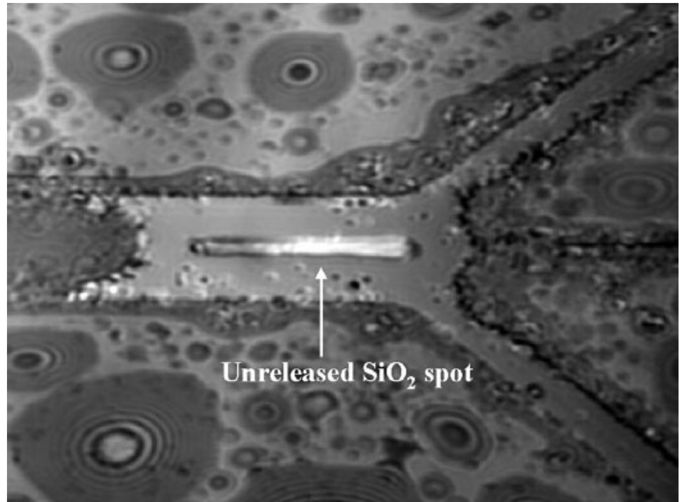


Fig. 32. Laser scanning image obtained at 100x magnification through the back side of the 500µm thick silicon substrate [44].

interferometry, speckle, micro-Moire, scanning laser Doppler velocimetry etc. A comprehensive overview, showing results on ink-jet printheads, microturbine, microrelays and micropumps is given in [45].

The addition of such motion analysis to electrical testing of MEMS is often indispensable to find the failure mechanism and mode. For example, optical inspection of the trajectory of an ink drop ejected by electrostatic drop ejectors (inkjet print heads) immediately provides information on a failed or malfunctioning ejector [45, 46]. Freeman et al. developed a computer microvision system which allowed them to investigate effects of stimulus and environmental conditions, as well as manufacturing process conditions, on the performance, aging, and ultimately, failure of MEMS [47]. A system allowing the monitoring of MEMS vibrations up to 15 MHz was presented in [48]. Fig. 33 shows sequential images of the motion of a microrelay, indicating clearly the problem areas [48]. (The images will be in black and white, so you might want to point out what's happening.)

Such systems can also be used to investigate stiction in various MEMS technologies. Fig. 34 shows electrical measurements (Fig. 34a) of the capacitance change (ΔC) induced by a switching RF-MEMS capacitive switch (depicted in Fig. 34b, photo in Fig. 34c). These data show an initial increase, followed by a sudden decrease, which is followed by a slower decrease. Without optical investigation of the motion of such switches, it is rather hard to predict the cause of these capacitance changes. Fig. 26d shows an image taken after the switch showed a decreased ΔC but was still functioning. Here one can see a bright area where the switch is still moving and a dark area where it is not moving anymore, i.e. is stuck. These pictures (movies) show that the switch at first is not completely pulled in for the used actuation voltage, the pull in increases due to charging of the dielectric, i.e. ΔC increases. Next, stiction, due to dielectric charging, occurs at the end part of the switch. This leads to a decrease in ΔC. The remaining part of the switch is still moving. Because further charging of the oxide takes place, the portion of the switch that sticks increases until the complete switch is stuck and does not move anymore [49].

Raman Spectroscopy

Raman spectroscopy is an optical technique that probes lattice vibrations (phonons in solids or molecular vibrations in gasses or fluids). Because the vibration frequencies are sensitive to internal and external perturbation, the technique can be used to study the composition, phase, crystallinity, crystal orientation, and in some cases, doping of materials [50]. Although Raman spectroscopy is not considered a failure analysis technique, it has some very useful applications for microsystem devices [51].

In a Raman spectroscope [50-53], the light of a laser is focused on the sample. The scattered light is collected and its spectrum is analyzed by a grating spectrometer and a CCD detector. If the sample is fixed on a XY-stage under the microscope, one or two-dimensional scans can be made by moving the sample and collecting a Raman spectrum at certain intervals. An autofocus system added to the microscope

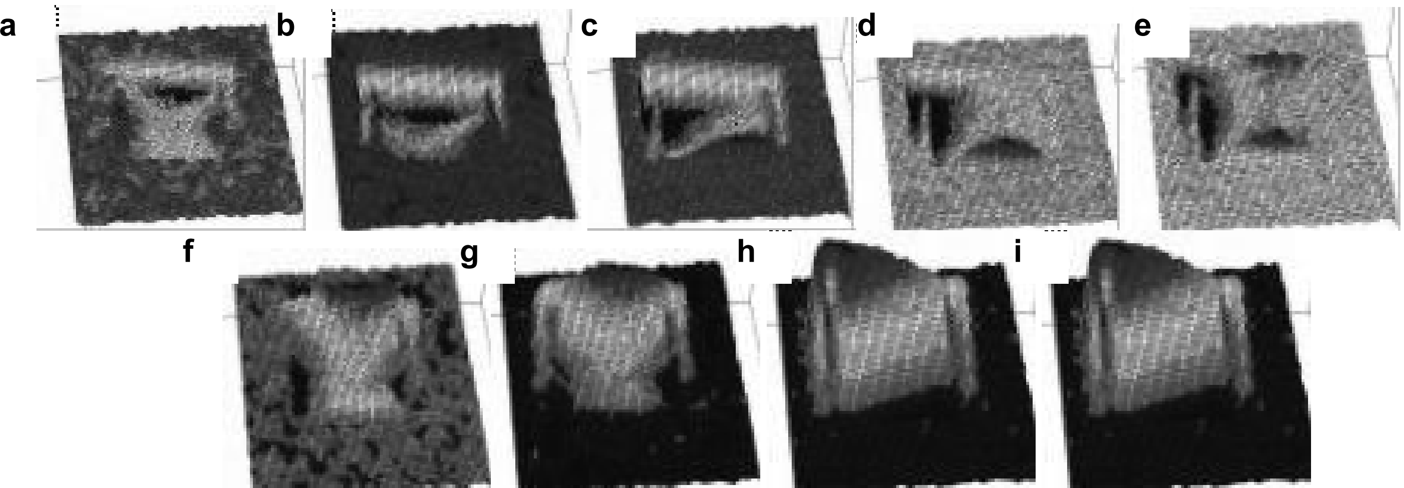


Fig. 33a-i. Sequential pictures taken from a moving defective micromachined relay vibrating at 1000 Hz [48].

Fig. 34. Combined electrical and optical lifetime testing of an RF-capacitive switch. **a)** The change in capacitance ΔC (arbitrary unites) as a function of the number of switching cycles. **b)** Scheme showing the layout of the switch. **c)** Photograph of the switch taken with the CCD of the optical system. The fixed part of the switch is at the bottom. **d)** Picture showing stiction at the top side (right side in b, top side in c) of the switch. The switch is still moving in the white area [49].

allows mapping of larger areas. Micro-Raman spectroscopy, where the laser light is focused through a microscope, allows investigation with µm spatial resolution.

Fig. 35 shows a typical Raman spectrum of crystalline silicon. The frequency of this peak is influenced for example by mechanical stress and temperature in the sample. The FWHM (Full-Width-at-Half-Maximum) of the Raman peak can provide information on the crystal quality. When there is damage in the silicon or there are non-uniform structures or conditions in the sample, the Si peak may become asymmetrical and the FWHM increases [50].

Failures in MEMS are often related to mechanical stress in the structure. One interesting characteristic of Raman spectroscopy is its sensitivity to strain. When the material is under strain it may cause a shift in the frequency of the Raman signal. To calculate the relation between Raman shift and strain, one has to solve the "secular equation" [51, 52], which is in general a rather complex job, often impossible. However, for silicon the following rule generally holds: compressive

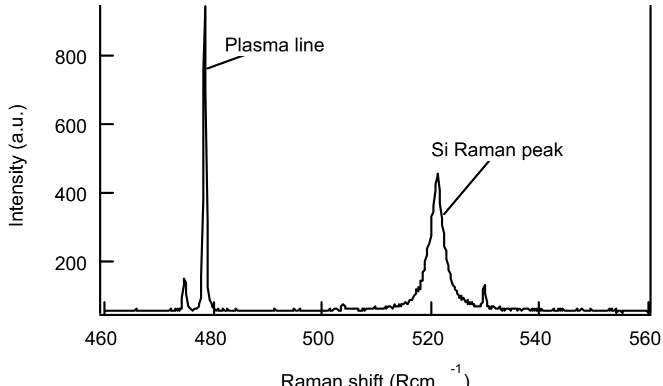


Fig. 35. Raman spectrum of crystalline silicon.

stress results in a positive shift of the Raman frequency ($\Delta\omega > 0$) and tensile stress in a negative shift ($\Delta\omega < 0$). For uniaxial or biaxial stress, one obtains a linear relation between stress and Raman shift.

Fig. 36 shows an example where micro-Raman spectroscopy is used to measure the stress in the membrane of

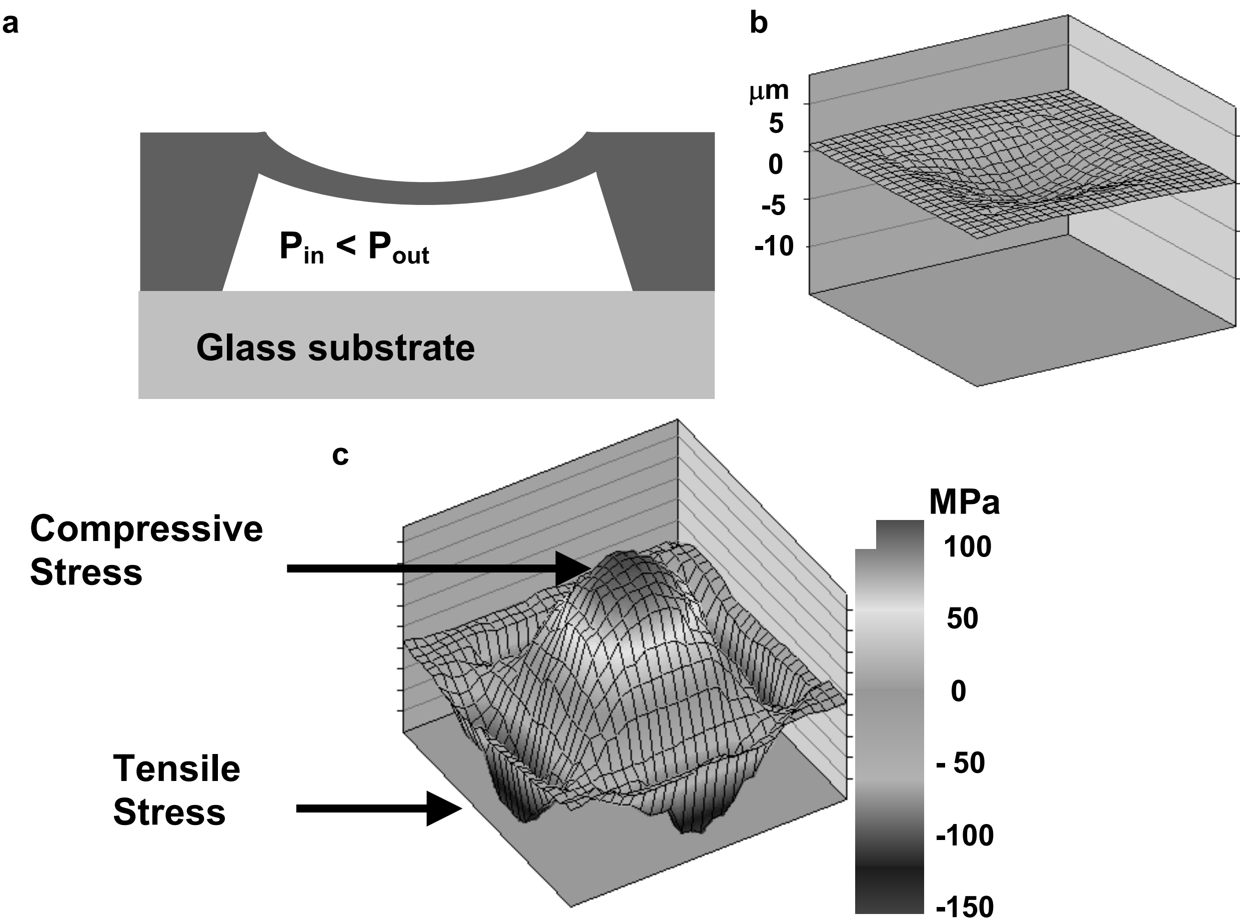


Fig. 36 a) Schematic of a Si membrane of a pressure sensor bonded to a glass substrate. The membrane is in under-pressure, **b)** deflection of the membrane measured using laser interference, **c)** biaxial mechanical stress in the membrane measured using micro-Raman spectroscopy.

a pressure sensor [53]. This sensor was processed using etched cavities in the silicon wafer and bonding the wafer to a glass substrate using an anodic bonding process (Fig. 36a). This bonding process introduced a negative pressure inside the chamber. The membrane was deflected inwards about 5 μm (Fig. 36b). A micro-Raman spectroscopy system equipped with an autofocus module was used to scan the surface to measure the Si Raman peak. Fig. 36c shows the mechanical stress in the membrane, calculated from the change of the Raman frequency ($\Delta\omega$) from its stress-free value. In this picture, a positive value means compressive stress. A compressive stress state is observed at the center of the membrane. A negative value observed near the sides of the membrane indicates a region of tensile stress. Similar experiments were performed on the backside of the membrane. The Raman spectroscopy technique allows probing of the silicon through the glass substrate. This makes it possible to study, the anodic bonding process by analyzing the glass/silicon interface.

Stiction and friction are of great concern for MEMS with impact and rubbing surfaces such as switches and microgears. The application of anti-stiction coatings can often reduce these problems. To obtain good performance, the coatings should be uniformly applied on all the exposed surfaces. This is particularly challenging when coating high-aspect-ratio areas with small gap spacing such as meshing gear teeth. In some cases, micro-Raman spectroscopy can be used to study the coating [54] as well as its uniformity across the device. Ager et al. [55] used the intensity of the Raman signal of an anti-friction carbon (DLC) coating to study its thickness variations on Ni-alloy gears (Fig. 37). It was shown that the coating was uniform on the top gear surface, but somewhat thinner at a certain depth inside the gear teeth.

Photoemission Microscopy (PEM)

PEM is very well known for failure analysis of IC's at the wafer or chip level. It can detect light emitted by diodes, MOSFETs in saturation, Fowler-Nordheim tunneling through gate oxides, gate oxide breakdown, latch-up, etc. There are three basic mechanisms for light emission: light emission from electrical field accelerated carriers, light emission from recombination of electron-hole pairs and black body radiation. Although the technique has limited applications for MEMS, it is certainly of use in some specific cases.

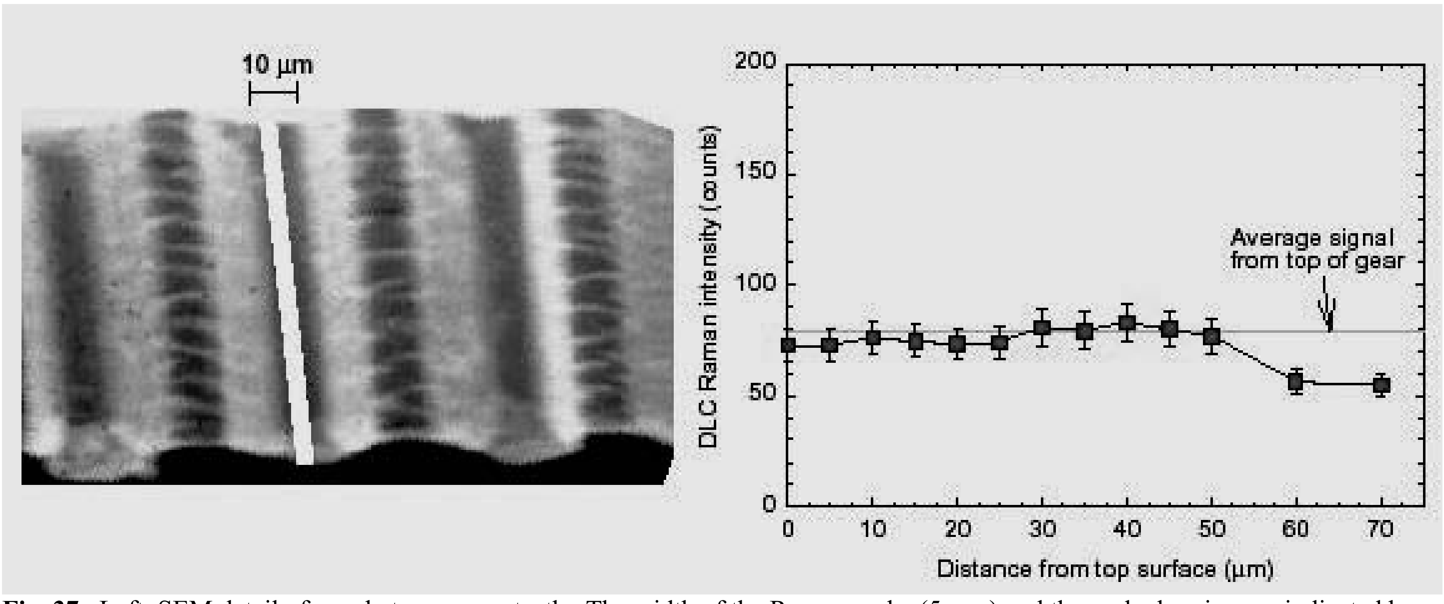


Fig. 37. Left: SEM detail of area between gear teeth. The width of the Raman probe (5 mm) and the probed region are indicated by the white line. Right: DLC Raman intensity between gear teeth as a function of the distance from the top surface of the microgear. The intensity observed from scans on the top of the gear is indicated by the solid straignt line (from [52]).

A first application domain is the detection of black-body radiation. PEM instruments, especially the older generation, are not really made for heat detection. The sensitivity of their detector is in general limited between about 500 nm and 1000 nm. In newer instruments this is extended towards longer wavelengths. But MEMS often use heating elements in which the temperature can locally and temporarily reach levels higher than 300 ^{o}C. It is possible to detect this heat using a PEM instrument, even with a detector from an older generation. An example shown in Fig. 38 reveals the heat distribution in a 6/60 Ti/TiN line (300 µm long, 1 µm wide), which was pulsed with 5.5 mA current [56, 57]. The heat distribution was uniform along the length of the line. Using spectroscopic PEM, the temperature of the line was measured as 380° C.

Two other applications are shown in Figs. 39(a) and (b), where failure analysis using PEM showed the position of a short in an accelerometer (Fig. 39a) and in a moving resonator (Fig. 39b) [58].

0-Level Packaging Hermeticity and Scanning Acoustic Microscopy (SAM)

MEMS require in general different packaging solutions than the ones commonly used for ICs. While in both cases the packages have to withstand standard tests such as thermal cycling, thermal storage and high humidity/high temperature storage; many MEMS packages need to fulfill additional specifications. The motion of a microsystem is affected by the environment in which it functions, and might be influenced by particles, humidity, pressure and gasses. IC's are protected against debris caused by the dicing process through a passivation layer that is deposited on top of the wafer at the end of the processing. This cannot be done for most MEMS

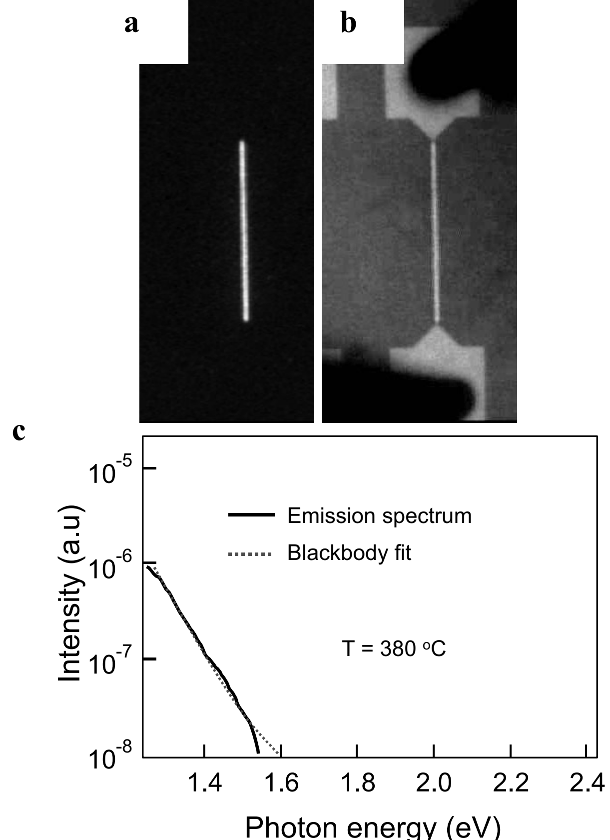


Fig. 38. Ti/TiN line with 26 V bias. **a)** Light due to blackbody radiation as detected with PEM, **b)** Overlay of emitted light and image of the device, **c)** spectrum of the light from **a**, measured with SPEM. Dotted line: result of fit of the blackbody equation to the spectrum (from [52]).

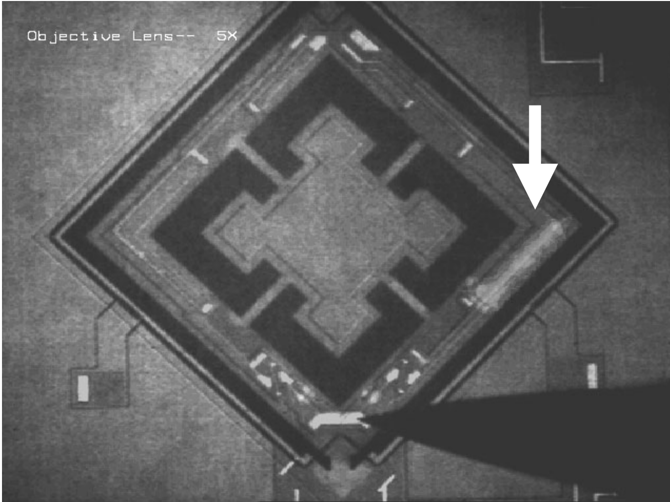


Fig. 39a. Emission due to field accelerated carriers observed in an accelerometer, indicating a short between different layers [54].

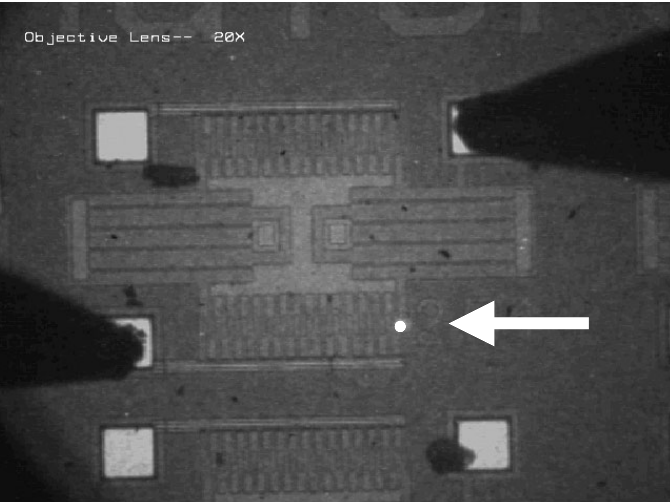


Fig. 39b. Emission (bottom right) seen as a flashing light during movement of a resonator, indicating the position where two comb structures touch [55].

containing moving parts. As a result, they often require packaging before dicing, i.e. 0-level packaging on wafer level by either wafer to wafer bonding or locally bonding of a little cap (ex. Si or glass) over the MEMS using a hermetic sealing ring [56, 57].

As already mentioned, high humidity levels can result in capillary stiction of the moving part of the device. Humidity can also cause indirect failure. It can for example change the sensitivity for charge trapping of insulator material used in capacitive RF-MEMS. This again might cause stiction of the moving part, this time due to charging of the insulator. So, it is often required that the package forms a hermetic seal against humidity. Hermeticity testing of cavities (fine leak and gross leak) is in general done using standard methods (i.e. MIL-STD-883D). However, because the cavities of such MEMS packages deal with volumes that are much smaller (at least 1000 x) than the ones described in the MIL standard, the standard test techniques are not applicable [58 - 60]. The upper detection limit of the fine leak tests was shown to be well below the lower limit of the gross leak test, leaving a gap in the detection region of these two techniques. Leaky cavities can in this way still escape from detection with these standard techniques. This detection gap can be bridged through the use of microbolometers or MEMS resonators as sensors inside the package [61]. An alternative way (Fig. 40) was presented in [62]. In this case the test is destructive: a hole has to be made through the substrate to check the hermeticity of the package. It is however a very good way for optimizing the packaging process.

For some packages, scanning acoustic microscopy (SAM) can be used to pin-point the leaky part. SAM is in general not applied a lot for MEMS or MEMS packaging failure analysis. The reason is simple: SAM detects the presence of air gaps, voids or delamination in IC packages. But because a MEMS device is mostly surrounded by air, SAM cannot be used. It is however a good technique to find problems in 0-level packages where a BCB [63] or solder sealing ring is used to bond a cap such as Si over the MEMS [63]. An example of a 0-level package (PbSn solder as sealing ring and Si as cap) that failed for a gross-leak test is shown Fig. 41. The figure shows a SAM picture of one corner of the cap. There are clearly white spots in the solder showing that it is not well sealing the Si cap, which explains why this sample did not pass the leak test.

Atomic Force Microscopy (AFM)

AFM provides very detailed topographic information about the surface roughness and surface topography of a sample. Surface roughness is an especially important aspect in optical MEMS and MEMS with impacting or rubbing surfaces (switches and gear transmissions). As surface roughness increases, the amount of reflected signal decreases

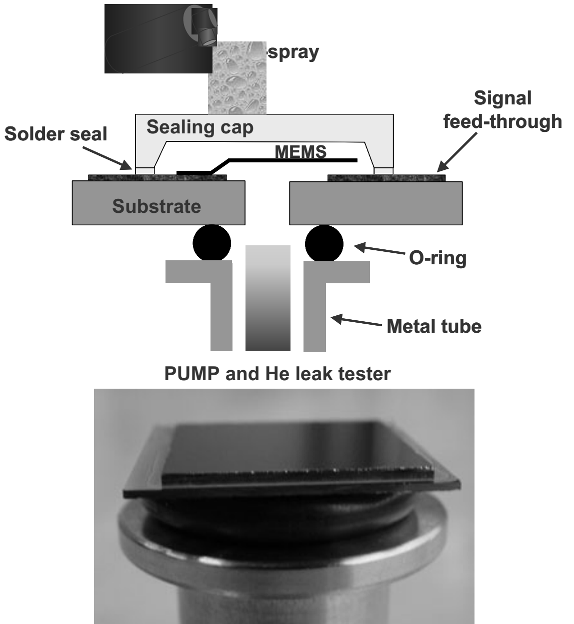


Fig. 40. Procedure to test the hermeticity of 0-level packages for RF-MEMS [56].

(by reflective loss and absorption. Surfaces that have been subjected to rubbing or impact will be revealed with a surface scan analysis. AFM analysis of a worn surface is shown in Fig. 42. This image shows the wear debris produced in a microgear of a failed binary counter. Fig. 43 shows the topography of a larger surface subjected to wear. Region (1) has a higher surface roughness and is an accumulation site for worn material. Region (2) is much smoother indicating it was the region subjected to wear. Region (3) is the typical roughness of a polysilicon device. A line scan shown in Fig. 44 revealed the topography across the region of Fig. 43 (dashed line).

In some instances, portions of the device must be removed to examine the area of interest. The FIB can be used to excise given components to enable AFM analysis. This is of particular interest if you wish to look at bottom level structures or surfaces of a 3 or 4 level system. This aspect is also important regarding the limitations of z-axis piezoelectric movement in the AFM. For large piezoelectric heads, typical z-axis displacement ranges from 8-10 µm. This range prevents imaging MEMS with a total z-height of 10 µm or more. Aside from modifications to characterize the lowest levels in MEMS, no special sample preparation or surface treatment is required for AFM analysis.

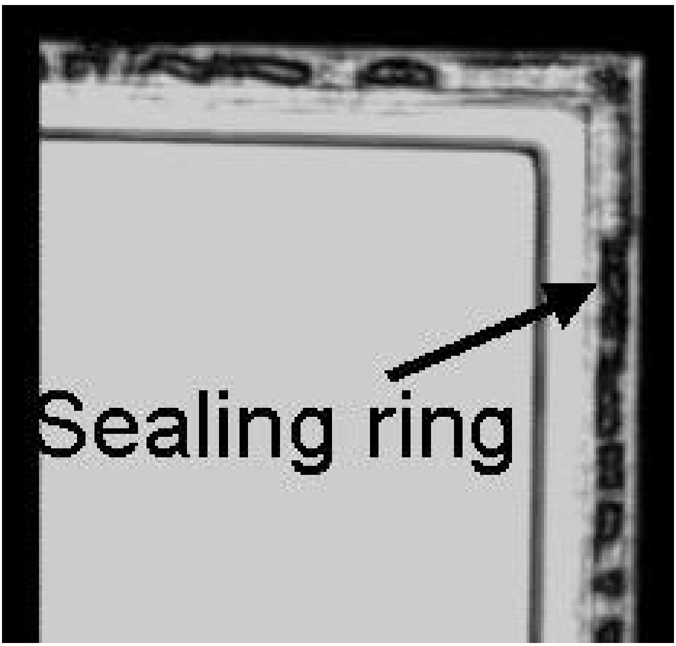


Fig. 41. SAM image of one corner of a 0-level package: a Si cap bonded to Si substrate with PbSn solder as sealing ring. The picture shows incomplete sealing by the solder [64]

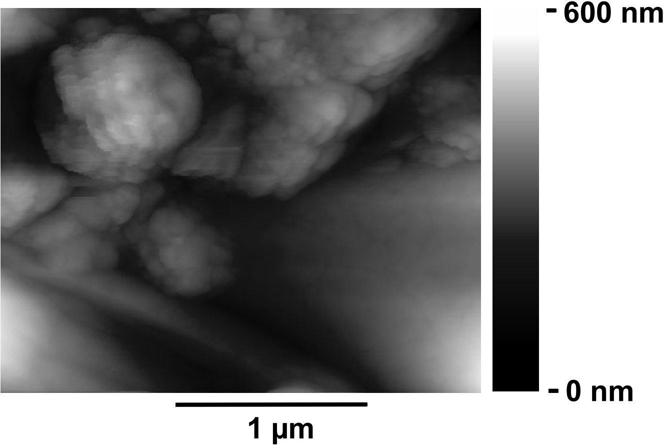


Fig. 42. AFM topographic image of wear debris produced along the rubbing region of a failed binary counter.

Conclusions

Although several failure analysis tools and techniques from the IC industry have been adapted and utilized for MEMS failure analysis, several more are needed to identify application specific failure mechanisms. Non-destructive failure site identification from adhered or cold-welded devices

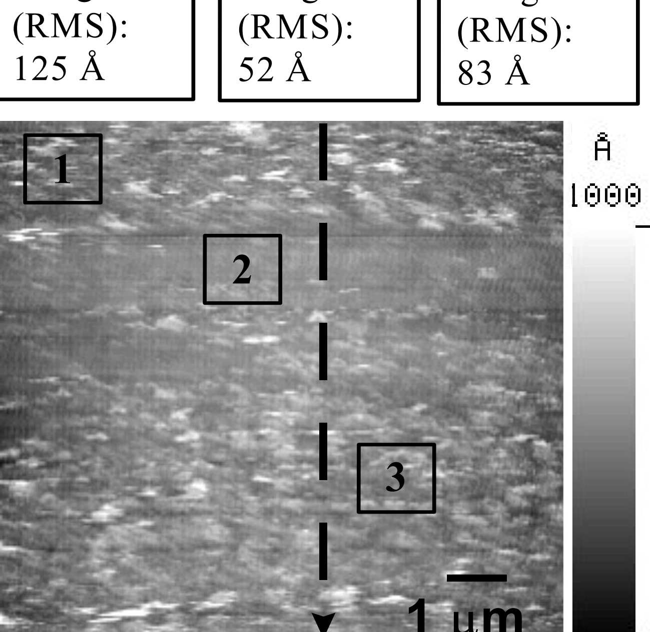


Fig. 43. AFM topography image of a worn surface revealed a wear track.

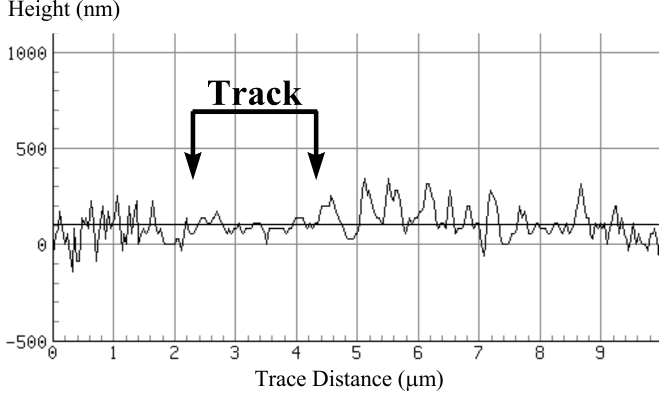


Fig. 44. Surface trace of the dashed line shown in Fig. 43. Note the smooth region indicating the location of rubbing along the surface.

will be needed as these devices become more commercially available. Current techniques are destructive and compromise the failure mechanism while rendering the rest of the device inoperable. Multi-functional tools able to operate several samples in parallel will also be required.

Acknowledgements

Several collaborators from IMEC Belgium, Analog devices, and Sandia National Laboratories contributed to the technical content of this chapter. The authors would like to thank Becky Holdford of Texas Instruments for reviewing this manuscript.

The authors would like to thank Merlijn van Spengen and Mahmoud Rasras of IMEC, Stuart Brown of Exponent, the microsystems engineering and science applications staff at Sandia for their processing efforts, Alex Pimentel, Teresa Gutierrez, Bonnie Tom Headley, Paul Kotula, and Michael Rye for their analytical support. Sandia is a multiprogram laboratory operated by the Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL95000.

References

[1] S. F. Brown, “Big Jobs are going to Micro-Machines”, *Fortune Magazine*, May 10, 1999, pp. 128[B]-128[F].

[2] R. M. Boysel, T. G. McDonald, G. A. Magel, G. C. Smith, and J. L. Leonard, “Integration of Deformable Mirror Devices with Optical Fibers and Waveguides”, Proc. of *SPIE*, **Vol. 1793**, 1992, pp. 34-39.

[3] W. Kuehnel and S. Sherman, “A surface micromachined silicon accelerometer with on-chip detection circuitry*”, Sensors and Actuators A*, **Vol. 45, No. 1**, 1994, pp. 7-16.

[4] N. Unal and R. Weschsung, “Inkjet Printheads: An example of MST market reality”, *Micromachine Devices*, **Vol. 3, No. 1**, January 1998, pp. 1-6.

[5] S. Arney, V. A. Aksyuk, D. J. Bishop, C. A. Bolle, R. E. Frahm, A. Gasparyan, C. R. Giles, S Goyal, F. Pardo, H. R. Shea, M. T. Lin, and C. D. White, “Design for Reliability of MEMS/MOEMS for Lightwave Telecommunications,” Proc. of ISTFA, 2001, pp. 345-348.

[6] P.J. French, P. M. Sarro, “Surface versus bulk micromachining: the contest for suitable applications,” J. Micromech. Microeng. (1998) pp. 45-53.

[7] B. Kloek, S. D. Collins, N. F. de Rooij, and R. L. Smith, “Study of electrochemical etch-stop for high precision thickness control of silicon membranes,” IEEE Trans. Electron Devices ED-36, (1989), pp. 663-669.

[8] I. W. Rangelow, P. Hudek, and F. Shi, “Bulk micromachining of Si by lithography and reactive ion etching (LIRIE),” Vacuum, vol. 46, number 12, (1995), pp. 1361-1369.

[9] R. Bosch, Patent No. 5501893: *Method of Anisotropically Etching Silicon*, 1996.

[10] L. Kenoyer, R. Oxford, A. Moll, “Optimization of bosch etch process for through wafer interconnects,” IEEE, (2003), pp. 338-339.

[11] S. D. Senturia in Ref 4, p. 11.

[12] J. J. Yao and M. F. Chang, “A surface micromachined miniature switch for telecommunication applications with signal frequencies from DC up to 4GHz,” Transducers, June 1995, pp. 384-387.

[13] J. P. Sullivan, T. A. Friedmann, and K. Hjort, “Diamond and amorphous carbon MEMS,” MRS Bulletin, April 2001, pp. 309-311.

[14] T. A. Friedmann, J. P. Sullivan, J. A. Knapp, D. R. Tallant, D. M. Follsteadt, D. L. Medlin, and P. B. Mirkarimi, Appl. Phys. Lett. 71 (1997) P. 3820.

[15] M. Mehregany, C. A. Zorman, S. Roy, A. J. Fleischman, C. H. Wu, N. Rajan, “Silicon carbode for MEMS,” Inter. Matls. Review, Vol. 45, No. 3, pp. 85-108.

[16] C. R. Stoldt, C. Carraro, W. R. Ashurst, D. Gao, R. T. Howe, R. Maboudian, “A low temperature CVD process for SiC MEMS,” Sensors and Actuators A 97-98, (2002) pp. 410-415.

[17] M. Mehregany, C. A. Zorman, “Micromachining techniques for advanced SiC MEMS,” Mat. Res. Soc. Symp., Proc. Vol. 640, (2001) MRS, pp. H4.3.1-H4.3.6.

[18] http://mems.cwru.edu/SiC/Pages/devices.html.

[19] J. Hruby, “LIGA Technologies and Applciations”, *MRS Bulletin*, April 2001, p. 1-4.

[20] H. Lorenz, M. Despont, N. Fahrni, N. LaBianca, P. Renaud, and P. Vettiger, J. Micromech. Microeng., 7 (1997) pp. 121 – 128.

[21] A. A. Avon, R. Braff, C. C. Lin, H. H. Sawin, and M. A. Schmidt, J. Electrochem. Soc. 146 (1) (1999), pp. 339-343.

[22] M. C. Gower, “Excimer laser microfabrication and micromachining ,” Optics Express 7 (2) 2000), pp. 56-60.

[23] http://www.dbanks.demon.co.uk/ueng/liga.html

[24] G. K. Fedder, “MEMS Fabrication,” ITC International Test Conference, paper No. 27.3, (2003), pp. 291 – 698.

[25] K. A. Shaw, Z. L. Zhang, N. C. MacDonald, “SCREAM 1: a single mask, single-crystal silicon, reactive ion etching process for MEMS,” Sensors and Actuators A v. A40, no. 1, Jan. 1994, pp. 63-70.

[26] V. K. Varadan, J. Xie, “Microstereo-lithography for polymer based MEMS,” Smart Structures and Materials, Proc. of SPIE Vol. 5055 (2003), pp. 167 – 176.

[27] V. K. Varadan “Polymer and carbon nanotube based MEMS accelerometer with modified organic electronics and thin film transistor,” Smart Structures and materials, Proc. of SPIE Vol. 5055 (2003), pp. 87 – 97.

[28] V. K. Varadan, and V. V. Varadan, “3D MEMS structures and their applications”, (Invited paper presented at the International Symposium on Microsystems, Intelligent Materials and Robots), Tohoku University, Japan, 1995

[29] V. K. Varadan, X. Jiang, and V. V. Varadan, *Microstereolithography and other fabrication techniques for 3D MEMS*, John Wiley Press, 2001.

[30] http://www.mems.sandia.gov

[31] D. M. Tanner, J. A. Walraven, L. W. Irwin, M. T. Dugger, N. F. Smith, W. P. Eaton, W. M. Miller and S. L. Miller, "The Effect of Humidity on the Reliability of a Surface Micro-machined Microengine", Proceedings of IRPS, San Diego CA, 1999, pp. 189-197.

[32] S. T. Patton, W. D. Cowan and J. S. Zabinski, "Performance and Reliability of a New MEMS Electrostatic Lateral Output Motor", Proceedings of IRPS, San Diego CA, 1999, pp. 179-188.

[33] Proceedings of the EOS/ESD Symposium, 1979-1999 and D.G. Pierce, "Electrostatic Discharge (ESD) Failure Mechanisms," *IEEE Int. Reliability Physics Symp*., Tutorial Notes, Topic 8, 1995, pp. 8.1-8.53.

[34] J. A. Walraven, E. I. Cole Jr., J. Soden, D. M. Tanner, R. E. Anderson, and L. R. Irvin "Electrostatic Discharge and Electrical Overstress Susceptibility in MEMS: A New Failure Mode," Proceedings of SPIE Micromachining and Microfabrication, MEMS Reliability for Critical Applications, 2000, pp. 30-39.

[35] J. A. Walraven, J. M. Soden, E. I. Cole Jr., D. M. Tanner, and R. E. Anderson, "Human body model, machine model, and charge device model ESD testing of surface micromachined microelectromechanical systems (MEMS)," *EOS/ESD 2001 Symposium*, pp. 3A.6.1 – 3A.6.11.

[36] Goldsmith C., Ehmke J., Malczewski A., Pillans B., Eshelman S., Yao Z., Brank J., Eberly M. Proc. IEEE MTT-S 2001 Intern. Microwave Symp., pp. 227-330, 2001.

[37] Yao J.J. Topical Review, "RF MEMS from a device perspective," Micromech. Microeng. 10, R9-R38, 2000.

[38] De Wolf I, van Spengen, W. M., "Techniques to study the reliability of metal RF_MEMS capacitive switches," to be published, Proc. of ESREF 2002.

[39] E. I. Cole Jr, J. M. Soden, "Scanning Electron Microscopy Techniques for IC Failure Analysis", Microelectronic Failure Analysis Desk Reference, 3rd Edition, 1999.

[40] K. A. Peterson, P. Tangyunyong, and A. Pimentel, "Failure Analysis of Surface-Micromachined Microengines", *Materials and Device Characterization in Micromachining Symposium*, SPIE Proceedings, Santa Clara CA, 1998, **Vol. 3512**, pp. 190-200.

[41] K. A. Peterson, P. Tangyunyong, and D. L. Barton, "Failure Analysis for Micro-Electrical-Mechanical Systems (MEMS)", Proc. of ISTFA, Santa Clara CA, 1997, pp. 133-142.

[42] J. A. Walraven, T. J. Headley, A. N. Campbell, and D. M. Tanner, "Failure Analysis of Worn Surface Micromachined Microengines", *MEMS Reliability for Critical and Space Applications*, Proc. of SPIE, Santa Clara CA, 1999, **Vol. 3880**, pp. 30-39.

[43] J. A. Walraven, E. I. Cole Jr., L. R. Sloan, S. Hietala, C. P. Tigges, and C. W. Dyck, "Failure Analysis of Radio Frequency (RF) Microelectro-mechanical Systems (MEMS)," Proc. of SPIE, San Francisco CA, 2001 pp. 254-259.

[44] J. A. Walraven, E. I. Cole Jr., and P. Tangyunyong, "Failure Analysis MEMS Using TIVA," Proc. of ISTFA 2000, pp. 489-496.

[45] Lellouchi D., Beaudoin F., Le Touze C., Perdu P., Desplate R., "IR confocal laser microscopy for MEMS Technological Evaluation," to be published, Proceedings of ESREF 2002.

[46] Krehl P., Engemann S. Rembe C., Hofer E.P. High-speed visualization, a powerful diagnostic tool for microactuators – retrospect and prospect. Microsystem Technologies 5, pp. 113-132, 1999.

[47] J. A. Walraven, P. C. Galambos, E. I. Cole Jr., A. A. Pimentel, G. Roller, A. Gooray, "Failure analysis of MEMS electrostatic drop ejectors," Proc. 27th ISTFA, 2001, pp. 365-372.

[48] http://umech.mit.edu/freeman/talks

[49] van Spengen W.M., De Wolf I., Puers R., Vikhagen E. Optical imaging of high-frequency resonances and semi-static deformations in micro-electromechanical systems (MEMS). Proc. 27th ISTFA, pp. 357-364, 2001.

[50] De Wolf I., van Spengen W.M., Modlinski R., Jourdain A., Witvrouw A., Fiorini P. and Tilmans H.A.C.. Reliability and failure analysis of RF MEMS switches. To be published in Proc. ISTFA 2002.

[51] De Wolf I. In "Analytical applications of Raman spectroscopy", Ed. M.J. Pelletier (Blackwell Science). Chapter 10. Semiconductors. Pp. 435-472, 1999.

[52] De Wolf I., Jian C., W.M. van Spengen. The investigation of Microsystems using Raman spectroscopy. Optica and Lasers in Engineering 36, pp. 213-223, 2001.

[53] De Wolf I. Topical Review: Micro-Raman spectroscopy to study local mechanical stress in silicon integrated circuits. Semicond. Sci. Technol. 1996; 11: 19.

[54] van Spengen W.M., De Wolf I. and Knechtel R., Experimental two-dimensional mechanical stress characterization in a silicon pressure sensor membrane using micro-Raman spectroscopy, SPIE Micromachining and Microfabrication, Sept. 2000.

[55] Maeda Y., Yamamoto H. and Kitano H. J. Phys. Chem. Vol. 99, No. 13, pp. 4837-4841, 1995.

[56] Ager III J.W., Monteiro O.R., Brown I.G., Follstaedt D.M., Knapp J.A., Dugger M.T. and Christenson T.R. Performance of ultra hard carbon wear coatings on microgears fabricated by LIGA. Proc. Materials research Society, 1998.

[57] Rasras M.S. De Wolf I., Maes H.E. Spectroscopic identification of light emitted from defects in silicon devices. J. Appl. Physics, Vol. 89, 1, pp. 249-258, 2001

[58] De Wolf I., Rasras M. Spectroscopic photon emission microscopy: a unique tool for failure analysis of microelectronic devices. Microelectronics Reliability 41, pp. 1161-1169, 2001.

[59] Unpublished material from IMECvzw, Leuven, Belgium.

[60] A. Jourdain, P. De Moor, S. Pamidighantam and H.A.C. Tilmans, "Investigation of the hermeticity of BCB-sealed cavities for housing (RF-)MEMS devices", *Proceedings MEMS 2002*, Jan. 20-24, Las Vegas, pp. 677-680, 2002

[61] P. De Moor, K. Baert, B. Du Bois, G. Jamieson, A. Jourdain, H. Tilmans, M. Van De Peer, A. Witvrouw, C. Van Hoof, “Low temperature zero-level hermetic packaging for MEMS based on solder and polymer bonding”, *IMAPS 5th Topical Technology Workshop on MEMS*, Related Microsystems and Nanopackaging, November 2003

[62] M. Nese, R.W. Berstein, I.R. Johansen, R. Spooren. Proceedings of Transducers ’95, pp. 94-99, June 1995

[63] De Wolf I., van Spengen W.M., Modlinski R., Jourdain A., Witvrouw A., Fiorini P. and Tilmans H.A.C.. Reliability and failure analysis of RF MEMS switches. To be published in Proc. ISTFA 2002.

[64] Courtesy of IMEC, 2004.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Failure Analysis and Reliability of Optoelectronic Devices

Robert W. Herrick
JDSU, San Jose, California, U.S.A.

1 Introduction

It is commonly said that optoelectronics fabrication today resembles integrated circuit manufacture 20-30 years ago. The explanation for such a lag centers on the smaller volume of production of optoelectronic devices and the relative immaturity of the technology. Failure analysis (FA) techniques are if anything perhaps further behind the need than was the case for silicon integrated circuits decades ago. An optoelectronics vendor who sells $50 million of optoelectronic parts in a year might be a leader in their market segment. Compare this with Intel, which sold over $35 billion in 2009.

With such a small revenue base, the optoelectronic vendor cannot afford a large, well-equipped and staffed FA lab. Optoelectronics suppliers tend to stick to a few FA techniques that are simple, cheap, and quick, and have worked well in the past. The physics of failure is not well understood. Controversies that raged in the 1970s about how dislocations grow still remain unresolved to this day. The parts were simply improved until they met customer's expectations, and then further research was discontinued.

This introduction is not intended to sound pessimistic; it seeks only to lower expectations for readers of this volume, many of whom may come from a background of integrated circuit failure analysis. By contrast, failure analysis of optoelectronics is a nearly unexplored frontier. While very challenging and interesting, at the same time there is relatively little background work on which to build.

This chapter does not contain a comprehensive listing of failure mechanisms for optoelectronic components. Excellent books and articles have published information more specific to the varied types of optoelectronic devices. We have tried to highlight a few key principles in this chapter, such as the need to identify failures as "wearout" or "maverick" failures a key concept regardless of the type of device being investigated. We also have covered the capabilities of many key failure analysis tools. The chapter includes many examples from degraded semiconductor lasers — the field in which the author has two decades of experience.

1.1 Types of optoelectronic devices

Optoelectronic components can be readily classified in a few categories. The first category contains active light-emitting components, such as semiconductor lasers and light emitting diodes

The second category contains electrically active but non-emitting components, such as photodetectors, modulators, and switches. The third category contains inactive components, such as waveguides, filters, or splitters, where no bias is applied.

In this chapter, we focus on the first category, and particularly on semiconductor lasers. Semiconductor lasers perform a remarkably difficult job, with minority carrier concentrations (and recombination rates) thousands of times higher than those observed in most other electronic components. These high carrier concentrations drive defects to grow and kill the laser unless the device is nearly perfect to begin with. And when the lasers do fail, it is often difficult to isolate the cause of failure. In fact, it often requires nearly atomic resolution to find the underlying cause.

Failures are rare for the other types of non-emitting optoelectronic devices. When failures do occur, their causes tend to be more obvious (e.g., contact corrosion or packaging-induced failures).

1.2 Types of semiconductor laser active regions

It is helpful to understand that there are three types of semiconductor lasers, from the standpoint of active region degradation. The first, and ideal type, would be those that have non-propagating defects. The defects do not propagate because they are not mid-gap recombination sites, and therefore, non-absorbing and not electrically active under most conditions [1]. Such lasers are vulnerable to damage from electrical overstress (EOS), for example, but after damaged, usually show little or no increase in degradation rate beyond what was observed immediately after they were damaged. Such lasers are still vulnerable to gradual degradation, but such degradation is far more predictable. Unfortunately, relatively few of the types of commercially-available lasers are from this family. The second type of laser is one with mid-gap recombination, but slow or no defect growth due to pinning of defects through compressive stress. An example would be GaAs-based lasers with InGaAs or InAlGaAs active regions [2]. The third type of laser is one with rapidly propagating defects. If any damage or flaw is present, the laser is vulnerable to potential failure. Substantial work on such lasers goes into screening through burn-in and visual examination. Such lasers include 850 nm GaAs vertical cavity lasers (VCSELs) and 1.3um AlGaInAs-on-InP (high temperature) in plane lasers. Understanding which class the laser of interest fits into can help you develop the right control strategy, and guide your FA approach.

1.3 Basics of semiconductor lasers

The author realizes that many of the readers may be experts in the failure analysis of silicon ICs, and have little or no background in the workings of semiconductor lasers. For this reason, we have added a few paragraphs explaining some of the rudiments of operation and common nomenclature.

Semiconductor laser design and fabrication take place in two separate planes. The vertical structure is determined by epi growth, while the lateral structure is determined by the masks used during wafer processing.

The first step in making a laser is epitaxial growth, usually in a high-temperature reactor with metal-organic chemical vapor deposition (MOCVD) depositing a few microns of material on a GaAs or InP substrate. In this step, cladding and waveguide layers are grown to contain the carriers and photons, and a "gain layer" is grown near the middle of the structure. Photons are created in this layer when current is passed through the device. This gain layer is usually less than 0.1 µm thick. It is also referred to as the "active layer" or "quantum well(s)", due to quantum mechanical effects present in most recent designs where very thin layers are used. Bounding layers are grown above and below the gain region to keep the photons continually passing through the gain layer, as well as to contain the electrons and holes. Two designs are popular: one is the "stripe" or "in-plane" lasers, where the photons travel back and forth along the plane of the active region and come out the cleaved facets, as shown in figure 1. The second design is the vertical-cavity surface-emitting laser (VCSEL), where the photons travel up and down, and are reflected off the mirrors (a.k.a. distributed Bragg reflectors or DBRs). See figure 2. In both cases, most of the photons are confined to within less than a micron above or below the gain layer, most of the time. Defects existing further above or below the active region than one micron tend not to cause problems unless they grow to extend closer over time.

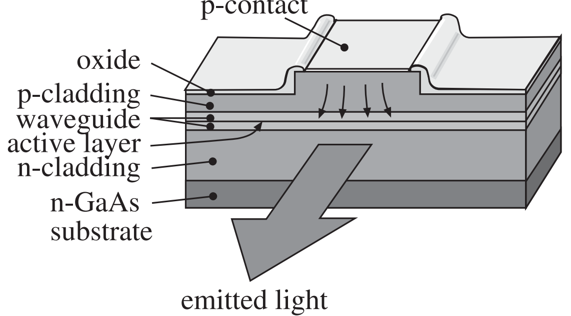


Fig 1: Picture of ridge-waveguide laser. Arrows below p-contact show where current flow is concentrated.

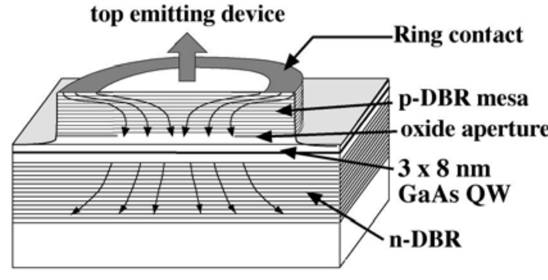


Fig 2: Picture of vertical cavity laser (VCSEL). Arrows below emitting aperture show where current is concentrated.

After the structure has been grown in the vertical plane, fabrication in the cleanroom takes place, and the device structure is defined in the lateral directions with photomasks. Much of the processing is similar to that used for silicon ICs, but the requirements are far less rigorous, and equipment from the 1970's and 1980's is often used. The process is designed to make most of the chip electrically non-conductive. This is usually done by introducing an insulating layer (such as an oxide layer) between the semiconductor and the contact metal. Alternately, proton bombardment can be used to increase the resistance of the semiconductor in the unpumped regions. Only a small portion of the wafer will be left unaffected in this step: either a narrow stripe the length of the chip for in-plane lasers, or a small dot for VCSELs. The area where current flows is known as the "emitting area," the "pumped area," or the "current aperture". The current and the photons are usually both confined laterally to within a few microns of this pumped area, and again, defects that are further away than a few microns do not usually affect operation. This sounds like a good thing, but herein lies one of the biggest problems faced by GaAs-based lasers: defects quite far away from the emitting area can pass undetected, only to slowly grow toward the emitting area, and kill the laser months or years after being fielded in a system. It is this process we will describe in detail in the next section.

After cleanroom fabrication is complete, the devices are either scribed and broken or diced with a wafer saw. Great care must be taken to not induce damage at the chip edge in many designs, since such damage can subsequently result in defects that will later grow to cause problems. A common chip size is 250 x 350 µm – visible by the unaided eye, but only barely – about the size of a grain of sand. Again, only a small fraction of the area is actually electrically active, as shown in figure 3.

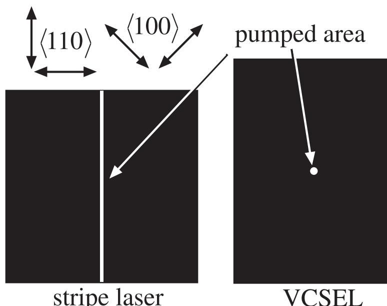


Fig 3: Top view of typical stripe laser (left) and VCSEL (right) show how small a portion of the total die area is actually used to create light. The die area is shown in black, with the pumped area shown in white to scale. At the upper left, the Miller crystal index directions are shown.

One additional piece of nomenclature needs to be brought up at this point: "Miller crystal indices" are generally used for specifying wafer orientations, and are not intuitively obvious to those who haven't had Material Science or Solid State Physics training in this subject. When you see "⟨100⟩ DLD" (or "dark line defect") it should be read as "a dark line defect running along the one-zero-zero crystal axis", or "a DLD running diagonal to the die edges", and *not* as "one hundred dark line defects"! The crystal indices are also drawn on figure 3. The defects only appear to grow in this direction from a macroscopic scale; microscopically, the axes are quite different.

2 Physics of Failure

In this section, we will briefly describe the material science behind some causes of device failure. For greater detail, the reader can consult Chapter 2 of the author's Ph.D. dissertation, and the references therein. [3]

2.1 Climb dislocations

GaAs-based lasers have always been vulnerable to climb dislocations. Three decades ago, threading dislocations from substrate defects were the nucleating factor for those early semiconductor lasers. Today, the main cause of failure in VCSELs is mechanical damage or ESD damage. These forms of damage nucleate ⟨100⟩ Dark Line Defect (DLD) networks, which grow out of the damage sites and cause device failure, as shown in figure 4.

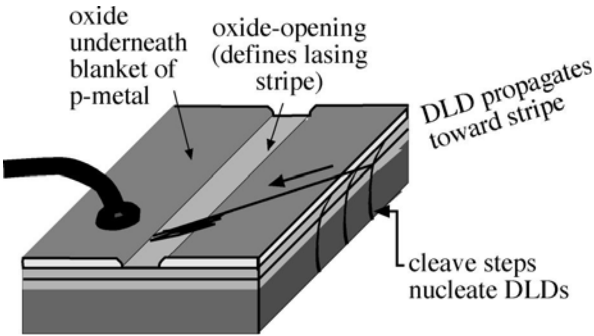


Fig 4: Drawing of ⟨100⟩ DLD originating from a crack at the edge of the device and traveling toward the stripe.

In general, a significant amount of direct carrier injection only happens in the electrically pumped region, or within a few carrier diffusion lengths of this region, where the diffusion length is typically on the order of 1-2 µm. Inside this area, DLDs grow extremely quickly, typically about 1 µm every few minutes. Outside this area, DLD growth is believed to progress by an entirely different mechanism. Spontaneous emission from the emitting area optically excites electron hole pairs on the side of the dislocation network facing the emitting source. This effectively "attracts" the DLD to grow toward the source. Such growth happens very slowly when the DLDs are far from the emitting area. It can take hundreds of hours for the DLDs to grow even one micron. Once the DLDs reach the emitting area, catastrophic failure occurs quickly. The time scales for ⟨100⟩ DLD growth in GaAs stripe lasers was shown nicely by an experimental result published by the late Robert Waters [4].

Note as well that properties of the DLD network (both high optical absorption and carrier recombination) are expected to prevent the network from growing outward, away from the emitting area. On VCSELs where DLDs originate inside the pumped area, DLDs don't usually extend more than 5 µm beyond the oxide edge.

2.2 Glide dislocations

Dislocation glide is very different from dislocation climb. While climb involves adding material to a growing dislocation network, glide is "conservative", meaning no material is absorbed or emitted by the dislocation [3].

Control of ⟨100⟩ climb dislocations focuses both on optimizing growth conditions and on eliminating mechanical damage to the surface and edges. ⟨110⟩ glide elimination concentrates primarily on eliminating stress inherent in the chip design (*e.g.*, dielectrics and metals) and device packaging (*e.g.*, soldering stresses).

On a microscopic scale, the DLD grows as shown in figure 5. ⟨110⟩ DLDs grow more slowly and are less electrically active than the ⟨100⟩ DLDs. Only the kink sites along the dislocation edge are electrically active. These make up less than 1% of the sites on the dislocation edge.

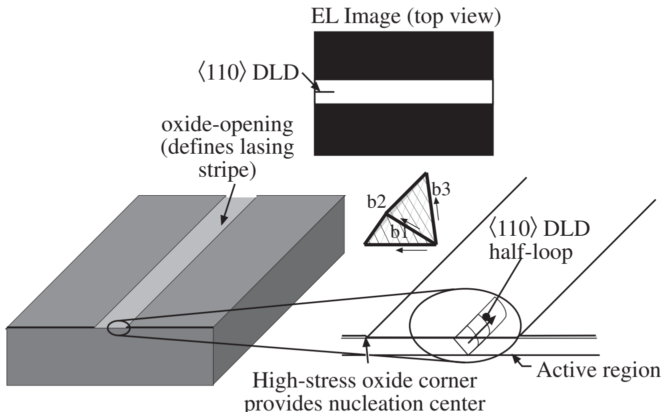


Fig 5: Growth of a ⟨110⟩ DLD in a typical stripe laser.

2.3 Dark Spot Defects (DSDs)

When failure analysts have studied device degradation only at low magnifications using electroluminescence (EL), they have often used imprecise terms, such as "dark spots". Depending on the device being used, the dark spots may be due to contact degradation, spots from ESD damage, or even by emerging dislocation loops not yet long enough to be considered a "dark line" defect. Without further study, it may be difficult to determine which of these causes, if any, are responsible for the DSDs. In this sense, "DSD" is a description too vague to be useful in most of the applications I have observed.

2.4 Gradual Degradation

Researchers have observed that using non-optimized growth conditions reduces the lifetime of lasers by a multiple of ten or more. In particular, arsenic-poor growth seems to lead to reduced lifetime. From this observation, some observers have hypothesized that group-III interstitials contribute to building the DLDs (the "extrinsic theory"). In most devices, where nothing exists to nucleate a DLD, low-level impurities from epitaxial growth gradually cluster and form microloops. The microloops grow to a diameter of 15-50 nm (see figure 6). They resemble the climb dislocations discussed above, inasmuch as an additional half plane forms the microlooop [5]. Dozens of these loops can eventually form in a 1x1 µm square [6], and increase optical loss within the lasing cavity. In addition, the concentration of carrier traps also rises as the device degrades. The traps contribute as well to the degradation of the device, even if they don't combine into larger, observable structures [5,7].

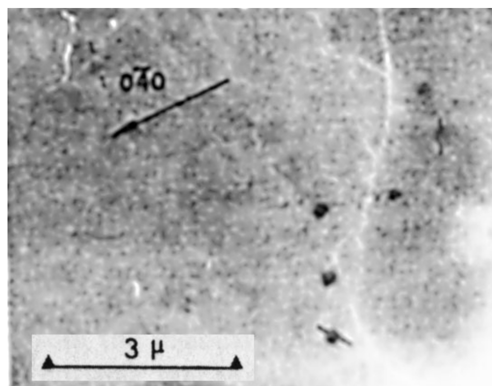


Fig 6: TEM of microloops from a gradually-degraded sample. Image courtesy of Prof. Pierre Petroff [8].

3 Common Failure Mechanisms

3.1 The difference between maverick and wearout failure mechanisms

When a failure occurs, one of the most important questions that will be asked is "should we expect to see more failures like this in the future, on lasers from the same batch?" The answer varies. When the failed part was used outside its design limits, the number of other failures expected may be very low. If the part was well designed but imperfectly manufactured, some other failures may be expected. If the part contains a design defect, many failures may be expected.

A wearout failure occurs when something in the part is used up or exhausted during the aging process. For example, tire tread wears off over the life of the tire. Such wear is expected. By using batch testing, one could predict when other tires from the same batch might fail due to tread wearout.

By contrast, a maverick failure mechanism occurs when a "weak part" was not built as designed. An example would be tires where the tread delaminated – this might happen on a fraction of a percent of bad tires. Thus testing only a small sample from a batch would probably not reveal how many bad tires existed, since the randomly sampled tires would probably all be good. Figure 7 shows another example of a maverick failure. Spittle contamination on the bondpad prevented the wirebond from making good contact as it would with a clean bondpad. The wirebond then becomes detached during operation, causing failure due to an open circuit at the wirebond.

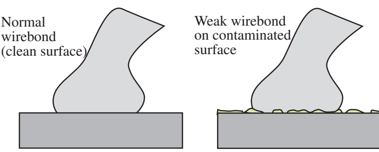


Fig 7: Weak wirebond (right) may have only a small fraction of its surface in metal-to-metal contact. It is then vulnerable to premature failure.

It has been said that rare maverick failures of this type are primarily what controls reliability [9]. If so, then trying to "test reliability in" cannot succeed. Even testing 10,000 parts, with no failures, cannot guarantee that future problems in manufacturing will not occur and will not cause a high failure rate next week or three months from now. To ensure that reliable parts will be delivered requires that the institution of best practices. Among them are failure modes and effects analysis (FMEA) to anticipate and eliminate potential failure mechanisms – for example by installing a spittle shield over the inspection and wirebonding microscopes. Configuration control – having a well-documented method of making the parts that is consistently followed – will also have positive effects.

Finally, we should consider "infant failures" as a possibly separate category. This small sub-population can be screened out. Accelerated operation for a few hours or days (a "burn in") will usually cause any weak parts to fail. They can then be detected before the defective parts are shipped to customers. Maverick failures may take longer to show up, and thus they may escape detection by a burn-in.

For many types of devices, maverick failures cause sudden catastrophic failure, while wearout causes slow, gradual degradation. This is depicted below in figure 8.

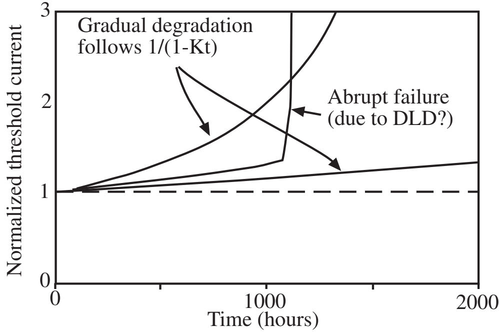


Fig 8: Graph of normalized threshold current versus time on lifetest shows that some wafers degrade gradually, while others suffer abrupt failure.

3.2 Common Laser Failure Mechanisms

As mentioned earlier, semiconductor lasers have minority carrier populations (and hence recombination rates) thousands of times higher than those observed in other semiconductor devices. Dislocations which would require red-hot temperatures to move without current injection can grow even at room temperature using only a normal amount of drive current. As a result, even minor imperfections can result in reliability failures. In spite of this challenge, the semiconductor laser industry has learned to reliably produce millions of lasers every year, for just a few dollars each, with hundreds of years of useful life. However, semiconductor laser reliability has cannot yet be "taken for granted" or "guaranteed by design," at least for low-cost lasers.

3.3 Wearout failure mechanisms

3.3.1 Epitaxial contamination

Semiconductor lasers are usually created in "growth chambers" using ultra-pure gasses. If any of these gasses has even relatively low (ppm) levels of contamination, the life span of the lasers can be significantly reduced. Oxygen leaks in the reactor lead to oxygen contamination in the wafer, which often also reduces the life span of devices.

Fortunately, the reduction in lifetime affects all parts on the wafer nearly equally. As a result, a lot acceptance test can easily screen out the defective parts. Regrettably, the lag between having an oxygen leak on the reactor and detecting it at the end of the process of wafer fabrication and testing may be as long as two months. This increases the likelihood that a large amount of inventory in the pipeline will have to be scrapped. Secondary ion mass spectroscopy (SIMS) is a popular technique for detecting epitaxial contamination It can detect contamination as low as one part per billion for many species.

3.3.2 Ionic contamination

Other problems might include ionic contamination originating in a wet solution used in wafer fabrication, copper contamination in the contact metallization, or other similar causes. Since the contamination is often dilute except in the (very small) pumped region, it can be technically difficult to identify the source of such contamination. Degradation and undesired alloying of the contact metallization has been observed as one primary failure mechanism for early long-wavelength LEDs and lasers.

3.3.3 Facet damage

In many edge emitters, degradation near the emitting facets may occur. This is particularly a problem in two types of lasers: "uncooled" AlGaInAs-on-InP 1.3um lasers used for 10Gb/s data links, and high power lasers. Dealing with the AlGaInAs lasers first, they have propagating climb dislocations, which are particularly likely to start from the facets. Special screening regimens based on driving to high optical power are commonplace as a way of eliminating "weak" devices.

For high-power semiconductor lasers, facet oxidation particularly affects the popular AlGaAs-based lasers, as does the increased intra-cavity absorption of lasing photons near the facet due to bandgap shrinkage. A positive feedback loop can occur where heating from the absorption creates further shrinkage, which, in turn, raises absorption even more, until catastrophic optical mirror damage (COMD or COD) occurs [10,11].

To improve reliability, special steps must be taken to prevent surface recombination. For lasers with modest output power needs (that is, with less than 30mW of single-mode power),

surface coatings may be used to prevent further oxidation. Lasers also sometimes eliminate metallization near the facet to reduce the surface current flow. For higher-power devices (with more than 100mW of single mode power), regrowth or special processing to eliminate the active region near the facet are popular. Still another technique cleaves the devices in an ultra-high vacuum and applies a coating immediately, before the facet has the opportunity to oxidize [12].

Auger sputtering or other surface-sensitive techniques are often used to detect facet oxidation. When examining a scanning electron microscope (SEM), image, it is not uncommon for facet oxidation to be visible where the laser power is highest [13]. In high-power lasers, power densities of $10MW/cm^2$ are not uncommon. Note that these densities are greater than the optical density at the surface of the sun). Even if COMD can be avoided, we can observe high degradation near the facet region. It is one cause of normal wearout failure on edge-emitting lasers [14].

Finally, it is worth noting that carbon deposition on the facets has been observed as a cause of darkening and reduced power in at least a few papers. It was detected by SEM imaging. The cure was to eliminate the use of solvents in packaging, which were the original source of carbon.

3.3.4 Normal Wearout

Over the past decade, normal wearout has continued to become more of a problem for many popular applications. One example is high data rate semiconductor lasers using on-off modulation. For the 10 – 25 Gb/s data rates commonly used, extremely high average bias levels are used that require special recipe optimization to obtain acceptable lifetimes. Another example is high-power applications, such as DVD burners used for writing data discs where more power enables faster write times. If all the other premature failure mechanisms mentioned above are avoided, the typical lifetime of the device can be tens or hundreds of years for those that do not need to be driven at high currents. The inherent failure mechanism can be observed only after a long period of aging the devices at extremely high temperature and high current to accelerate the aging process. For edge-emitting lasers, typical failure mechanisms include gradual degradation of the active region owing to microloop formation. In such cases, the microloops act as electrical recombination centers, and more importantly, they add optical loss, absorbing lasing photons which pass near them.

This degradation is often concentrated near the facets. EBIC studies have shown the microloops to be concentrated within 15 μm of the facets [14], as shown in figure 9. Additional studies have compared aged lasers with unaged companion lasers. The aged lasers had been severely degraded during life testing. However, after cleaving off the near-facet regions on both ends, their performance was identical to those of the companion lasers, showing there was no degradation in the bulk of the material.

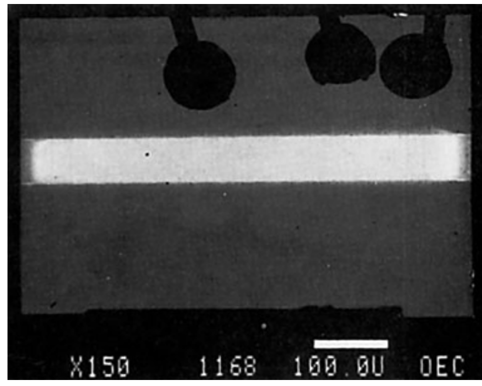


Fig 9: EBIC image (top view) shows gradual degradation near the facet as darkening along the right and left ends of the stripe (after [4]).

Extensive failure analysis has been performed on laboratory-aged VCSELs, with the surprising result that no measurable degradation of the active region is observed even in devices where peak power has fallen dramatically. The explanation involves the "current shunting failure mechanism." The belief is that dopants are passivated (probably by forming complexes with mobile hydrogen), thus raising the resistance near the device center. When this occurs, it pushes current away from the center where it is needed, and thereby lowers the laser's efficiency. This hypothesis has not been *directly* confirmed (*e.g.*, by mapping effective dopant concentrations in a cross-sectioned VCSEL). We have only indirect evidence, namely a lack of measurable degradation of the active region by EBIC, CL, TEM, etc. See references 3 and 15 for more details.

Note also that poor design (*e.g.*, excessive dopants), and hard use (poor cooling or high drive currents) can also shorten substantially the lifetime of these devices. GaN-based lasers, such as those used in "Blu-Ray" disc drives show evidence of gradual degradation as well, with point defects evidently arising from threading dislocations. [16].

3.4 Maverick Failure mechanisms

3.4.1 Epitaxial defects

Threading dislocations were one of the main causes of failure in the early days of semiconductor laser growth, as shown in figure 10. Since then, defect densities of the substrates have been dramatically reduced, to a point at which they now cause failure only rarely. When threading dislocations are present in GaAs-based lasers now, they are not usually caused by substrate quality limitations, but rather by particulate contamination which falls on the substrate before epitaxial growth.

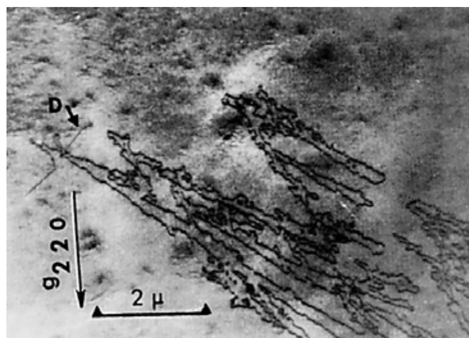


Fig 10: Plan-view (i.e., top-view) TEM of a ⟨100⟩ DLD growing out of a threading dislocation. Threading dislocation is highlighted with the letter "D". Photo courtesy of Prof. Pierre Petroff of UCSB, after [17].

One type of laser where threading dislocations from the substrate *are* still a limiting factor is the GaN-based blue semiconductor laser. At the time of this writing, low defect density GaN substrates are not yet available, but are an area of active research.

3.4.2 Mechanical damage

After 10,000 or more devices are processed on a single wafer, they must be broken into individual devices (or small arrays) by a process known as "singulation." The die edge often can be damaged in this process, usually by sawing or a "scribe and break" step. Any scratches or cracks can serve as nucleation sites for dislocation networks. Beyond mesa or chip edge damage, such cracking can also include internal breakage, such as at the semiconductor / oxide interface for oxide VCSELs [18, 19]. Mechanical damage then nucleates ⟨100⟩ DLDs in GaAs-based lasers, as shown in figure 11 below. This EBIC image is a top-view of a stripe laser similar to that shown in figure 4 earlier.

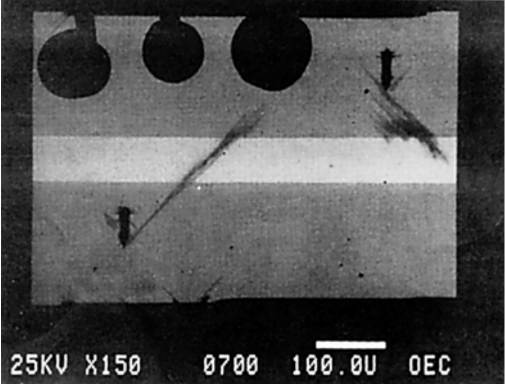


Fig 11: Plan-view EBIC of 60x600 µm GaAs QW stripe laser which was intentionally damaged with two diamond scribes. Note the ⟨100⟩ DLDs that travel out of the scribes, and toward the lasing stripe (after [4]).

EBIC can't be used on VCSELs, and other techniques such as backside TIVA are more appropriate for visualizing DLD networks. This will be discussed at length in section 5.5 of this chapter. Plan view TEM is another way of looking for a point of origin, if the DLD network is not too large (i.e., less than 40 x 40 µm). Tracings of TEM photos are shown below in figure 12, and record the extent of the DLD network that failed due to mesa edge damage.

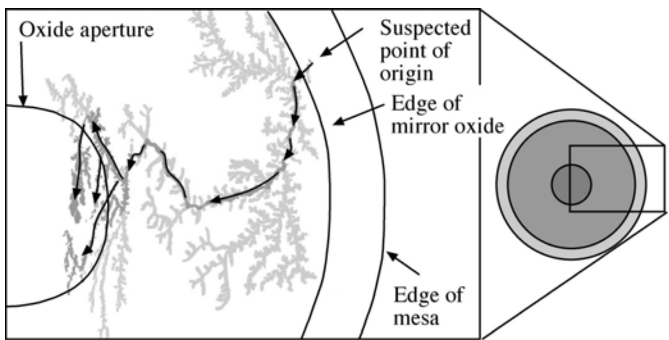


Fig 12: Plan view TEM images were traced to show the extent of the ⟨100⟩ DLD network. Drawing at right shows what portion of the device had DLDs. Arrows show how the DLDs traveled from the right edge of the mesa into the emitting area inside the oxide aperture. The branches to the sides of the arrows (shown in light grey) play no important role in the final device failure. (after [20]).

3.5 Failure from overstress (EOS or ESD)

A common cause of failure in semiconductor lasers is electrostatic discharge (ESD) or Electrical overstress (EOS). ESD tends to involve very short pulses that are less than the thermal time-constant of the device; EOS involves longer pulses, and usually shows signs of thermal transients, such as localized melting in the device. Extensive coverage of ESD and EOS is available in other chapters of this volume, and is also the subject of an excellent annual symposium [21] and good books [22]. However, it is important to point out that while most semiconductor lasers are indeed quite vulnerable to ESD and EOS, it is still overdiagnosed by failure analysis personnel. Caution must be used in making the diagnosis, and distinctive signatures that are only observed from ESD or EOS should be present before making a conclusive diagnosis that one of them was present.

In VCSELs, when ESD is applied to devices using laboratory test equipment, damage is focused along the aperture edge. This can be seen in electroluminescence images from the top. During normal operation, the damage sites expand quite slowly for a while, but eventually form a network of ⟨100⟩ DLDs. Once a DLD network forms, it travels rapidly across the device, and kills it within a few minutes.

ESD damage can be seen clearly in cross-sectional TEM images [23], such as figure 13, where breakdown leaves evidence of damage – provided the cross section goes through the damaged area, which can be tricky to arrange for if the damaged area is not large. The excellent "survey" by David

Mathes is highly recommended reading for those interested in understanding what various types of damage look like in TEM. [24] . Longer-pulse EOS can also be a significant issue in many lasers as well, due to over-driving during testing or use [25].

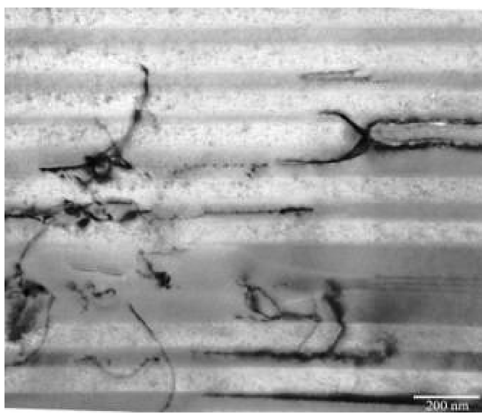


Fig 13: TEM cross-section image of ESD-damaged sample shows evidence of ESD damage, near the oxide tip (upper right). In addition, the quantum wells appear distinct toward the right, but they have intermixed in the opening to the left due to melting from the ESD pulse.

3.6 Common LED Failure Mechanisms

The use of light emitting diodes (LEDs) has grown rapidly in importance in the recent past, particularly in high-power solid-state lighting applications. LEDs tend to have different failure mechanisms than semiconductor lasers, owing both to the fundamentals of operation and to the packaging [26].

In LEDs, we frequently observe the diffusion of dopants into the active region during operation can cause a gradual drop in efficiency. In fact, the LEDs with the best initial performance (with low dopant concentrations in the active region) often degrade the most rapidly. Intentional introduction of small amounts of contaminants such as oxygen can sometimes be used to form complexes that prevent dopant migration. Dopant diffusion can often be diagnosed using SIMS.

Packaging-induced stress has been another common source of problems. The plastics used to package most LEDs have a much higher coefficient of thermal expansion (CTE) than the chips or lead frames they surround. The plastics are usually cast at higher temperatures than normal operating temperature they then shrink as they cool creating static stress at normal operating temperatures. Temperature cycling can also lead to fatigue problems that result in broken wirebonds. One popular solution is to cover the die in a soft silicone gel so that the plastic doesn't contact the die. Unfortunately the ionic contamination in the gel can create problems of its own. Packaging-related problems account for a significant fraction of the failures – one question that can be asked is whether the die still puts out the specified amount of power after the packaging has been removed.

Much of the growth in LED sales comes from blue and violet GaN-based LEDs that can create white light by using phosphors. Early work showed that the plastics that were used "yellowed" with age, absorbing increasing amounts of the blue light. The yellowing dominated the degradation observed [27].

3.7 Common Detector Failure Mechanisms

Among photodetectors, dark currents are typically on the order of nanoamps (nA), with very low minority carrier concentrations. Thus we need not worry about growth of dark-line defects. Contact degradation and corrosion are more relevant problems. They can usually be easily detected in high temperature reverse-bias tests and biased 85/85 tests, respectively. Internal degradation is usually easily measured as increased dark current, while visible degradation on the top surface from corrosion is generally easily visible with optical microscopy. [28]

3.8 Common Failure Mechanisms in Passive Optoelectronics

Passive optoelectronics usually have failure mechanisms caused by package-related failures. Examples include the shifting of optical fibers that couple the light in, delamination of optical coatings or gel-coats, and changes in the properties of materials owing to their absorption of moisture.

4 Determining the cause of failure

Because every device is different, it is difficult to make general recommendations on how to determine the cause (or causes) of failure. The analyst must first have a strong understanding of the physics of failure, and of the most useful techniques for the diagnosis of common failure mechanisms.

The industry suffers when analysts have only elementary understanding of device operation, or of reasonable and likely failure mechanisms. They might see small pieces of dust on the top of a laser, for example, and conclude that the diffusion of chemicals from the dust killed the laser. Or they might suggest that the particle represents ejecta from an ESD event. In this case, we could disprove both hypotheses using elemental analysis (*e.g.*, EDX). In general, however, the analyst will get in trouble if he "looks only where the light is good", rather than the likely location of the real problems. In failure analysis, having a wide range of tools, and knowing how to use them is crucial. "If your only tool is a hammer, then every problem looks like a nail."

4.1 Building up a "failure library"

Ideally we want to be able to identify common failure mechanisms using quick and non-destructive tests, such as power *vs*. current (L-I) characterization, reverse leakage current, or electroluminescence imaging. In many cases, combining a few such characterization methods will allow us to identify positively a single particular failure mechanism. However, such "quick and easy" methods of characterization

rarely can provide the sort of in-depth root-cause information that we need to understand *why* the failures occurred, or how they can be prevented. For this, we need more expensive and time-consuming procedures such as TEM to find or confirm the root cause. With luck, each individual failure mechanism can be correlated with a few easily measured and distinctive characteristics. Then using in-depth FA to confirm the root cause on such a "fingerprint", we can build a "failure library". It would allow us to say, "If you see this sort of thing in electroluminescence, and this sort of thing in the reverse leakage curve, the likely root cause is such-and-such."

Success requires that analysts keep an open mind and a willingness to revisit earlier conclusions if the original diagnosis does not pan out. A humble attitude is a useful asset in failure analysis. Too often analysts sweep conflicting data under the carpet. Often those seeming contradictions are trying to tell you something important.

4.2 Building a "failure tree"

Once the dominant failure mechanisms are understood, one can build a failure analysis tree. All skilled analysts, perhaps without realizing it, use such a decision tree to guide their thinking, even if they haven't gone to the trouble of documenting it.

The failure tree allows analysts to start with quick and non-destructive tests that distinguish between the most likely failure mechanisms. With luck, a few tests can confirm most of the common failure mechanisms at the outset. Additional tests may be needed to confirm the diagnosis, or to find the origin of the failure, but in many cases the process can quickly narrow down the number of relevant possibilities.

In general, the first tests will be a light-out vs. current-in scan, and if that confirms low or no power out, then the next two steps will usually be a electroluminescence image at a moderate level (1-10% of operating, or 20-50% of threshold level), and a reverse-bias I-V trace with a paramter analyzer to look for junction damage. Where to go from there is usually indicated by where other recent analyses have gone, or what failure mechanisms are most common. A failure tree requires good documentation if failure analysis is to be delegated to personnel who do not understand the physics of failure present in the particular device being analyzed.

4.3 Product Return History

Even the best tools can fail when no evidence remains to reveal how the chain of failure started. In such cases, understanding the product return history is often critical when we try to determine the likely failure mode. Some failure modes take a long time to show up, while others will cause failure nearly instantly. Information on whether returns are just coming from one customer, or are distributed across the customer base, may yield important clues about whether the failures are application-specific.

5 Common Optoelectronic FA tools

5.1 Picking the appropriate tool

It is not possible to make general recommendations about FA tools, because different types of optoelectronic devices require very different techniques. Often techniques that work well for one device won't work at all for another. For example, electroluminescence (EL) images are the most important failure analysis technique for LEDs or VCSELs, but luminescence can't easily be imaged from edge emitting lasers, since metal covers both the top and bottom.

As mentioned above, it is valuable and sometime essential to gather return history. The best histories cover both the return in question (*e.g.,* how long was the module in service or on reliability test before failing?) and from returns of other parts which might have a similar failure mode (were failures limited to a particular customer / application or date code?).

At some point in the FA process, destructive analysis is often necessary. We want either to obtain a cross section view (i.e., a view from the side) or a plan view (*i.e.*, a view from the top). In such cases, the decision should depend on whether we know which layer the defect started in. Cross section views may show unusual failure modes. They may also be helpful when we study a device for the first time. However, after some experience is gained, we usually know at which layer the where defects start.

For example, for GaAs-based lasers with $\langle 100 \rangle$ DLDs, the defect network always grows in the quantum wells. In this case, it is almost always more useful to obtain a plan view section. A plan view section allows a view of the entire active region.

By contrast, if a TEM cross section is taken, only a few percent of the material from the pumped area of the active region remains. The rest has been sputtered or ground away. Thus, if a single point of origin were responsible for the failure, one would have to be very lucky to not destroy it with the vast majority of the material cut away during the sample making process. By comparison, the chance of seeing it in the plan view section would be quite good. Examples of each type of sample are shown in figure 14.

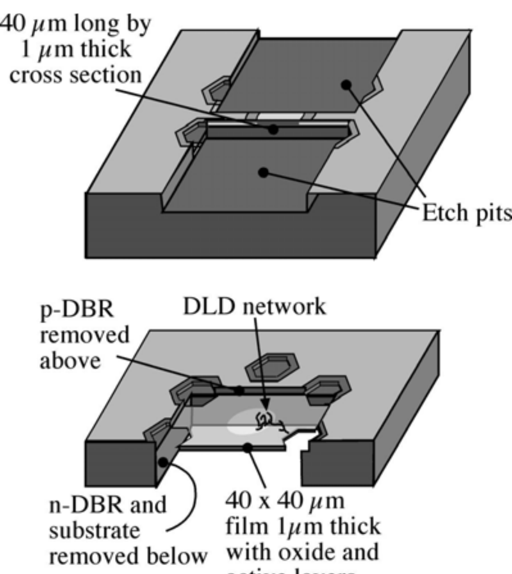


Fig 14: Drawing of oxide VCSEL prepared by FIB for cross-sectioning (top) or plan-view imaging (bottom). Plan view image shows DLD network originating from the etch hole and traveling into emitting area.

If a cross sectional view is appropriate, some sort of failure localization (*e.g.,* an EL image taken from the top) is highly useful. It lets us know *where* to cut through the device. If plan view data is desired, scanning beam imaging (with a laser or e-beam) can be done from the top or from the bottom. Of course, it is always preferable to collect data in a non-destructive manner. Minimizing sample preparation also reduces turn-around time. Top scanning often requires no sample preparation, and is advantageous in this respect. However, sometimes the top contact metal blocks the access of the scanning beam to much of the die. In such cases, some sort of substrate removal can be highly useful to allow better access to the beam, as shown in figure 15 below.

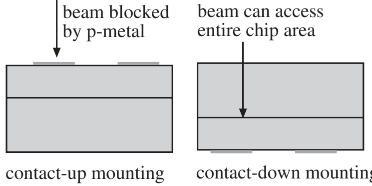


Fig 15: Beam can be blocked by metal contact (left); using special sample preparation, we can embed the sample, and polish off backside, allowing full beam access from the backside while preserving the top contact (right).

5.2 Simple Characterization Techniques

5.2.1 Electroluminescence Images

Electroluminescence is defined as the light emitted from a light-emitting diode or laser when a forward bias is applied. By using a microscope camera sensitive to the wavelength emitted, we can obtain an image of the emission pattern. For a given device, half a dozen or more distinctive patterns of failure may exist. Each pattern probably corresponds to a specific failure mode, or group of failure modes. Examples from oxide VCSELs are shown in figure 16. Chosing an appropriate current is important. Usually, a current lower than normal operating currents is desirable, and always well below lasing threshold for the initial scan. At lasing currents, filamentation occurs, causing bright and dark spots that are due to laser gain dynamics, rather than the presence of permanent defect clusters.

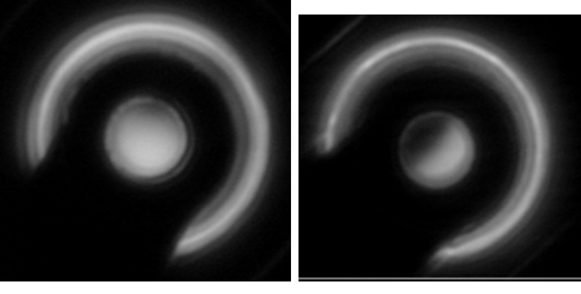


Fig 16: EL from wearout device (left) shows only subtle darkening at the edge; $\langle 100 \rangle$ DLD failure (right) shows profound darkening at the upper left edge.

Sometimes a diagnosis is made easier by noting quantitative details that may escape casual observation. For example, in VCSELs, many analysts could easily confuse the two devices shown in figure 24. However, in many devices aged in life tests until threshold current doubles or triples (like the one shown at left) might require only 0.1mA to get a good image, while $\langle 100 \rangle$ DLD failures like the one shown in the right might require 1mA or more. A linescan is another useful quantitative tool. It notes how rapidly and completely luminescence is extinguished as one moves past a dark line. This is shown in figure 17. The line scan of unaged device shows a symmetrical nearfield pattern. Devices from "ALT" (accelerated life testing) show only a mild drop in luminescence efficiency, even from deeply degraded devices that have totally failed. By contrast, lightly degraded devices which have $\langle 100 \rangle$ DLDs (even ones which still put out several mW of power) show nearly complete extinction of light on one side of the device.

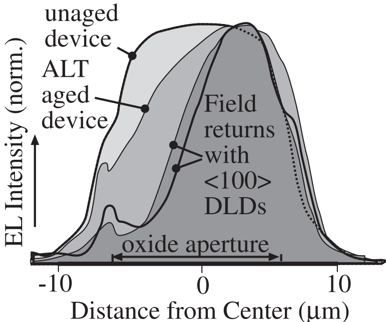


Fig 17: Line scans of unaged VCSELs, laboratory ("ALT") aged VCSELs, and VCSELs with $\langle 100 \rangle$ *DLDs. The line scans were taken from the devices shown in the previous figure, and oriented from the darkest to the lightest points.*

For in-plane lasers, if metal coverage on the topside is not complete, a topside EL scan is usually a good first step in FA after confirming below-normal power. This can usually show whether degradation is isolated to the front facet, the back facet, or the middle of the laser, although without much precise detail. Backside EL gives a far clearer picture as an alternate to EBIC, provided that the substrate is transparent (e.g., a 1.3 or 1.55 µm laser on an InP substrate.) Backside EL is usually preferred if dealing with a lot of surface topography, such as a ridge lasers.

To perform backside EL, the laser can be removed from the header (usually by heating the submount and pushing the die off with a needle probe). Alternately, it could be embedded in a plastic puck along with its mount. Either way, the backside metal is then be polished off. Bare die are placed over a "stripe mask," with the edge of the die making electrical contact to the substrate, and the stripe along the middle of the die being visible through the slide it is mounted on. An EL image is captured with an infra-red sensitive camera and Near-IR optimized microscope objectives.

5.2.2 Filtered EL images

In some cases, placing a bandpass filter in front of the camera can be helpful. This is particularly true for VCSELs. The cavity is designed to bounce photons back and forth hundreds of times before they escape. However, select a wavelength at least 40-50 nm below the emission wavelength (*e.g.,* 790 nm for an 850 nm VCSEL). At this wavelength, most photons are emitted on the first pass, without undesirable reflections. Figure 18 shows the great improvement in image contrast. Where $\langle 100 \rangle$ DLDs are present, the optical absorption is so strong that a high contrast image can be obtained without the use of filters.

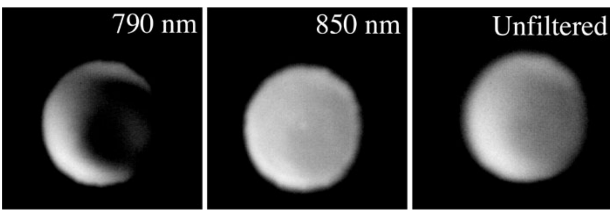


Fig 18: Image with spectral filter outside the mirror stop band (left) shows high contrast, where images at the lasing wavelength, or images without a filter, show low contrast. [3]

Even when the device is imaged with no current, such a filter can show sub-surface details such as the oxide aperture or cracks inside the structure. [19]

5.2.3 Reverse-biased photoemission microscopy (EMMI)

Elsewhere in this volume, entire chapters discuss emission microscopy, whereby a reverse bias is applied, and photon emission takes place at areas where damage to the die has occurred. While this is not a popular technique for FA of optoelectronic die, it has been successfully used, particularly where mechanical damage is suspected. This technique would ordinarily be done at a service lab that has a special dark-box, and a special cooled camera allowing very long integration times. Areas with defect structures that are non-radiative will sometimes show emission in reverse-bias, giving a "complementary image". [29]

5.2.4 EL at very low drive currents

If devices are biased at very low currents, different features of operation often appear. We can then use a cooled astronomical camera to produce EL images of the device at currents as low as 1 µA. If the dark areas remain unchanged from their appearance at 100 µA or above, the cause probably lies in the active region. On the other hand, if image contrast significantly falls, the problem might lie in the mirror layers above or below the active region. Operating at low currents eliminates "current crowding" effects, and allows the current to become more uniformly distributed in the active region as current is reduced.

5.2.5 Microscopic inspection

A careful microscopic inspection often reveals the presence of mechanical damage or contamination responsible for the device failure. It is also quick, non-destructive, and inexpensive.

5.2.5.1 Nomarski microscopy

A Nomarski microscope can enhance the contrast of steps in height by using interferometry. Its output can be very helpful in detecting epitaxial defects and subtle cracks which are often invisible with ordinarily bright-field microscopy.

5.2.6 L-I data

Plotting a graph that shows how light output varies with drive current lets the analyst determine quickly how much damage has occurred. We note how much threshold current and peak output power have changed from their normal undegraded values. (Note that some failure modes are catastrophic and quickly result in total failure. In these cases, no output power will be measured)

When the device is operated at very low currents (µA), we can observe whether its efficiency is reduced. As mentioned in section 5.4, at very low currents, the device can escape current crowding effects, and allow us to distinguish active region degradation from mirror degradation. Emcore VCSELs have shown that the L-I curve displays "kinks" i.e., that its slope rapidly changes from positive to negative for parts damaged by significant amounts of ESD. See fig 19 for examples.

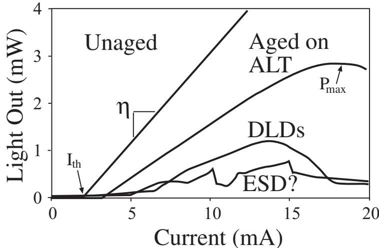


Fig 19: Typical light vs. current (L-I) data for good and degraded parts. The curve with kinks is from a field failure suspected to have failed due to ESD.

5.2.7 Electrical characterization

Many failure modes lead to changes in the device's electrical characteristics. One possibility would be an increase in device resistance, owing to contact degradation. Another would be a drop in turn-on voltage resulting from contact annealing or from ESD that creates leakage pipes. These causes can be analyzed by comparing the current *vs.* voltage (I-V) curve of the device that failed to the curve generated by an undegraded device of the same vintage. Such a comparison is necessary because the batch-to-batch variations are often larger than the changes due to degradation. Thus a reference part is often necessary.

A particularly helpful enhancement uses a parameter analyzer rather than a curve tracer. Data are collected either by using a log I *vs.* V scale (going down to a picoampere of forward current). Alternately, voltage rather than current can be used as the independent variable, with steps in voltage of 0.05V or smaller.

In reverse bias, leakage on good devices is generally minimal, with leakage dominated by the very small number of thermally generated carriers and perhaps edge leakage from the perimeter. For many 850 nm VCSELs, leakage from good devices is generally under a few pA at room temperature. If dark-line defects have formed, significant additional leakage occurs in the DLD network, and up to a few nA can be observed before the device reaches its avalanche breakdown voltage. (The exact amount of leakage depends on the size of the DLD network.) If ESD occurs, leakage can be much larger, owing to damage to the depletion region [23]. See figure 20 for examples.

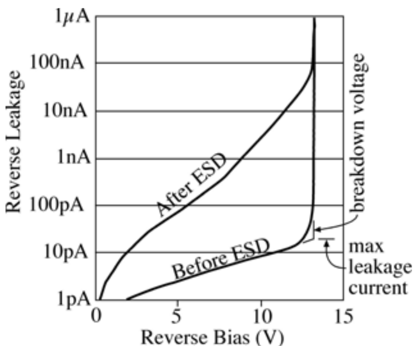


Fig 20: Observations of reverse leakage on an Emcore VCSEL show very low leakage of undegraded device, and three orders-of-magnitude higher leakage after ESD damage at -420V HBM. Leakage increases further after aging.

5.2.8 Thermal Microscopy

In some cases, internal short circuits cause device failure. These short circuits can be diagnosed by using a number of techniques commonly employed in diagnosing short circuits in integrated circuits. These techniques include liquid crystal imaging and fluorescence microthermal imaging. The reader can learn more details from other articles in this volume.

5.3 Scanning Electron Microscopy Techniques

This section covers techniques using the scanning electron microscope (SEM). The first subsection covers standard imaging, in the secondary-emission imaging (SEI) mode. The subsequent sections cover two more specialized modes (EBIC and CL) which require non-standard additions to the microscope.

5.3.1 Standard scanning electron microscopy

Scanning electron microscopy (SEM) is useful for detecting many of the causes of failure, including surface contamination, delamination, oxidation, and melting from overheating. SEM's real weakness is that it only shows a picture of the surface. Usually the causes of failure are hidden below the surface, and thus go unseen by the SEM.

5.3.2 EBIC Electron-beam induced current imaging

Electron-beam induced current (EBIC) imaging requires an electron microscope outfitted with special electrical feed-throughs and a current amplifier. The electron beam generates electron hole pairs, which are swept out through the electrical contacts. Undamaged material provides a certain collected current, which depends on a variety of factors, especially the injected electron beam current. In damaged material with DLDs, carriers recombine at the dislocations before they can

be collected, and a much lower current is measured. These areas appear dark in the EBIC scan. Note that carriers are collected only within the depletion region of the part. Important defects can exist above or below the depletion region, and EBIC would not detect them. This could prove to be useful information, or a disappointing limitation, depending on the researcher's situation.

EBIC is not widely used in electron microscopy. The accessories required to add the capability (*e.g.*, feedthroughs and sample mounts) must be either custom fabricated or purchased from specialist firms.

In using EBIC, one important consideration is the magnitude of the accelerating voltage used in the microscope. For cross-sectional use, where the junction is at the surface, the highest spatial resolution can be obtained by using a low accelerating voltage (<10kV), since this minimizes the size of the area in which carriers are created. By contrast, when EBIC is used in plan-view imaging, edge-emitting lasers typically have junctions 1-1.5 µm below the surface. Higher accelerating voltages of 30-40 kV are needed to reach the junctions.

In devices such as VCSELs, where the mirrors above the junction tend to be 3-4 µm thick, an ordinary SEM lacks the necessary acceleration voltage to probe down to the junction. create a well-formed beam at the junction. Furthermore, the contact metal scatters the electron beam, weakening and dispersing it. For this reason, EBIC is not very useful on VCSELs.

EBIC can be used in cross-section on cleaved device edges. It is most often used as a tool to diagnose the size and position of the depletion region relative to the position of the layer structure.

More common techniques (such as Polaron or SIMS) do not readily reveal information on depletion region position and width (particularly as to the *activated* dopants). The presence of undesirable contaminants or reactor malfunction can often be deduced from EBIC images. Many types of electro-optic phase modulators are also critically dependent on junction positioning. One enhancement greatly increases the amount of useful information. It makes EBIC scans with forward and reverse bias applied, examining how the depletion region expands to both the n- and p-sides when reverse bias is increased. Figure 21 illustrates an example, where the GRINSCH waveguide has light p-doping. The depletion region starts out centered below the quantum well, but expands upward through the waveguide region as a reverse bias is applied. The EBIC technique with varying bias levels was a key to making an effective modulator with this structure [30]

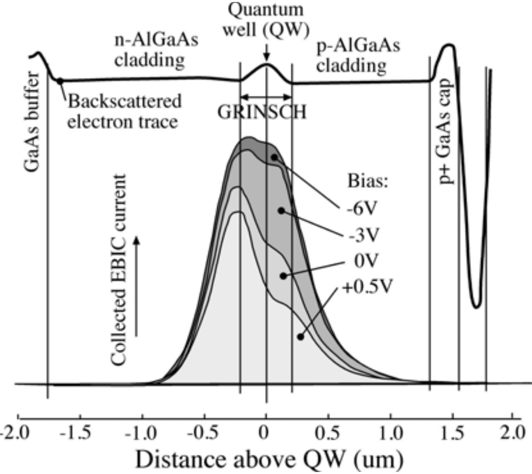


Fig 21: EBIC linescans superimposed on backscattered electron trace show position and width of depletion region for a variety of bias conditions.

As discussed earlier, most laser and LED failure originates and propagates in the gain layer Thus the "height" information is usually not useful. Plan-view imaging is generally more revealing, as discussed in section 5.1.

Plan-view EBIC helps greatly in analyzing junction-side-up edge-emitting lasers. If the diode is mounted junction-side down (for improved heat-sinking), it must first be taken off the heat sink to allow access to the junction. All solder must be removed. The laser geometry is usually a broad stripe (up to 100 µm across) for high-power lasers, or a narrow stripe (3-5 µm wide) for index-guided lasers. As discussed in the earlier section that dealt with the physics of failure with $\langle 100 \rangle$ DLDs, DLDs that extend beyond the current injection region often lead to the source of DLD nucleation. This can easily be verified if a crack, scratch or epi pit is observed at the origin of the DLD, as seen in figure 22, which is a top-view of the device drawn in figure 4.

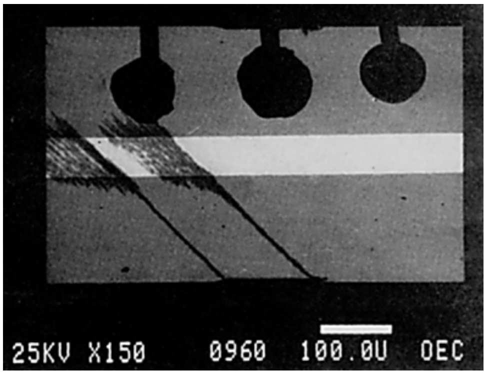


Fig 22: Plan view EBIC shows a top view of a stripe after aging. Defects originate at cleaving cracks at the bottom, and travel up and to the left. Once the defects cross the stripe, they quickly cause catastrophic failure. A drawing of this laser is also shown in figure 4 (after [4]).

5.3.3 Cathodoluminescence

Cathodoluminescence (CL) is light generated in an electron microscope by an impinging electron beam. Unlike EBIC, CL emission emerges not only from the depletion region, but also from all layers that have a direct bandgap. Where dislocations exist, non-radiative carrier recombination greatly reduces luminous efficiency nearby. Unlike most other characterization techniques, electron beam current can be uniformly injected throughout the sample, regardless of doping or composition. And unlike many other techniques, no electrical connectivity to the sample is required. However, there still needs to be a path for light to escape. Also, many optoelectronic devices use cladding material without a direct bandgap, so no luminescence would be seen in the cladding layers of such devices.

5.4 Transmission Electron Microscopy

5.4.1 Focused Ion Beam preparation of samples for Transmission Electron Microscopy

Transmission electron microscopy (TEM) has always been essential to a real understanding of laser degradation. The technique has now been used for nearly four decades to study semiconductor laser degradation [17]. Physical features only hundredths of a micron across can easily be resolved and identified precisely as precipitates, microloops, or some other structure. The identification helps greatly in eliminating the source of degradation. Since TEM requires a sample 1 μm or thinner, sample preparation is destructive by nature. It requires grinding or ablating all material beyond the section being studied.

When the early TEMs were produced, they were “lucky results” requiring extremely talented technicians and heroic sample preparation. More recently, the use of focused ion beam machines (FIBs) has made sample preparation satisfyingly replicable. Now we can be confident enough to section field returns since the technique has a success rate of better than 90%. In FIB, a gallium beam cuts away the undesired material. Other chapters in this volume that deal with TEM discuss further recommendations about FIB TEM sample preparation. Unless confident of the location of damage, and the uniformity of the damage, the normal flow would be to perform a plan view TEM taking in all of the possible nucleation and DLD growth area. Then cross sections can be taken out of the plan view section to isolate the type of dislocation that started the DLD. Some excellent examples of this type of work were published by its originator, Terry Stark [31].

5.4.2 TEM on cross sectioned samples

Cross-sectional TEM can be useful where the layer of origin is not known. However, the investigator needs a method of failure localization to specify where the section should be taken. Otherwise the cross section is unlikely to include the point of origin for the failure network. (Recall that all but a 1 μm or thinner section needs to be cut away to make an acceptable TEM sample, and thus more than 98% of the active region will be destroyed in the process. This is okay if a representative sample is all that’s needed, but not okay if a unique feature needs to be examined.)

One attractive feature associated with the use of a FIB is that a wedge can be plucked from the device, and a TEM sample made from the wedge This saves the time that would otherwise be spent in sawing and grinding using conventional sample preparation techniques. TEMs of aged VCSELs are shown below in figure 23. The DLDs shown in Fig. 23 are typical of what is seen in most VCSEL field failures.

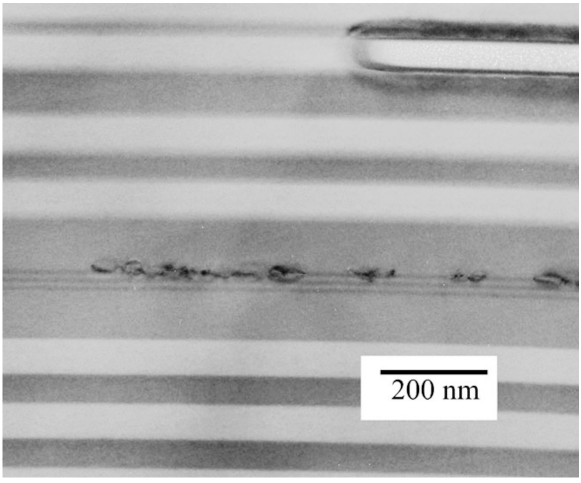


Fig 23: Cross sectional TEM image of failed oxide VCSEL. Oxide aperture appears upper right; three distinct (thin) quantum wells can be seen running the width of the device, with $\langle 100 \rangle$ DLDs centered in the upper quantum well.

5.4.3 TEM of plan-view samples

As mentioned earlier, if the layer in which the degradation started is known (usually the active layer for DLD networks), plan-view imaging is much more useful than cross sectional imaging. A plan-view sample shows the entire active region in the pumped area, rather than eliminating most of it. Note that a plan view sample probably has a practical limit of 40 x 40 μm (at least for routine samples), both due to FIB tool removal rates and the brittleness of thin semiconductor sample. In some cases, the source of degradation can be further away, and this limited size can be a significant drawback. In this case, the backside TIVA technique can inspect the entire die dimensions (usually roughly 250 x 350 μm).

Plan-view imaging was first used on VCSELs by Dr. David Mathes, who studied dozens of degraded VCSELs as part of his Ph.D. dissertation, and continued to work on them in industry [32,33]. Service houses now routinely provide plan-view TEM. It can resolve some VCSEL field failures where

other techniques cannot conclusively show the cause of failure. A plan-view image of an extensive $\langle 100 \rangle$ DLD network is shown in figure 24 below, which was a typical field failure.

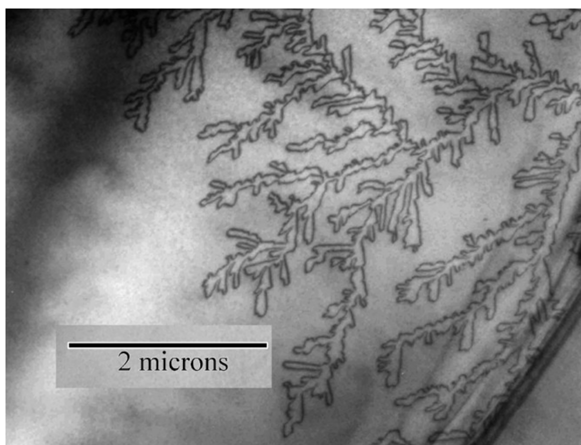


Fig 24: Plan-view TEM image of failed VCSEL, showing $\langle 100 \rangle$ DLD network from the top.

In principle, plan-view TEM should provide conclusive results. However, analysts recognize that interpreting the images is often difficult and sometimes inconclusive. Sample bending ("potato-chipping") provides undesired dark fringes that can easily be confused with DLDs. Extensive DLD networks often obscure the nucleation points. Analysts often overlook subtle features, such as moiré fringes, or edge darkening near oxide apertures, since they don't know what to look for. To establish a base for knowledge, it is often necessary to first create of a library of samples degraded by known causes. Even when this has been done, a significant fraction of devices studied show no clear cause of DLD origin.

5.5 Scanning laser techniques

5.5.1 Description of laser scanning microscope (LSM)

Laser scanning microscopes (LSMs) have been used for more than twenty years to evaluate epitaxial uniformity and defect concentration, on LED epitaxy. They have also been used a failure analysis tools on silicon ICs, where the backside of the IC has been polished or milled away. However, until recently, laser scanning microscopy had not been used as a routine FA tool in optoelectronics failure analysis. By combining the stimulus from the laser beam with a current or voltage signal collected through the device contacts, we can obtain maps of defect networks, often with little or no sample preparation.

While many excitation and biasing schemes are available, they fall into two broad categories. The first category uses a scanning laser with a photon energy greater than the bandgap of the device being studied. Here the beam creates electron-hole pairs, which can be collected in a manner similar to EBIC.

Most recently, we have concentrated our efforts on the second mode, thermally induced voltage alteration (TIVA). Here, the beam energy is less than the bandgap of the semiconductor. The beam merely heats the device as it traces a raster pattern across the sample.

5.5.2 Thermally-induced voltage alteration (TIVA)

TIVA has been widely used to detect failures in integrated circuits (ICs) . In such failure analysis, the backside of the package is polished off, as well as most of the substrate. Although preparing the sample can be troublesome, TIVA allows the analyst to see directly through to the junction, without the contact and gate metallization blocking the beam.

TIVA has been used for both topside and backside inspection of failed VCSELs. As with EBIC, it works best with defects in the depletion region (i.e., in or near the active region). TIVA is less sensitive to degradation in the middle of a VCSEL mirror (microns above or below the active region), where little change in the junction characteristics would be expected. TIVA shows devices without defects, or ones with no degradation in the active region, as totally featureless (except for bright areas at the etched edges). However, DLDs show up as bright areas in TIVA due to the reduced voltage for a fixed current.

In backside TIVA, the DLDs can often be followed to the point of origin. In many cases TIVA can identify the underlying cause of the failure. As with ICs, using backside sample preparation helps with VCSELs, since it allows clear access without the contact metal's blocking the beam. This is shown in Fig 25 below. Similar work has been published elsewhere [31, 34]. The rightmost image shows the DLD network, which originated from a defect at the bottom right, and traveled toward the center until it killed the laser. The image in the center is of another laser from the same array, where some sort of damage or delamination was present at the upper left edge, although it did not propagate in toward the center during aging.

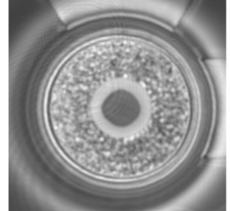
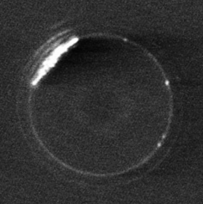
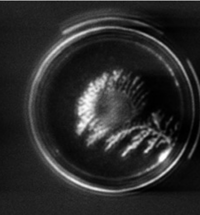
  

Fig 25: Photos of the backside of a mesa VCSELs taken by a laser scanning microscope (LSM) using a 1.06um wavelength. Reflected light image (left) shows large mesa with small central area in the middle with oxide aperture where current is confined. Middle and right images show normal ring at mesa edge, plus evidence of damage in the bright areas.

While backside TIVA allows the analyst to see the full area of the die, preparing the sample takes time (typically about 4 man-hours per sample in our lab). The primary advantage of TIVA is lower cost (\$500 vs. \$2,000 for plan view TEM), quicker sample preparation, and larger image area. However,

in many cases, the analyst will want to submit the TIVA die for TEM later for in-depth inspection of the nucleation area. Most commonly, the analyst might want to answer the question of whether the defect originated in the middle (perhaps from ESD or EOS), or whether it originated from the edge, perhaps because of mesa edge or etch hole damage.

6 Common Steps to Assure Product Reliability of Optoelectronic Components

Optoelectronic devices exist at the bottom of the supply chain. As a small component that can be mass produced (perhaps 20,000 die per wafer), lasers might sell for just a few dollars each, or much less than a dollar in consumer applications like compact disc (CD) players. In spite of their low cost and seeming simplicity, the lasers are one of the most difficult components to produce reliably, given the extreme stresses they operate under with normal bias. The lasers enable data switching equipment whose unit price often ranges from thousands of dollars to over a million dollars.

In these applications, failure can cause expensive downtime. Repairs usually cost thousands of dollars. If failures exceed perhaps ten or a hundred parts per million (ppm), they can cause an equipment supplier to be removed from the approved vendor list. Optoelectronic suppliers thus face enormous potential liabilities if they supply potentially defective parts. To prevent such defective parts from reaching the customer, companies routinely spend more to screen and test parts for reliability than they spend actually making the parts. Cost-effective FA is also particularly important, since it is expensive to fix problems unless they can be quickly and positively identified.

6.1 Design

As discussed above in section 1.2, selecting an appropriate materials system is critical to making reliable parts. Some materials, like InGaAsP (used in 1.3 and 1.5 µm lasers) don't allow dark line defects to propagate. Even if significant ESD or mechanical damage occurs, continued operation won't cause any further degradation.

By contrast, GaAs-based lasers (such as most VCSELs or CD-lasers) would be killed by fast-growing dark-line defects if the die is damaged. In spite of this serious shortcoming, the cost and simplicity advantages to GaAs-based lasers are significant, and their weaknesses have proved manageable in many applications. Another good example of material system choice comes from contrasting two materials used for blue and green emitters. GaN blue LEDs are becoming commonplace. They seem virtually indestructible, with most types of threading dislocations being non-propagating [16, 40]. ZnSe-based green LEDs by contrast contain a soft material that generates defects within hours. Consequently, researchers have abandoned II-V materials like Zn-Se.

Most VCSELs are used as direct-modulated lasers, with the drive current switched between "nearly off" and what is often a very high "on state" current. With this kind of modulation, nearly every parameter (rise time, fall time, eye opening, link margin, etc.) can be improved by simply increasing the drive current. However, higher current makes the part less reliable, owing both to the higher current density, as well as to the increase in junction temperature.

Thermal design and keeping junction temperatures low are important elements in electronics reliability. They are especially important in the design of semiconductor lasers, which already operate with much higher minority carrier populations than do most other semiconductor components.

Many cost reduction efforts involve replacing metal (*e.g.,* in the ferrule which holds the laser) with much cheaper plastic. The plastic, however, may have a thermal conductivity three orders of magnitude lower than the metal it replaced. Fortunately, thermal modeling, and a focus on thermal properties, has become more common in the last few years.

Mechanical design – particularly minimizing mechanical stresses from packaging – is another key element in ensuring the reliability of electronic circuit boards. Packaging-related failures provide a significant source of problems in many optoelectronic devices as well. One common design optimization tries to choose packaging materials whose coefficient of thermal expansion (CTE) is as closely matched as possible to the CTE of the semiconductors used.

Another consideration is the choice of solder. Soft solders minimize stress on the parts, but tend to form voids and whiskers over time owing to their increased mobility [35]. Choosing materials that don't corrode when they contact one another is another important consideration.

Design for manufacturability is another key. We seek both low costs and the prevention of quality problems. While it is always easy for system designers to "cherry pick" lasers from the very center of the distribution, the practice should be avoided. Additional effort in other parts of the design (*e.g.,* designing the receiver IC to allow the widest possible range the range of photodetector capacitance) greatly widens the range of optoelectronic parts that can be used. Greater design tolerances also leaves buyers open to a wider range of competing suppliers, rather than the one or two suppliers willing to accept a customized "cherry picked" specification.

6.2 Basic Principles in Reliability Testing

Telcordia specification GR-468 covers reliability qualification of most active optoelectronic parts [36]. Most engineers regard the tests as a "necessary evil" – a hurdle that must be passed, but a test with little real value to detecting failure mechanisms that might be seen under normal use condition. My experience has been that a great deal of collected wisdom has gone into the creation of the specifications, and that the more one knows about reliability and failure analysis, the more respect one has for the authors of the Telcordia document. The real need is for a book entitled "*Why* we selected these tests for GR-468." Unfortunately, most of the

lessons originate in embarrassing mistakes that could have been avoided if the GR-468 tests had been used. Accordingly, such a volume is unlikely ever to be written. In addition, corporate qualification regimes, especially by the larger customers like Cisco, have extended the requirements, and addressed limitations in the GR-468, particularly the lack of rigor in wearout test requirements.

Reliability tests can generally be classified in three broad categories. The first deals with general aging tests, usually using high-temperature and increased electrical stress. The second contains mechanical tests that seek weak solder connections, wirebonds, open electrical vias, and the like. The mechanical tests normally rely on temperature cycling, or the use of low temperatures. The third category includes corrosion tests.

6.2.1 High-temperature aging tests

If a part is vulnerable to aging during normal use, the process of aging can normally be accelerated by high temperature (which enhances defect mobility) and increased electrical stress (i.e., the application of higher-than-normal current or voltage). An acceleration model shows how quickly the process of aging will proceed as currents and temperatures vary. The failures revealed by such tests – and particularly the *early* failures from the tests --- should be studied in detail to understand the dominant wearout failure mechanisms and maverick failure mechanisms that may be present. For lasers, threshold current and operating power are monitored over time. When they change by more than a pre-determined amount (e.g., a 20% increase in threshold current), we record the time failure occurred. The range of failure times are plotted on a lognormal plot, as shown below in figure 26, which used high temperatures and drive currents 3x the normal use condition. The mean times to failures (MTTFs) can then be computed, and put on an Arrhenius plot, along with data from other drive currents. We can use this to create a model and extrapolate what the expected time to failure under actual use conditions will be, as shown in figure 27.

Note that with the exception of high-power facet damage mentioned earlier, most laser failure mechanisms surprisingly show little or no dependence on laser power – either intra-cavity power or emitted power. Rather, the failure mechanisms appear to depend on current density and junction temperature. One revealing study used lasers of different cavity lengths, each of which had vastly different internal powers for the same current density. In spite of the variations in internal powers, they displayed almost no difference in degradation rates [38]. This relationship has also been observed with VCSELs as well, [39] and more recently in GaN blue lasers. [40]

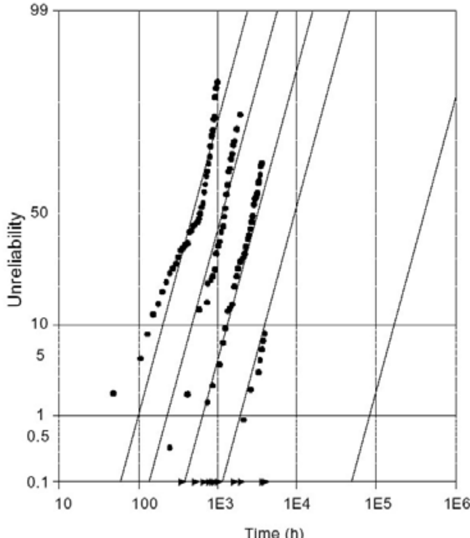


Fig 26: Lognormal plot of reliability versus time on lifetest for 4 different temperatures: 150°C, 130°C, 110°C, and 90°C (from left to right). Courtesy of Chris Helms at Emcore, after [37].

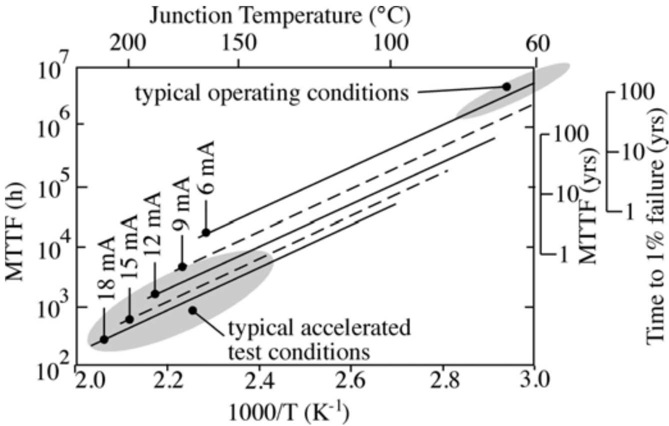


Fig. 27: Arrhenius plot showing how lifetime varies with temperature and drive current, using a typical VCSEL acceleration model, with 0.7eV activation energy, and degradation proportional to drive current squared, assuming a 12 μm round oxide VCSEL. Each line shows expected range of lifetimes from use condition (60℃) to a 150℃ oven temperature. The five lines shown cover the range of currents from use current (6mA) to the maximum used in our tests (18mA).

Laboratory testing is not omnipotent. While it can detect fundamental design flaws, low level maverick problems (*i.e.*, <100ppm) often remain undetected, even with samples as large as 10,000.

Many fiber optic modules are also strictly limited in how much acceleration can be applied, since there is no ability to increase electrical stress from the outside (the laser driver has

fixed drive current), and very limited temperature range. It is not unusual to observe a maximum acceleration factor of only 10-times on fiber optic modules. Thus testing for a wearout lifetime of ten or more years cannot be done on a timely basis. Furthermore, unlike IC testing, many fiber optic parts cost nearly $1000 each, making mass-volume testing an expensive task.

One approach that avoids the cost and acceleration factor restrictions tests individually the active components (i.e., IC, photodetector, and laser) used in the module, with special attention to the laser. Here, acceleration factors of one thousand times or more are obtainable, and samples of many thousand items can be economically tested.

Another limitation involves the inevitable delays in initial detection of a quality-control problem. Significant populations of parts may be fielded before the problem is detected, requiring expensive recall and repair efforts. That makes trying to eliminate potential problems in advance (e.g., by using FMEA) a worthwhile approach.

Finally, note the importance of ascertaining the current-and-temperature combinations at which the lasers will actually be used. It is not uncommon to see citations of lifetime at 25°C, in circumstances where the laser will actually be used at temperatures of 60 or 70°C. The device's lifetime might actually be more than 20 times shorter at the actual operating temperature than at 25°C. Lasers can also be misrepresented by citing the mean time to failure (MTTF), rather than the more relevant time to 1% or 0.1% failure. Other electronic assemblies often can be modeled with a constant failure rate, and thus the time to 0.5% failure is one-hundredth the time to 50% failure. This is very inaccurate with VCSELs, where the failure rate abruptly increases once the wearout lifetime is approached. Using "customer-centered" criteria in calculating likely lifetime is always preferable [41].

6.2.2 Mechanical tests

Solder joint failures are the most common failure mechanisms observed in electronics. They can occur quickly if inadequate amounts of solder are dispensed or inadequate wetting takes place. But even good solder joints will eventually fail due to fatigue, as shown in figure 28. The most popular method for testing involves temperature cycling. One thousand cycles from –40°C to +100°C would be a popular test condition for fiber optic transceivers. Note that subcomponents such as the laser or photodiode rarely fail these tests. If problems are encountered, they are normally with the printed circuits or ICs (e.g., pulled wirebond, cracked capacitor, open via). Depending on how quickly strain relaxes, rapid changes in temperature ("thermal shock") may be more or less stressful than thermal cycling.

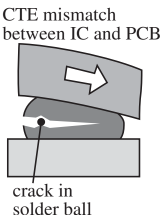


Fig 28: Solder bump cracks from metal fatigue induced by cycling of high lateral stress forces.

Low-temperature operating life can also create a high mechanical stress condition. Encapsulants are normally poured and cured at high temperatures, so that mechanical stress becomes progressively greater as temperature falls.

For larger assemblies, "shake and bake" – a combination of vibration and temperature cycling finds weak solder joints quite effectively. However, our experience shows that most chambers are oriented towards large systems, and small optoelectronic parts do not have sufficient mass or flexibility to be usefully stressed in vibrational testing.

6.2.3 Corrosion tests

The third type of reliability testing looks for corrosion. While hermetic packaging can keep atmospheric moisture out, it tends to be quite costly when compared to popular plastic packaging. Plastic packaging, by contrast, is readily permeated by atmospheric water vapor. The water, when combined with a bias, can lead to chemical reactions that change the circuits in ways that eventually cause failure.

Testing in chambers with 85ºC and 85% relative humidity is popular, with a typical requirement of 1000-2000 hours of test time. Most engineers who lack experience in reliability testing criticize these magnitudes. They say, "My part won't ever to be subjected to such high temperatures and humidities. Therefore, such testing is unrealistic."

By contrast, I have seen parts released to production in spite of failing after a hundred hours of 85/85, only to result in significant numbers of corrosion-related field failures less than a year later, and the product being pulled off the market.

Fortunately, most large customers are sufficiently risk-averse to insist on passing this test, regardless of whether the device engineers understand its relevance. Extensive testing on Si and GaAs ICs over a range of temperatures and humidities has shown that 85°C/85%RH accelerates corrosion by about 100 times faster than the worst-case conditions of outdoor use. The corrosion continues to occur at lower temperatures and humidities, but obviously proceeds much more slowly. A sample acceleration model for oxide VCSELs can be found in ref [42], with more general information on GaAs IC corrosion contained in [43].

Both biased and unbiased testing are usually done. Biased testing reveals flaws such as electrolytic corrosion, while unbiased testing avoids self-heating effects that lower the relative humidity at the surface.

The "damp heat" or "85/85 test" is a good test for qualification. When product improvement or product screening is performed, a faster test is desirable. Here, "pressure pot" testing can help. By using a pressurized chamber, temperatures above 100°C can be used, and another factor of 10 times in acceleration can be obtained. For example, 130C, 85% R.H., 2.3 atmg is a common qualification condition for IC package corrosion, and works well on many optoelectronic devices as well. Then, only about one hundred hours (a few days) gets the same results that would otherwise require over a month of testing at 85/85. Note, however, that unlike 85/85 testing, pressure pot testing does tend to cause embrittlement in most plastic parts we have studied. They often break like chalk, or they fog owing to chemical changes. These are essentially experimental artifacts, since they won't be observed at 85/85 or lower temperatures. However, this limitation is often not so serious as to make pressure pot testing useless.

While the tests described here are generally intended to be non-condensing, the power temperature cycle often subjects cold parts to hot humid air, which leads to condensation on the part while it is biased. Similar real-world condensation can cause failure in the field. Thus this test uncovers possible weaknesses in this aspect of the product design.

A variation of corrosion testing uses special corrosive gas elements found in polluted cities. Some installations in large industrial cities in low-income countries have shown premature failures not observed elsewhere. Special Telcordia tests with a pollution-simulating gas try to detect such problems.

6.2.4 Electrical overstress testing

Customers want to know if they need to take special precautions to protect parts from ESD damage. Thus it is important to understand how robust the product is to withstanding electrostatic discharge.

Their precautions can be divided into two general categories: (1) simulations of handling of subcomponents by assemblers, and (2) simulations of actual use by the final customer. For the first category, a human body model (HBM) simulates hand insertion of parts; machine model (MM) simulates robotic assembly. The cumulative discharge model (CDM) simulates charged parts discharging on a contact surface. In the second ("customer use") category, air discharge wands can be used to simulate an unprotected end user touching the outside of the finished product to see if finger contact would cause a loss of data or damage to the product.

6.3 Screening

If reliability cannot be "guaranteed by design", it may be necessary to add steps to the manufacturing process to ensure that only reliable parts are shipped, and to screen out potentially weak parts. Achievement of these goals depends on the frequency of the deviation. It may affect all parts from a given lot, all devices on a wafer, or it may differ even among neighboring dies. The particular findings determine an effective strategy to prevent defective parts from being shipped to the customers.

6.3.1 Destructive versus non-destructive testing

If parts with infant failure mechanisms are present in significant quantities, it is necessary to test 100% of parts with a burn-in. The burn-in is normally greatly accelerated to keep it short (and economical), but it must be sufficiently long to identify and reject the large majority of weak parts. [44]

Burn-in is limited in scope by the fact it must be non-destructive. Therefore, an additional wafer sample must undergo destructive testing to ensure that the finished parts have adequate wearout lifetime. [41]

6.3.2 Screening a growth lot or a fabrication lot

We can imagine a case where all parts from a given epitaxial growth lot might have oxygen contamination, or all parts from a wafer fabrication lot might have contaminated metallization. In such cases, only a representative sample from each lot requires testing. A sample size of fifteen parts is usually considered the smallest acceptable sample for measuring the mean and the deviation of laser lifetime. Larger samples are often tested to allow non-uniform problems to be better detected. For low-cost parts, a sample size of 30 – 40 lasers is common. The wearout sample tests are usually destructive tests. This test subjects the samples to aging roughly equivalent to the mission life of the part, using high acceleration factors (usually 1000x or higher, to allow a pass/fail decision to be made in a week or so.). Non-destructive testing can generally find only those products that fall an order of magnitude short of meeting mission life requirements, and would have failed in the first year or two or normal operation.

6.3.3 Wafer Level

Laser products are immature from a manufacturing basis when compared to ICs, and significant wafer-to-wafer variations still occur within a lot, although they are not common. When these variations can occur, every wafer needs a sample selected. It is important to pick the sample from across the wafer, since variations may occur from one edge to the other.

6.3.4 Part Level

Weak parts can still occur anywhere, in isolation from their neighbors. Therefore 100% burn-in is an important aspect of the screening process for lasers. Aging for the equivalent of several months to a year of normal use is routine. Then the devices are checked to quantify how much change in the device characteristics has occurred. All items should also be visually inspected so that the ones with mechanical damage will not be shipped to customers.

7 Recommended Reading

7.1 Books on reliability

Franklin Nash. Dr. Nash did the original qualification of the undersea fiber optic links. His book [9] contains many examples specific to semiconductor lasers. Its chapter 8 presents an example of an analysis. The book is also excellent in describing where various failure distributions are appropriate, as well as in its treatment of limitations of reliability testing to *prove* high reliability.

7.2 Books on in-plane laser and LED failure analysis

Artech House has published two excellent books. While they are somewhat out of date, they have a very comprehensive list of failure mechanisms observed in various types of LEDs and lasers.

Mitsuo Fukuda's book (ref [10]) includes a primer on the physics of operation, and reliability testing, while Osama Ueda's treatise ("Reliability and Degradation of III-V Optoelectronic Devices, 1996) is more highly focused on pure materials science and failure analysis.

7.3 Acknowledgements:

I would like to thank my coworkers (too numerous to mention individually) at McDonnell Douglas, UCSB, Agilent Technologies, Emcore, and Finisar for their assistance in this work.

8 References

1. S. Yamakoshi, M. Abe, O. Wada S. Komiya, T. Sakurai, "Reliability of high radiance InGaAsP/InP LEDs operating in the 1.2-1.3 µm wavelength," *IEEE J Quantum Electron.,* vol 17, #2, pp. 167-173, (1981).
2. H. Wang, A. A. Hopgood, and G. I. Ng, "Analysis of dark-line defect growth suppression in InxGa1-xAs/GaAs strained heterostructures.," *J Appl Phys*, vol. 81, pp. 3117-3123, (1997).
3. Robert W. Herrick, "Degradation in Vertical Cavity Lasers," a Ph.D. dissertation, available from University Microfilms, Ann Arbor Michigan, 1997.
4. R. G. Waters, "Diode laser degradation mechanisms: a review," Progress in Quantum Electronics, vol. 15, pp. 153-74, (1991).
5. K. Kondo, O. Ueda, S. Isozumi, S. Yamakoshi, K. Akita, and T. Kotani, "Positive feedback model of defect formation in gradually degraded GaAlAs light emitting devices," IEEE Transactions on Electron Devices, vol. ED-30, pp. 321-326, (1983).
6. O Ueda, "Degradation of III-V Optoelectronic Devices," J. Electrochem. Soc., Vol. 135, pp.11C-22C, (1988).
7. J.W. Tomm, A. Barwolff, A. Jaeger, T. Elsasser, J. Bollmann, W.T. Masselink, A. Gerhardt, and J. Donecker, "Deep level spectroscopy of high-power laser diode arrays," J Appl Phys, Vol. 84, #3, pp. 1325-1332, (1998).
8. P.M. Petroff, "Device Degradation and Recombination Enhanced Defect Processes in III-V Semiconductors," Semiconductors and Insulators, Vol. 5, pp. 307-319, (1983).
9. Nash, Franklin R., "Estimating Device Reliability," Kulwer Academic Publishers (Boston, 1993).
10. Mitsuo Fukuda, "Reliability and Degradation of Semiconductor Lasers," published by Artech House, 1991, pp. 128-136
11. M. Fukuda, M. Okayasu, J. Temmyo, J. Nakand, "Degradation behavior of 0.98-µm strained quantum well InGaAs/AlGaAs lasers under high-power operation", IEEE J. Quantum Electron., V. 30, pp. 471-476, (1994).
12. M. Gasser and E. E. Latta, "Method for mirror passivation of semiconductor laser diodes (U.S. patent #5,063,173)," . USA: International Business Machines Corporation, Armonk, N.Y., 1991.
13. M. Okayasu, M. Fukuda, T. Takeshita, S. Uehara, "Stable operation (over 5000 h) of high-power 0.98-µm InGaAs-GaAs strained quantum well ridge waveguide lasers for pumping Er3+-doped fiber amplifiers," IEEE Photon. Technol. Lett, V. 2, pp. 689-691, (1990).
14. W. J. Fritz, L. B. Bauer, and C. S. Miller, "Analysis of aluminum gallium arsenide laser diodes failing due to nonradiative regions behind the facets," Proc. of the Intl. Reliability Physics Symposium 1989, pp. 59-64, (1989).
15. R.W. Herrick and P.M. Petroff, "Gradual degradation in 850-nm vertical-cavity surface-emitting lasers," IEEE J. Quantum Electronics, Vol.34, no.10, pp. 1963-9 (1998).
16. S. Tomiya, . Hino, T. Miyajima, O. Goto, and M. Ikeda, "Defects and degradation of nitride-based laser diodes", Proc. SPIE, Vol. 6133, 613308 (2006)Tomiya, SPIE 2006
17. P.M. Petroff, W. D. Johnston, Jr., and R. L. Hartman, "Nature of optically induced defects in Ga1-xAlxAs-GaAs double-heterojunction laser structures," Appl. Phys. Lett., vol. 25, pp. 226-8, (1974).
18. H.Q. Jia, H. Chen, W.C. Wang, W.X. Wang, W. Li, Q. Huang, J. Zhou, and Q.K. Xue, "Improved thermal stability of wet-oxidized AlAs," Appl. Phys. Lett., Vol. 80, #6, pp. 974-976, (2002).
19. Suning Xie, Robert W. Herrick, Danielle Chamberlin, S.J. Rosner, Scott McHugo, Grant Girolami, Myrna Mayone, Seongsin Kim, and Wilson Widjaja, "Failure Mode Analysis of Oxide VCSELs in High Humidity and High Temperature", Journal of Lightwave Technology, Vol 21, pp. 1013-1019, (2003).
20. C. Helms, W. Luo, I. Aeby, R.W. Herrick, A.Yuen, "Oxide VCSEL Reliability at Emcore", Proc of SPIE, vol. 5364, p. 183 (2004).
21. ESD / EOS Symposium, see www.esda.org for further information.
22. G. Theodore Dangelmayer ESD, "ESD Program Management," second edition, Springer, (1999).
23. J. Krueger, R. Sabharwal, S. McHugo, K. Nguyen, N.X. Tan, N. Janda, M. Mayonte, M. Heidecker, D. Eastley, M. Keever, and C. Kocot, "Studies of ESD-related failure patterns of Agilent oxide VCSELs," Proc of SPIE, vol. 4994, (2003).

24. D. Mathes, J. Guenter, B. Hawkins, B. Hawthorne, C. Johnson, "An atlas of ESD failure signatures in Vertical Cavity Surface Emitting Lasers," *Proc. 31st ISTFA*, p. 336, (2005).
25. J. Guenter, B. Hawkins, R. Hawthorne, R. Johnson, G. Landry, K. Wade, "More VCSELs at Finisar," Proc of SPIE, Vol. 7229, paper 772905 (2009).
26. M. Meneghini, A. Tazzoli, G. Mura, G. Meneghesso, and E. Zanoni, "A Review on the Physical Mechanisms That Limit the Reliability of GaN-Based LEDs," IEEE Trans. Electron Dev., Vol. 57, p. 108 (2010).
27. D.L. Barton and M. Osinski, "Life Tests and Failure Mechanisms of GaN/AlGaN/InGaN Light Emitting Diodes," Proc. Of CLEO-1998, pp. 440-441, (1998).
28. R.B. Comizzoli, J.W. Osenbach, G.R. Crane, G.A. Peins D.J. Siconolfi, O.G. Lorimor, and C.-C. Chang, "Failure Mechanism of Avalanche Photodiodes in the Presence of Water Vapor," J. Lightwave Technol., Vol. 19, #2, pp. 252-265, (2001).
29. D. K. McElfresh, L. D. Lopez, and D. Vacar, "Reverse-bias emission sheds light on the failure mechanism of degraded VCSELs", J. Appl. Phys. 99, 123113 (2006).
30. Kun H. No, Richard J. Blackwell, Robert W. Herrick, and Joseph L. Levy, "Monolithic integration of an amplifier and a phase modulator fabricated in a GRINSCH-SQW structure by placing the junction below the quantum well," IEEE Photonics Technol. Lett., Vol. 5, #9, pp. 990-993, (1993).
31. T. J. Stark, P. E. Russell, and C. Nevers, "3-D Defect Characterization Using Plan View and Cross-sectional TEM/STEM Analysis", Proc. of 31st ISTFA, p. 344, (2005).
32. David T. Mathes, "Materials Issues for VCSEL Operation and Reliability," a Ph.D. Dissertation, published University of Virginia, Dept. of Materials Science and Engineering, 2002.
33. D. Mathes, J. Guenter, J. Tatum, R. Johnson, B. Hawkins, C. Johnson, and B. Hawthorne, AOC moving forward: the impact of materials behavior," Proc. SPIE 6132, 613203 (2006).
34. C. Lei, N. Li, C. Xie, R. Carson, X. Sun, W. Luo, L. Zhao, C. Helms, D. Jensen, and C. Liu, "Emcore VCSEL failure mechanism and resolution", Proc. SPIE, Vol. 7615, 761504 (2010).
35. M. Fukuda, O. Fujita and G. Iwane, IEEE Trans Comp Hybrids Manufactur. Technol., Vol CHMT-7, p. 202, (1984).
36. "Generic Reliability Assurance Requirements for Optoelectronic Devices Used in Telecommunications Equipment", Document GR-468-CORE, published December 1998, available from Telcordia.com
37. I. Aeby, D. Collins, B. Gibson, C.J. Helms, H.Q. Hou, W. Lou, D.J. Bossert, and C.X. Wang, "Highly reliable oxide VCSELs for datacom applications", Proc of SPIE, Vol. 4994, pp. 152-161, (2003).
38. R.G. Waters and R.K. Bertaska, "Degradation phenomenology in (Al)GaAs quantum well lasers," Appl. Phys. Lett., vol. 52, pp. 179-181, (1988).
39. J. K. Guenter, R. A. Hawthorne, D. N. Granville, M. K. Hibbs-Brenner, and R. A. Morgan, "Reliability of proton-implanted VCSELs for data communications," Proc. of SPIE (Fabrication, Testing, and Reliability of Semiconductor Lasers), vol. 2683, pp. 102-13, (1996).
40. L. Marona et al., "Comprehensive study of the reliability of InGaN-based laser diodes," Proc of SPIE, Vol 6485, 648504, (2007).
41. R.W. Herrick, "Oxide VCSEL reliability qualification at Agilent Technologies", Proc of SPIE, vol. 4649, pp. 130-141, (2002).
42. R.W. Herrick, "Reliability of Fiber Optic Datacom Modules at Agilent Technologies," Proceedings of the 52nd annual ECTC conference, pp. 532-539 (2002).
43. Pollino, Emiliano, Microelectronic Reliability, Vol. II, pp. 374-387, Artech House, (Norwich, MA, 1989).
44. F.E. Jensen and N.E. Petersen, "Burn-In: An Engineering Approach to the Design and Analysis of Burn-In Procedures", publ by J.E. Wiley, 1982

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Solar Photovoltaic Module Failure Analysis

G. B. Alers
Department of Physics
University of California, Santa Cruz
Santa Cruz, CA 95064 USA

Abstract

Post-mortem analysis of photovoltaic modules that have degraded performance is essential for improving the long term durability of solar energy. A general procedure for analyzing a failed module might include (a) electrical characterization followed by (b) visual inspection (c) thermal imaging and (c) electroluminescence imaging. These inspections methods help to identify the physical location of a failure site. Once the site is identified then additional microanalysis can be performed if necessary.

Introduction

Power generation from solar photovoltaics is increasing at a rapid pace. The projected lifetime of a module is critical for determining the annual cost of a photovoltaic installation. Initial capital investment can be spread through the lifetime of the system. A longer projected lifetime has a direct impact on lowering the annual cost the of the installation. Making lifetime projections requires a detailed understanding of failure mechanism for solar panels and components. Therefore, failure analysis is critical to reducing the long term cost of solar energy.

Solar photovoltaic panels are manufactured with the series connection of many cells made of a variety of materials and assembled into an array. Failure analysis techniques of full modules need to be general enough to adapt to all forms of modules. Crystalline or polycrystalline Si panels use multiple cells assembled onto a glass superstrate to form a module. Thin film photovoltaic materials including amorphous Si, CdTe, Cu(In, Ga)Se (CIGS) can be deposited over a large area and patterned into cells or fabricated into individual cells and then mounted to a glass superstrate much the same way that Si cells are assembled. Concentrator photovoltaics use high efficiency multi-junction solar cells and light concentrated with either a lens or reflector. Organic or thin film photovoltaics that are processed as a roll and then sealed in plastic to form flexible photovoltaic sheets.

Each different type of photovoltaic technology will have a unique set of failure modes related to both packaging failures and cell failures. Crystalline Si photovoltaics tend to be dominated by packaging related failures because the basic Si photovoltaic material is very well developed. NREL has compiled a list of common failure modes in photovoltaics which is summarized in Table 1. [1]

Common Failure Mode for crystalline Si PV modules
Cracked cells (bonding processes, strain, etc.)
Solder joint or gridline interface failure
Reduced adhesion and corrosion / delamination
Slow degradation of ISC
Fatigue of ribbon interconnect
Junction box failure (poor solder joints, arcing, etc.)
Busbar adhesion degradation, electrical contact, etc.
Glass edge damage of frameless modules
Light-induced cell degradation
Effect of glass on encapsulant performance
Front surface soiling
Mechanical failure of glass-glass laminates

Table 1: Common failure modes for crystalline Si PV modules.

Thin film photovoltaic failures tend to be dominated by cell level failures of the materials and interfaces such as thin film stability issues. These failures are summarized in Table 2 for thin film photovoltaics including CdTe and Cu(CIGS. [1]

Common Failure for thin film PV modules
Cell layer integrity- back contact stability
Electrochemical corrosion of SnO2:F
Fill-factor loss (series resistance and/ or recombination)
Busbar adhesion degradation, electrical contact, etc.
Shunt hot spots at scribe lines before and after stress
Weak diodes, hot spots, nonuniformities in absorber
Notable sensitivity of TCO to moisture
Moisture ingress failure of package
Cell–to-cell interconnect
Edge shunting

Table 2: Common failure modes for crystalline Si PV modules.

Electrical Tests for Failure

The first indication that a photovoltaic module has failed is a reduced power output under load. However, power output is just one of the parameters that characterize the performance of a photovoltaic. The full current - voltage characterization of the module provides the first glimpse at the failure mode. Figure 1(a) shows a typical IV curve for a full power solar module. Modules are comprised of many cells in series to increase the output voltage. If one of the cells in the series fails then the output current of the full module can fail. To avoid this situation bypass diodes are placed in parallel with a set of cell in a string so that current output can bypass the failed string and other cells will still contribute power. In this case, the output voltage of the module will drop proportionately to how many cells are in the string that failed

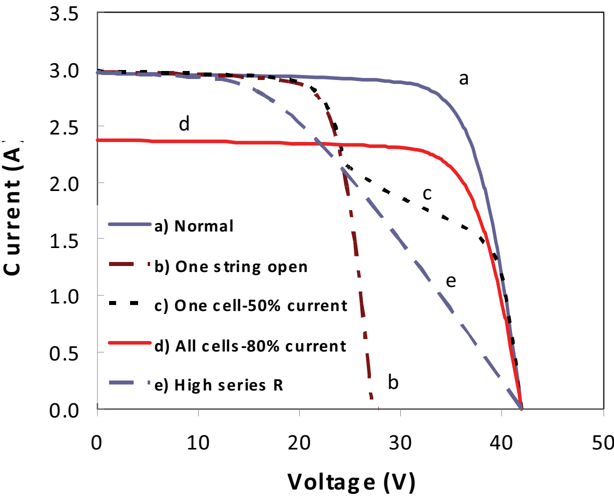


Figure 1: IV characteristics of common failure modes for a 72-cell module with 3 amps/cell of current output

resulting in an IV curve shown in Figure 1(b). This is the failure mode associated with an open circuit contact for a particular cell. If the performance of one cell starts to degrade due to poor contacts or materials degradation then this cell will start to limit the current output of the module. Figure 1(c) shows the IV characteristic with one cell at 50% output. If all cells degrade equally such as occurs with light induced degradation then the current output of the full module degrades as shown in Figure 1(d). Degradation of a contact can cause very high resistance (greater than a few ohms) then the series resistance of the module will increase and one would obtain an IV curve that looks like figure 1(c) with a non-vertical slope at I=0. A uniform degradation in the current output of all cells would result in a decrease in current output as illustrated in figure 1(d). If a few cells in the module have

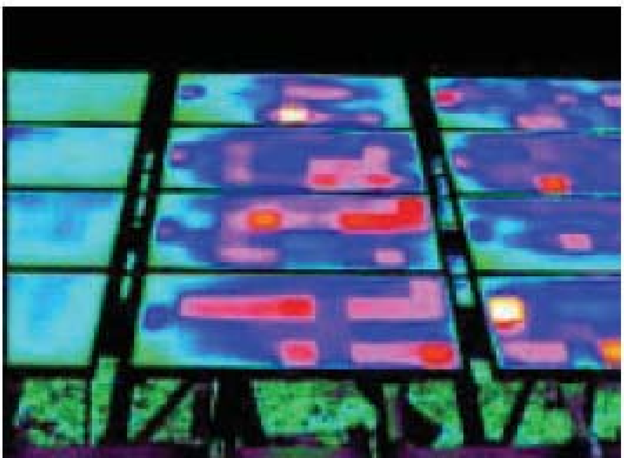


Figure 2: Thermal image of solar panels under full sun illumination and short circuit load conditions showing cells with high power dissipation.

decreased current output then short circuit current might not be reduced but output power with a load would be degraded as is shown in cases (c) and (e). This is the failure mode that would require imaging of the module to identify which how many cells are degraded and to what degree.

Thermal imaging of modules

Many different techniques have been developed to characterize solar cells and their defects. The most commonly used is thermal imaging with an camera sensitive to infrared black body emission in the 3-15µm range.[2] Poor contacts, shunted cells or local shorts will show additional joule heating relative to the neighboring environment and will appear a bright spots with thermal imaging. Figure 2 shows an example of an infrared image taken of a set of crystalline silicon solar cells mounted in an outdoor test field. Cells that output less than nominal current become reverse biased by the output of the other cells and start to behave as resistors in the circuit and dissipate heat. This heat is visible as the bright cells in the module. The current for heating of cells in a module can be from the sun in operation or with an external bias applied to the module.

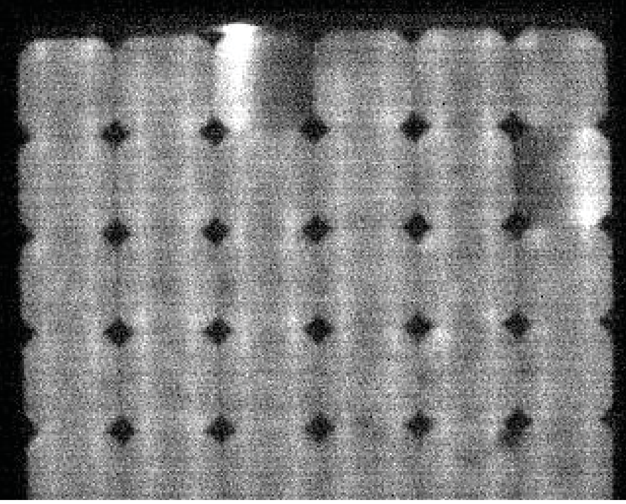


Figure 3: Thermal image with no illumination and external forward bias of the top fraction of a single crystalline solar panel with failed solder joints at two of the cells.

Forward bias thermal imaging

To maximize local joule heating of high resistance elements in the module a bias condition must be used with maximum current through the resistive elements. This can be achieved either under full illumination with short circuit loading or with an external power supply and minimal illumination. Poor solder contacts will have a higher than normal resistance and will exhibit joule heating as illustrated in the thermal image in Figure 3. Bright spots correspond to high infrared emission due to heating. A poor solder connection on a cell at the top of the module appears as a bright spot as is apparent in the top cell. Cells with no current

due to a poor contact or degradation appear dark in infrared images. Figure 3 also shows an example of a cell that has two parallel buss contact lines. One of these lines was broken and therefore no current passes through that portion of the cell (dark region) and additional current passes through the connect cell (bright region). Therefore the cell half with poor contact appears dark and the cell half with good contact appears brighter because of higher current density.

Reverse bias imaging and shunts

Regions of high series resistance are best identified with high current and forward biasing. Local shunting (or shorting) of a cell can occur from ohmic paths in parallel with the cell that bypasses current flow. The shunts can behave as resistors and dissipate heat equally in both directions. Alternatively, non-linear "weak diodes" can act as shunts in forward bias conditions only. Ohmic shunts are best observed in reverse bias conditions to achieve the greatest contrast in current density with minimal current flow through the cells and current flowing through the shunt. Weak diodes appear primarily in forward bias and appear as localized regions of higher than normal current density. Shunting can occur from localized materials degradation or defects and can result in a substantial loss in power for the module. Figure 5 shows an example of a weak diodes in a CdTe cell from McMahon et al. [3] The bright spots in the thermal image correspond to locations with exceptionally high current density under reverse bias. These locations correspond to excessive leakage current with a corresponding loss in power output. When the bright spot is removed by scribing out this area then the power efficiency increases from 6% to 9%

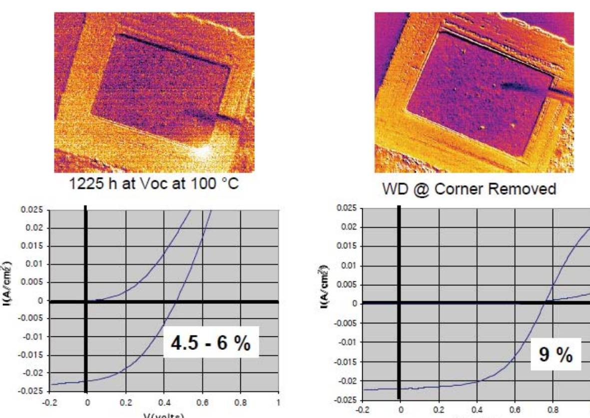
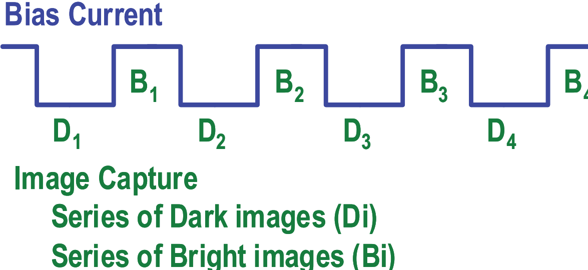


Figure 5: Thermal image and IV curve of a CdTe cell that contains a weak diode. When the weak diode is removed the power efficiency increases from 6% to 9%. From McMahon [4].

Lock-in thermal imaging

One fundamental problem with static infrared thermal imaging is thermal spreading. The heat generated from a localized hot spot will eventually reach thermal equilibrium



Final Image = Σ_i [Bi – Di]

Figure 6: Simple illustration of lock-in imaging technique

with its environment to minimize temperature gradients. Therefore, when a bias is first applied there is a narrow window of time in which defects heat locally but do not spread laterally losing contrast. One solution is lock-in thermography (LIT). LIT is a well-established failure analysis tool and is typically done with infrared (IR) cameras. [1] LIT techniques provide a wide rang of applications for qualitative and quantitative analysis of solar cell parameters. IR-LIT can produce thermal images with 10 μK temporal resolution and 5-10 μm spatial resolution. By using an AC modulation heating can be confined to smaller regions depending on the thermal time constant of the hot spot.

The lock-in technique applied to imaging processing is summarized in figure 6. An infrared image is taken with no bias on the sample. Then a bias is applied and a second image is taken. A difference of the two images is taken and averaged over many pulses. Therefore one obtains an image of thermal emission that is in phase with the applied pulse. The frequency and phase of the image collection relative to the pulsing can also provide information on time constants for local heating of defects. Figure 7 shows a comparison between a static infrared image (left) and lock-in thermography image (right) applied to a poly-crystalline solar cell. Localized hot spots appear near the upper contact bar and edge of the cell. [5]

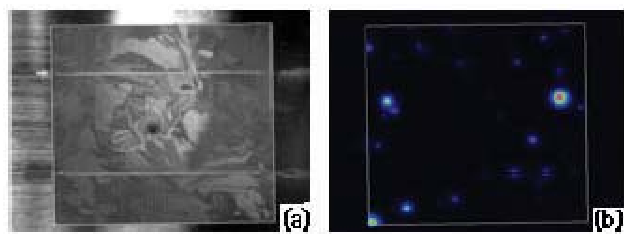


Figure 7: Example of lock-in thermography showing local shunts in a poly-crystalline Si cell. From Berman et al.[5]

Emission Imaging

PV materials are designed to efficiently convert photons to energy. The inverse is also true that solar materials will convert input energy into photons. The input energy can be in

the form of electrons for electroluminescence, photons for photoluminescence or RF radiation for microwave reflection decay measurements [10-13]. Emission imaging of photovoltaics covers the full spectrum from the infrared to visible range with each range providing different information. In the infrared, thermal emission can be used to image local heating as discussed above. In the near infrared emission from defect states and traps within the bandgap can occur and can be used to map dislocation density, for example. Emission at the bandgap occurs from band-to-band transitions and is the dominate emission for electroluminescence. Finally, hot carriers and pre-breakdown emission can occur at wavelengths less than the bandgap.

Electroluminescence (EL) imaging is perhaps the most commonly used for failure analysis because the intensity is proportional to both carrier lifetime (EL efficiency) and current density. Poor contacts and non-uniform current can easily be detected with electroluminescence. Imaging needs to be done with a CCD camera that is sensitive to the appropriate wavelength of emission. This can be a problem when using a Si CCD camera to image EL emission from Si. Figure 8 shows an example of EL imaging of a degraded module. One particular cell is found to have a high contrast of EL emission. High magnification images of the cell shows that it has broken contact lines resulting in non-uniform EL in the region of the line breaks.

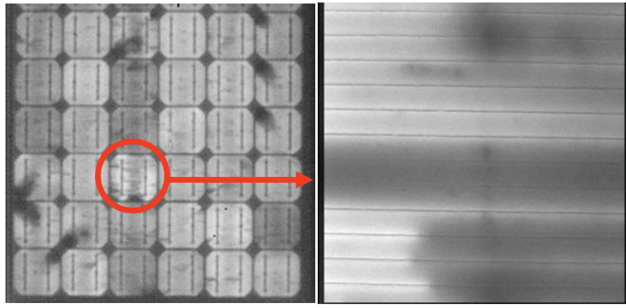


Figure 8: Electroluminescence (EL) image of a cell in a failed single crystalline silicon module. Cells with variation in EL emission had broken AL contact strings shown in the image to the right.

Photoluminescence (PL) imaging has recently been gaining acceptance as a method to map out carrier lifetime in PV materials. Both EL and PL imaging use the fact that emission intensity is proportional to carrier lifetime. Regions in the material that have a long minority carrier lifetime have a higher steady state density of free carriers resulting in stronger radiative recombination. In this case the PL intensity is not a direct measurement of lifetime but needs to be calibrated relative to a more direct method like microwave reflection photoconductive decay. Carrier lifetime is most valuable as a tool to inspect incoming wafers for metallic impurities. Figure 9 shows a comparison of carrier lifetime measurements in an intentionally contaminated Si wafer. Scanning conductivity probes can be used to measure lifetime directly, but the scanning process is very slow. PL images can be obtained quickly and correlate very well with other techniques for measuring carrier lifetime.[5]

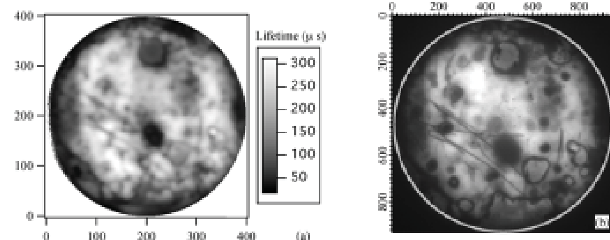


Figure 9: Carrier lifetime measurements obtained with conductivity (left) and photoluminescence (right). Variations in carrier lifetime occur from the non-uniform distribution of impurities. From Berman et al. [5]

Thermal Reflectance Imaging

In principal, infrared imaging has a spatial resolution limited by the wavelength of the emission at 3-10mm. Unfortunately, this resolution cannot always be obtained. Fully packaged modules or solar cells with glass superstrates/substrates can block direct imaging of the active region since typical glass used in fabrication is not transparent to IR light of 3-10 µm. 6 This limits the resolution of LIT to the thickness of the glass or ~3 mm. Thermoreflectance (TR) imaging is a LIT technique that uses the change in a materials reflectivity due to a change in temperature to obtain a high spatial resolution thermal image. [7-9] s technique is typically done with visible light and can produce 200-400 nm spatial

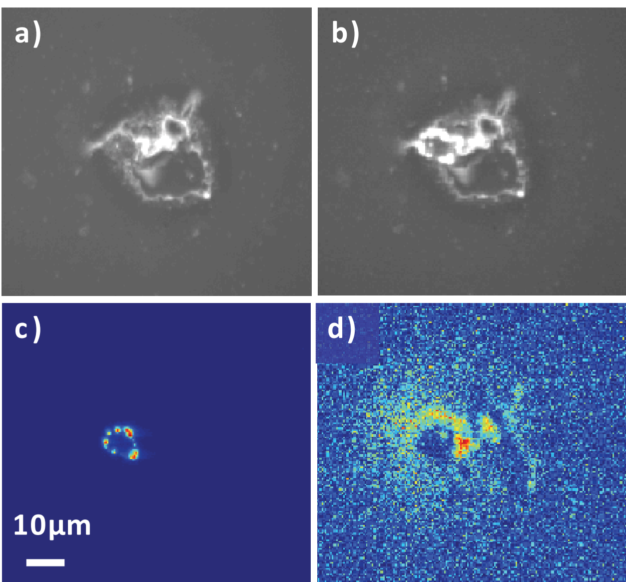


Figure 9: Simultaneous images in brightfield (a), thermal(b) and electroluminescence (c) and combined (d) of a defect in a cell obtained with thermal reflectance imaging. Lock-in thermal reflectance allows for the simultaneous imaging of both thermal and electroluminescence.

resolution and 10 mK temperature resolution after averaging with the use of typical 12-bit CCDs. TR imaging can be used to see through materials such as glass (with visible light) or silicon (with near-IR light >1200nm) in order to obtain high resolution thermal images of areas of interest.[14] Using a silicon-based CCD, visible electroluminescence (EL) is also visible during thermal imaging. Performing locking imaging with a pulsed bias and no illumination gives lock-in EL. Performing lock-in imaging with pulsed bias and illumination gives a thermal reflectance image. Combine pulses of current and light gives images in bright field, thermal and EL simultaneously. This versatile, noncontact characterization tool can help locate a broad range of thermal and EL defects with sub-micrometer spatial resolution. The higher spatial resolution of thermoreflectance imaging, along with its ability to simultaneously obtain visible electroluminescence, to locate and characterize various defects in photovoltaic cells and mini-modules.

Conclusions

The first signature of a failed solar module is loss of power. Examination of the IV characteristics of the module is the first indication of failure mode including loss of contact in a string Text about conclusions reached resulting from the research.

Acknowledgments

Much of the work summarized here is from the National Renewable Energy Laboratory in collaboration with P. Hacke, N. Bosco, D. Albin, T. McMahon and S. R. Kurtz. The thermal reflectance results were done in collaboration with D. Kending and A. Shakouri at the University of California at Santa Cruz.

References

[1] See NREL website: www.nrel.gov/pv/performance_reliability/pdfs/failure_references.pdf

[2] D.L. King, J.A. Kratochvil, M.A. Quintana, and T.J. McMahon, "Applications for infrared imaging equipment in photovoltaic cell, module, and system testing," Proceedings of the 28th IEEE Photovoltaic Specialists Conference, Anchorage, AK, Sept 2000, p. 1487

[3] T. McMahon, Proceedings of the IEEE Reliability Physics Symposium (2008)

[4] D. S. Albin, T. J. McMahon, T. J. Berniard, J. Pankow, S. H. Demtsu, and R. Noufi, "Experiments involving correlations between CdTe solar cell fabrication history and intrinsic device stability," Proceedings of the 31st IEEE Photovoltaic Specialists Conference, Buena Vista, FL, Jan 2005

[5] G.M. Berman, N. Call, R.K. Ahrenkeil and S.W. Johnston, Proceedings of the Fall Materials Research Society meeting (2008)

[6] J. Bauer, O. Breitenstein, and J.-M. Wagner, "Lock-in Thermography: A Versatile Tool for Failure Analysis of Solar Cells," *ASM International*, no. 3, pp. 6-12, 2009.

[7] D. Lüerßen, J. A. Hudgings, P. M. Mayer, and R. J. Ram, "Nanoscale Thermoreflectance With 10mK Temperature Resolution Using Stochastic Resonance," in *Proceedings of the 21st IEEE SEMI-THERM Symposium*, San Jose, 2004.

[8] G. Tessier, S. Hole, and D. Fournier, "Quantitative thermal imaging by synchronous thermoreflectance," *Applied Physics Letters*, vol. 78, no. 16, pp. 2267-2269, Apr. 2001.

[9] J. Christofferson and A. Shakouri, "Thermoreflectance based thermal microscope," *Review of Scientific Instruments*, vol. 76, 2005.

[10] O. Breitenstein, J. P. Rakotoniaina, M. H. Al Rifai, and M. Werner, "Shunt Types in Crystalline Silicon Solar Cells," *Progress in Photovoltaics: Research and Applications*, vol. 12, pp. 529-538, Jul. 2004.

[11] D. Shvydka, J. P. Rakotoniaina, and O. Breitenstein, "Lock-in thermography and nonuniformity modeling of thin-film CdTe solar cells," *Applied Physics Letters*, vol. 84, no. 5, pp. 729-731, Feb. 2004.

[12] O. Breitenstein, J. Bauer, J. Wagner, and A. Lotnyk, "Imaging Physical Parameters of Pre-Breakdown Sites by Lock-in Thermography Techniques," *Progress in Photovoltaics: Research and Applications*, vol. 16, pp. 679-685, Jul. 2008.

[13] M. Kasemann, W. Kwapil, M.C. Schubert, H. Habenicht, B. Walter, M. The, S. Kontermann, S. Rein, O. Breitenstein, J. Bauer, A. Lotnyk, B. Michl, H. Nagel, A. Schütt, J. Carstensen, H. Föll, T. Trupke, Y. Augarten, H. Kampwerth, R.A. Bardos, S. Pingel, J. Berghold, W. Warta, and S.W. Glunz, "Spatially resolved silicon solar cell characterization using infrared imaging methods," in *Proceedings of the 33rd IEEE Photovoltaic specialists conference*. 2008. San Diego, CA, USA. p. in print.

[14] J. Christofferson, K. Yazawa, and A. Shakouri, "Picosecond Transient Thermal Imaging Using a CCD Based Thermoreflectance System," *14th International Heat Transfer Conference*, August 2010. Washington, DC, USA

[15] D. Kendig, G. B. Alers, and A. Shakouri, "Thermoreflectance Imaging of Defects in Thin-Film Solar Cells", Proc. IEEE International Reliability Physics Symposium (2010).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

DRAM Failure Analysis and Defect Localization Techniques

Martin Versen, University of Applied Sciences Rosenheim, Germany

Abstract

Dynamic Random Access Memory (DRAM) failure analysis is a three step process of electrical test and diagnosis, localization and physical failure analysis. Device internal electrical measurements are controlled by the DRAM command interface. Diagnosis results for memory array failures are classified by bitmapping techniques that also provide localization. The limits of bitmapping are overcome by TIVA (Thermally Induced Voltage Alteration), OBIRCH (Optical Beam Induced Resistance Change) and Soft Defect Localization (SDL) for failure in the devices' periphery of the array. The SDL techniques requires adaptation to memory test systems in order to provide high speed comparison allowing SDL image acquisition times of a few minutes.

1 Introduction

Electrical diagnosis of failures is the first step to a successful physical failure analysis of the failure provoking defects. In this paper, an introduction to the dynamic random access memory (DRAM) operation is given with a focus to localization techniques of the defects combined with some physical failure analysis examples and case studies for memory array failures.

In the second part of the paper, the electrical measurement techniques are discussed for array failure analysis. The electrical measurements are performed with DRAM internal circuits and external time control by the DRAM command interface. In the third part, know-how-based analysis techniques of array failures by bitmap classification are presented that provide localization of defects with a reasonable spatial resolution most of the times. The limits of bitmapping are discussed in the forth part that lead to well known localization techniques like TIVA (Thermally Induced Voltage Alteration) and OBIRCH (Optical Beam Induced Resistance Change), that provide localization wherever the methods are applicable. Soft defect localization techniques complete the toolset in the fifth part.

2 Electrical Measurement Techniques

Diagnostic procedures both in product development and in mass production have three steps. Firstly, electrical data from measurement is obtained that is analyzed and that leads secondly to a localization of defects. The third step is the physical failure analysis itself. The information of the physical defects is crucial as a feedback both to system design and technological manufacturing.

Today, there are a number of different DRAM devices types that differ mainly in their memory density and access bandwidth which implies the DRAM array control interfaces like SDRAM, DDRx and Graphics DRAM.

The principle access to the memory cells in the memory array is always the same and is controlled by DRAM commands. A state diagram of the array access commands is depicted in Fig. 1. The different states are shown in red, while the commands are printed in capital black and abbreviated:

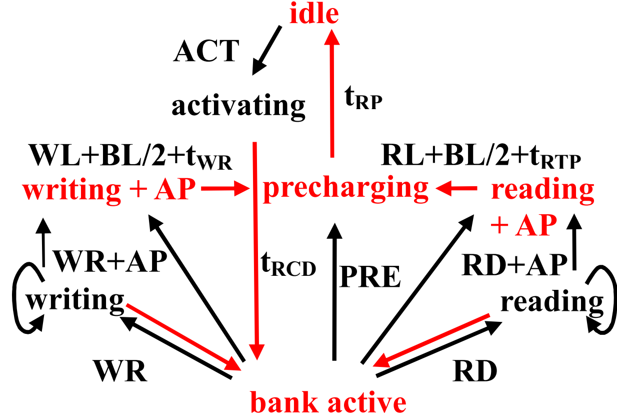


Figure 1: State Diagram of a DRAM device

ACT and PRE are row commands for activation and precharge. AP stands for autoprecharge which combines the precharge with a column command like read (RD) or write (WR). The most important timings between the different commands are t_{RCD} (row to column delay), t_{RP} (row to precharge) and t_{WR} (write timing) that parametrically influence the result of an electronic test of the DRAM cells.

A schematic of a DRAM cell is shown in Fig. 2. The cell consists of a capacitor and a normally off NFET transistor. The two terminals of the capacitor are connected to the plate or reference voltage V_{REF} and to the source or drain of the transistor. The gate electrode of the transistor is driven by the wordline or wordline driver which actually selects the cells and transfers the data to the cell capacitor. The data is provided by the sense amplifiers (see Fig. 4) at the end of the bitline. The data is driven through the bitline and the third terminal of the transistor into the cell capacitor. The data that is stored in the cell is either a high or a low voltage level, V_{BLH} or V_{BLL} (BitLine High or Low) in respect to the reference level V_{REF} which is

$$V_{REF} = \frac{V_{BLH} + V_{BLL}}{2} \quad (1).$$

Test objectives of the cells are for instance a high resistive path between the bitline and the cell that inhibit a write or read access to and from the cell.

Parametric variations of the oxide thickness or doping inhomogenities under the NFET may lead to additional current leakage from the cell so that the charge is lost during the retention time. Other problems may be in the capacitor itself. There may be defective cells of parametric variations of the capacitance of the capacitor. The DRAM cell is tested by use of the sense amplifier which is controlled by the activation (ACT) command (see Fig. 1). The ACT command is issued together with an address, the row address, which is decoded to the wordline address of the cell under test. The initial state of the wordline is the wordline-low voltage level V_{WLL} that turns the selection NFET in Fig. 2 to its off-State so that no current is flowing and no charge and information is lost from the capacitor. With the activation, the wordline is pulled up from V_{WLL} to the wordline high level V_{WLH} that opens the selection NFET in Fig. 2 and allows charge equalization between the cell capacitor and the (parasitic) capacitance of the bitline which is at the reference voltage level V_{REF} before the access. This equalization of charge sharing process is illustrated in Fig. 3. It shows a diagram of the bitline and reference bitline voltage vs. time during the activation process. The voltages in this diagram are normalized to the reference voltage for clarity. The bitline high voltage is $V_{BLH} = 2 \times V_{REF}$, while V_{BLL} is 0V.

The wordline is opened at around 6ns and the discussed charge sharing process is finished after 3ns at t_0=9ns in Fig. 3.

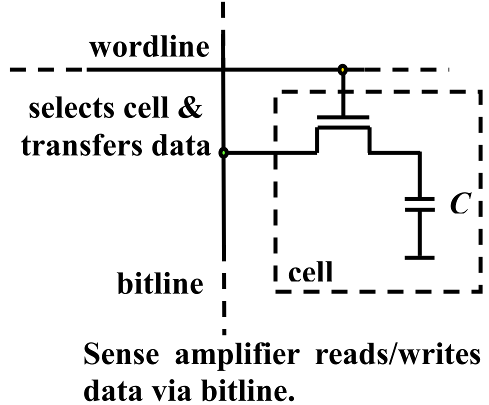


Figure 2: Schematic of a DRAM cell.

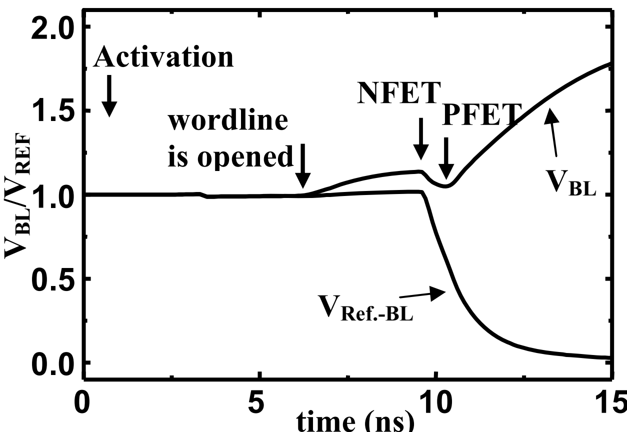


Figure 3: Timing diagram of the bitline voltage during an activation process.

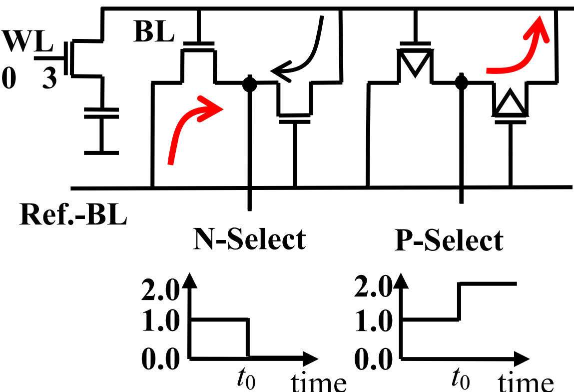


Figure 4: Schematic of a sense amplifier circuit with a cell capacitor and transistor on the left.

The voltage difference ΔV between the bitline voltage V_{BL} and the reference bitline voltage $V_{Ref.-BL}$ at round 9ns is the usable signal that can be amplified by the sense amplifier

$$\Delta V = \frac{C_{Cell}}{C_{Cell} + C_{BL}} \left(V_{Cell} - V_{REF} \right) \qquad (2).$$

A sense amplifier schematic is shown in Fig. 4 together with a cell capacitor and transistor on the left for clarity reasons. Before activation, the NFET and PFET Select Signals, N-Select and P-Select, are the reference voltage V_{REF} so that nearly no source drain voltage difference exists. The transistors are not-conducting and the charge sharing process between cell and bitline is undisturbed. Once the N-Select and P-Select signals are pulled down to 0V and up to V_{BLH}, respectively at t_0, there is a significant source drain difference of V_{REF}. Current may flow if the NFET gate source voltages are high enough to open the transistor. In the depicted case, the bitline voltage is higher than the voltage on the reference line, so that the gate electrode of the left NFET which means that a higher current is going through the left NFET pulling the reference bitline faster to ground that the bitline. This is indicated by the bold arrow below the left NFET in Fig. 4. The voltage on the reference bitlines eventually reaches the threshold voltage of the right PFET that opens and pull the bitline voltage up to V_{BLH} (see bold arrow on the right PFET in Fig. 4). The slope of the bitline voltage vs. time diagram shows a change of sign at around 11ns (see Fig. 3).

This measurement technique depends on the size for the usable signal ΔV in equation (2) and on the already discussed defect scenarios: if the stored cell voltage V_{Cell} is to low due to a resistive path in the connection of the cell, the amplification may fail. Also, if charge is lost during the retention time due to an oxide variation of the selection transistor, V_{Cell} is lowered and ΔV decreases. Also, if the cell capacitor has a low capacitance C_{Cell} the capacitive relationship between cell and bitline capacitance decreases and thus ΔV is decreased. All of these defect scenarios are electrically measurable with the internal function of the DRAM device. If ΔV is not large enough, the amplification process becomes instable, unreliable and a measured test result deviates from the expected result.

3 Bitmapping

The described method for the electrical test for the DRAM cells is applied to all cells of a DRAM cell array that have different cell addresses. If a failure occurs, the fail information is kept together with the fail address, which can be the wordline address X and the bitline address Y. This fail information is can be visualized in bitmaps, which that a map of failing bits is shown. The coordinates of the maps are the X and Y addresses of the cells. Four sample bitmaps are shown in Fig. 5.

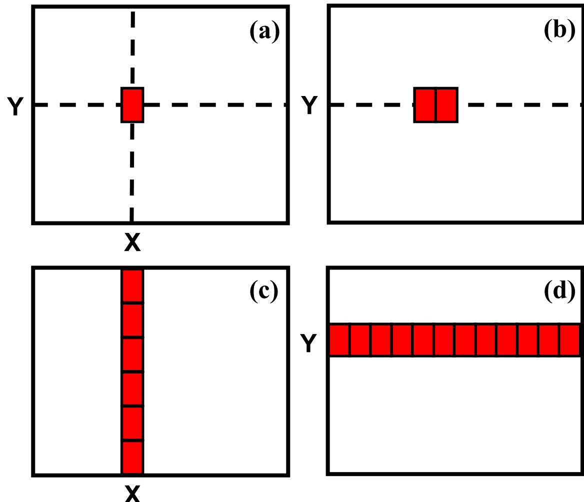


Figure 5: Bitmap examples: Single cell (a), Paired cell (b), wordline oriented (c) and bitline oriented fail (d).

The upper left bitmap (a) shows a single cell failure, which is characterized by its X and Y address and which provides also a localization with a lateral resolution of ±*F*, the smallest feature size of the DRAM device. Additional information regarding parametric dependencies either from electrical test or from technological parameters [1, 2] that influence the usable voltage difference ΔV leads to an even better resolution that includes also the layer information.
Bitmap (b) is a paired cell example. In this case, the two failing cells have a common column address Y and differ in the row address by 1. The localization is better than ±*F*, because a defect is assumed at the position of a common layout element of the two neighboring cells. The cases (c) and (d) show bitmaps for a wordline and a bitline oriented failure, respectively. The assumption that a defect causes all of the connected cells to fail always leads to common layout elements of the electrical paths, although the resolution decreases, as more layout elements can be in the critical electrical path. The layout elements are known from layout data and have a physical ±*x*/±*y* position referenced to the center of the rectangular die. A better isolation of the failure site can only be achieved by know-how based FA or "design of experiment" methods: a list of possible (layout) failure sites is created and the hypothetical failure sites are either verified or falsified by experiments. These experiments include both design simulation and hardware tests. The resolution is increased by exclusion of false failure sites.
The bitmapping technique is firstly used for classification of failure types in order to size the impact of a failure types to the yield and secondly for a localization of the defects for physical failure analysis that lead to technological improvements. Bitmapping is the working horse of DRAM yield improvement (e.g. [1, 2]). The physical failure analysis flow and selected techniques depend on the failure type. The two major techniques include in plane SEM and cross sectional FIB/SEM or TEM inspections. Two examples for in-plane and cross section SEM inspection after FIB preparation are shown in Fig. 6 and 7. In Fig. 6, an example for a wordline oriented fail with the bitmap 2(c) is presented, while a bitline oriented fail example in shown in Fig. 7.

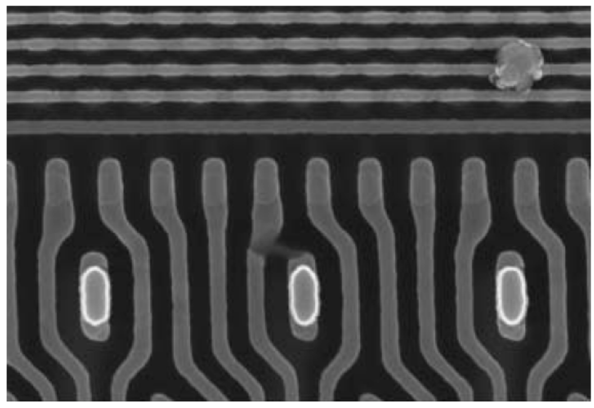


Figure 6: Failure site inspected by SEM [3].

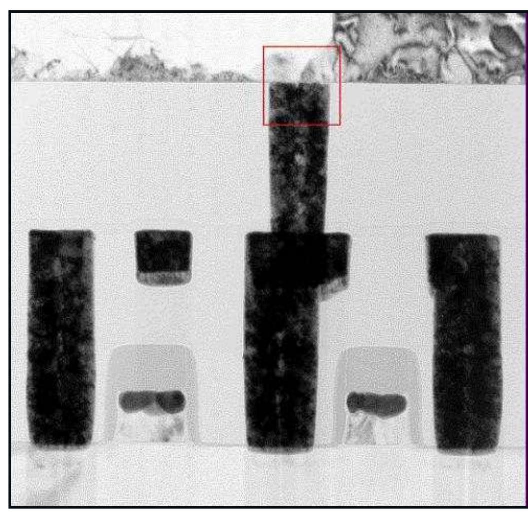


Figure 7: Failure site after cross sectioning by FIB again inspected by SEM [4].

4 Limits of Bitmapping

The limits of bitmapping appear if the localization is so bad that a defect may be anywhere in the die. An example is shown in Fig. 8 representing a bitmap in which one eighth of a DRAM is failing. A failure like this usually considered as a hard fail. The fail addresses of the cells in the array do not provide any defect position that could be used for localization. A block fail may have different reasons like contact issues and chip cracks. Parametric test data from contact test, input leakage and I_{DD} test together with optical inspection of the decapsulated part reveal the problems most of the time. On the other hand, other problems due to local defects in the chip internal supply networks that cause massive power losses are electrically, but not physically localized by parametric test data.

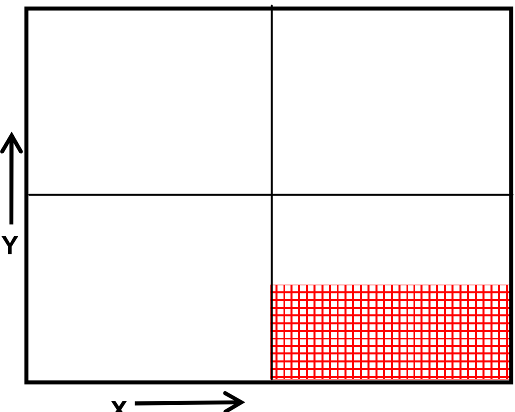


Figure 8: Bitmap example with a block fail.

Current and leakage current measurements characterize the symptoms of these problems and provide the two terminals for standard methodologies for laser scanning localization technique like TIVA and OBIRCH that complete the failure analysts' toolbox. A constant current or voltage source is applied to the two terminals of a device under test (DUT) and a voltage or current drop or increase is measured at the device, respectively [5-9]. These methods are discussed elsewhere in this desk reference.
A similar failure type is described with it's bitmap in Fig. 9. A logic problem in the periphery, in the logic part of the DRAM, provokes cell fails in the array. Similar to the previously discussed problems, any position of the fail addresses is not a defect position, because there is no common layout element to all of these fails. The only elements that are used for all of these depicted, failing cells in Fig. 9 are also used for other completely functional cells.

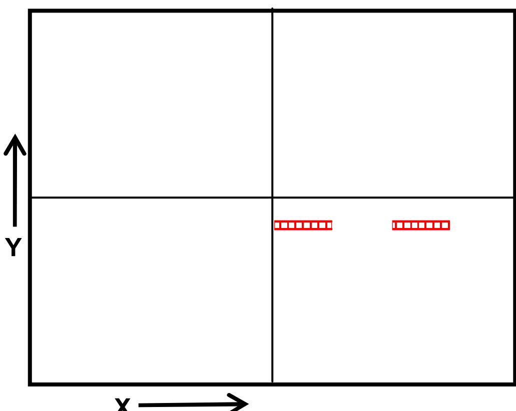


Figure 9: Bitmap examples of a fail that is caused by a logic problem in the periphery.

Logic problems that are caused by temperature dependent defects are called "soft fails" and cannot be localized by a TIVA or OBIRCH, because the methods do not apply to the internal switching status of an electrical device but to the analogue supply.
Tester assisted soft defect localization is necessary for this type of fails, which are more rare, as the design rules for the logic part of DRAM are relaxed compared to the smallest feature size *F* in the array.

5 Soft Defect Localization

Soft defect localization (SDL) is a well known method that allows a localization of functional fails within the periphery of the electronic devices. Laser scanning localization with a test system utilizes the trigger or synchronous input of the microscope electronics shown in Fig. 10: a functional pattern is executed and any local defect may cause a fail depending on the pattern. The tester evaluates the result (pass or fail) from the switching state of the DUT and sends a digital signal together with a trigger to the microscope electronics. This technique allows the localization of any temperature dependent defect that causes any function failure, but unfortunately standard and the improved methods can not be used for DRAM. This procedure implies that the tester pin electronic is working at a speed of a pixel well time of the laser scanning setup.

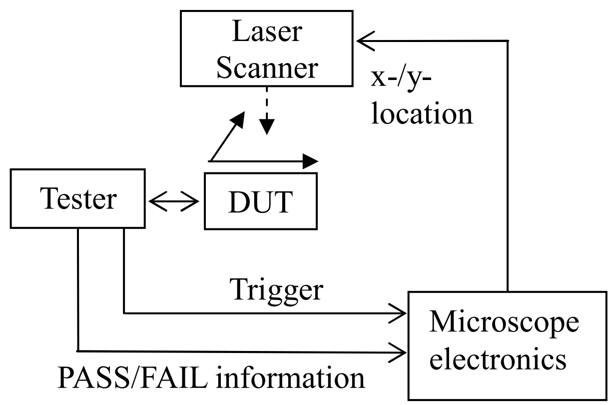


Figure 10: Laser scanning localization setup with synchronous input.

Unfortunately, DRAM test systems often use algorithmic pattern generator for test execution and evaluation. Although an executable pattern can be in µs or ms time range, the test evaluation time that is required before the next test is executed (also called test-to-test index time) amounts roughly to 300ms which is much longer than a practical pixel dwell time.
Consequently, the pixel dwell time must be adjusted. If a 128x128pixel image should be acquired with this test-to-test index time, a scan time of 82min (=128x128x 300ms) would be required. The method is therefore not applicable to DRAM devices with classical memory test systems, because the stability of a laser scanner is generally only guaranteed within a time of approximately 15 to 20min.
A possible solution next to a use of a different test system is an additional instrumentation of a fast pin electronics that can handle DRAM output data. A modified SDL setup is depicted in Fig. 11, in which the tester only delivers a triggers signal that is set within the algorithmic patter to indicate the valid data point of time. The data of the DUT is compared with expected data that is programmed in the fast pin electronic. A possible instrumentation is shown in [10].

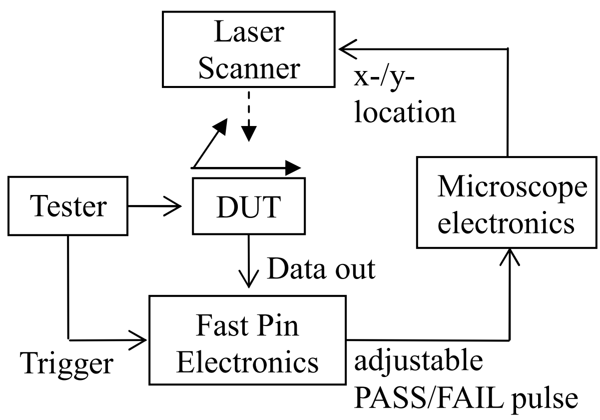


Figure 11: Soft Defect Localization (SDL) Laser scanning localization setup with a fast pin-electronics as an additional instrumentation.

Fast pin electronics is necessary, as today's DRAM devices have specified data rates of more than several hundreds of Mb/s. Application cases show that pattern run times of 10 to 100µs are achievable which is also in a reasonable range for the pixel dwell time. A pixel dwell time of 128µs implies a recording time of 34s for a 512x512 pixel image. The acquired image is overlaid with a backside infrared image of the die to reference a local heating spot to the physical surroundings (see Fig. 12).

After localization, the already mentioned inspection methods are applied for physical failure analysis: in plane SEM and cross sectional FIB/SEM or TEM inspections. A TEM inspection example is shown in Fig. 13.

This localization method could also be applied for defect localization within the cell array, but bitmapping seems to be faster in terms of turnaround times especially if more than one sample is available. As this method is not restricted to production fail samples, it will apply to important samples for which the resolution is absolutely required like for customer returns and reliability fails samples.

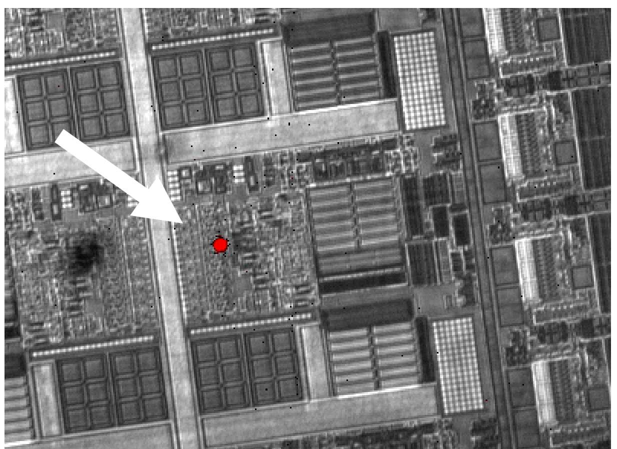


Figure 12: Overview overlaid images of the backside and SDL image in x20 magnification. The arrows mark the position of the observed spot [10].

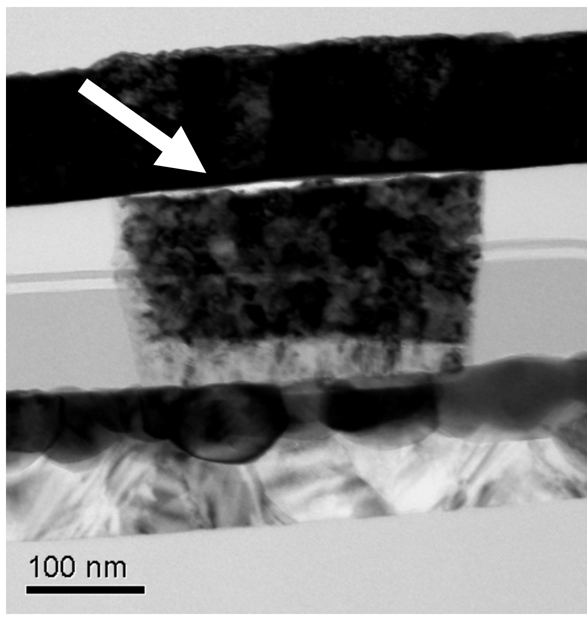


Figure 13: TEM image of a fail gate contact [10].

6 Summary

An introduction to the electrical diagnosis of DRAM devices for a successful product development has been given. Defects and parametric variations of cells are the test objectives . The time control is realized by external commands. The primary sense amplifier is a voltage measurement setup. Fails are recorded in bitmaps that allow easy characterization and easy localization by know-how based analysis through layout information. Bitmapping is the working horse of DRAM yield and quality improvements. The limits of bitmapping are overcome by laserscanning localization techniques like TIVA, OBIRCH and Soft Defect Localization for the DRAM periphery. This last method provides requires extra instrumentation with fast pin electronics.

7 Literature

[1] J. Vollrath, U. Lederer, T. Hladschik, "Compressed Bit Fail Maps for Memory Fail Pattern Classification", *ETW Proceedings*, 207-212 (2000).

[2] J. Vollrath, U. Lederer, T. Hladschik, "Compressed Bit Fail Maps for Memory Fail Pattern Classification", Journal of Electronic Testing **17**, 291-297 (2001).

[3] M. Versen, F. Ertl, D. Diaconescu, B. Straube, W. Vermeiren, F. Hobsch, „Fehleranalyse von Kurzschlüssen in DRAM Wortleitungstreibern durch Extraktion von Layout Parasitics und ΔIDD3N Messungen", *Proceedings of Testmethoden und Zuverlässigkeit von Schaltungen und Systemen*, 103-107 (2009).

[4] M. Versen, A. Schramm, F. Schamberger and I. Klein, Defective Contacts in DRAMs: From Electrical to Physical Failure Analysis, Electronic Device Failure Analysis, 2006, Vol.8, Issue 1, pp. 6-14.

[5] E.I. Cole Jr. et al., "Resistive Interconnect Localization", *ISTFA Proceedings*, 43-50 (2001).

[6] M.R. Bruce et al., "Soft Defect Localization (SDL)", *ISTFA Proceedings*, 21-27 (2002).

[7] F. Beaudoin et al., "Laser Stimulation Applied to Dynamic IC Diagnostics", *ISTFA Proceedings*, 371-377 (2003).

[8] E. I. Cole Jr., „Beam-Based Defect Localization Methods", Desk Reference of Microelectronics Failure Analysis, 5th Edition, 406-416 (2005).

[9] F. Beaudoin, R. Desplats, P. Perdu, C. Boit, „Principles of Thermal Laser Stimulation Techniques", Desk Reference of Microelectronics Failure Analysis, 5th Edition, 417-425 (2005).

[10] M. Versen, A. Schramm, J. Schnepp, S. Hoch, T. Vikas, D. Diaconescu, "Laser Scanning Localization Technique for Fast Analysis of High Speed DRAM devices", *ISTFA Proceedings*, 227-232 (2008).

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Failure Analysis of Passive Components

Stan Silvus
Southwest Research Institute®
San Antonio, TX USA

INTRODUCTION

Failure analysis of passive components is a very broad topic because of the numerous device types and constructions that exist. Articles in previous editions of the ***Microelectronics Failure Analysis Desk Reference*** (References 1 and 2) have presented the fundamentals of many types of passive components along with appropriate analysis techniques and procedures. Rather than repeat the previously presented material, the present article will supplement these earlier treatises by providing detailed information on selected passive-component types. Accordingly, this article includes typical construction features and failure mechanisms of the selected component types, as well as some case histories that illustrate common failure mechanisms of these and other component types.

COMPONENT DESCRIPTIONS

Surface-Mount Thick-Film Resistor

Construction. A screen-deposition process is usually used in fabricating surface-mount thick-film resistors. The starting material is a thin wafer (substrate) of alumina ceramic. Metallic terminations (pure silver or silver-palladium alloy) may be deposited and fired first, and the resistive material, often a ruthenium oxide based material, may be deposited and fired second; however, some manufacturers reverse this order. Usually, thick-film resistors are fabricated with lower-than-desired resistance, and laser trimming is used to adjust resistances to their final values. Various trimming patterns are employed in the industry; in some cases, there are two straight trim kerfs projecting into the resistive material from opposite edges, but in others, a single straight or L-shaped trim kerf is used. To provide mechanical protection, the resistive material is covered by a fired overglaze; sometimes, trimming is performed before the overglaze has been applied, but in other cases, the trim kerf may cut through previously deposited overglaze.

After individual resistors have been separated from the substrate, a barrier layer (typically nickel) is deposited on the end terminations, and the barrier layer is coated with solder or tin. A typical chip resistor is shown in Figure 1; this resistor was trimmed before deposition of the overglaze. Note that thick-film resistor networks are fabricated by a similar process.

Failure Modes and Mechanisms. Surface-mount thick-film resistors may fail open or shorted, or they may exhibit parametric changes. They may be electrically overstressed, a failure mechanism that is usually accompanied by visible signs of overheating. Delamination between layers of a thick-film resistor may permit intrusion of solder flux, cleaning solvents, or water, and these foreign materials may support corrosion or electrochemical migration or may cause parametric changes.

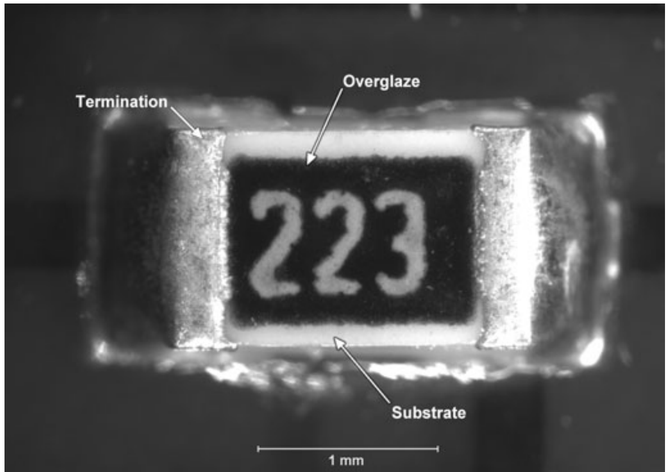


Figure 1. Typical thick-film chip resistor mounted on circuit board (Trim kerf covered by overglaze)

The trim kerf, which may extend as much as half way across the resistive material, causes current flowing through the resistor to be pinched, thereby increasing local current density by as much as 200% (refer to Figure 2). Under steady-state conditions, the effects of current pinching are accounted for in the device ratings, but in some applications, there may be short-duration transients during which current and, hence, power dissipation are much greater than the device ratings. As a result of current pinching, heating will be concentrated at the tip of the trim kerf as indicated in Figure 2, yielding a localized hot spot. If energy associated with the transient event is sufficiently high, the under side of the overglaze may be locally melted, and the molten material may infiltrate the grain boundaries of the resistive material. Each time that this occurs, resistance of the device increases. Eventually, resistance of the device will increase to the point that circuit function is altered. An indicator of this type of failure is that removing the overglaze causes partial recovery of the resistance.

Because terminations on thick-film resistors are usually silver based, electrochemical migration and tarnish are possible failure mechanisms (References 3 and 4). Tarnish occurs in applications in which sulfur-containing gases are present in the environment; vulcanized rubber is a common source of sulfur-containing gases. Residues of solder fluxes, including "no-clean" types, under a surface-mount component support electrochemical migration, so cleanliness is the key to preventing this failure mechanism. Usual instruments (e.g., the Omegameter® and ion chromatography) used to check circuit-board cleanliness yield values that are averaged over the entire surface area of the circuit board, so an acceptably low average ionic-contamination reading does not guarantee absence of high concentrations under small surface-mount components.

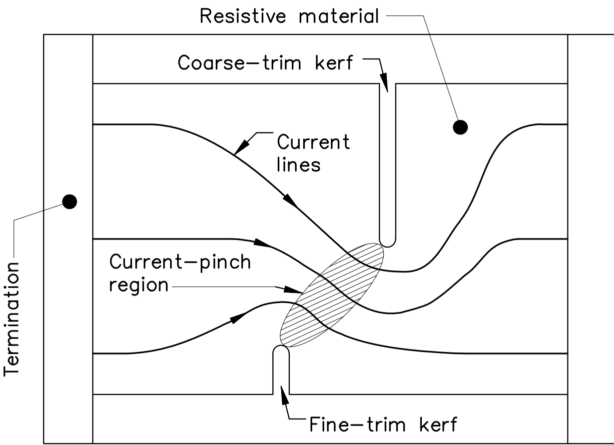


Figure 2. Current-pinching effect in surface-mount thick-film resistor

Surface-Mount Stacked-Film Capacitor

Construction. Stacked-film capacitors are made by stacking numerous sheets of metallized polymeric film. Metallization (zinc, aluminum, or a combination of these metals) on each sheet is patterned so that many capacitors may be stacked simultaneously. Additionally, each sheet is coated with a thin adhesive (often, a wax). After the required quantity of sheets has been assembled, nonmetallized cover sheets are applied to both sides of the stack, and the resulting structure is compressed and heated; this operation bonds the sheets in the stack. Next, the stack is diced to yield many individual capacitors, following which terminations, usually zinc, are applied by plasma spraying. Finally, a solderable coating is applied to the terminations. A typical surface-mount stacked-film capacitor is shown in Figure 3.

Failure Modes and Mechanisms. Like all capacitors, stacked-film devices may fail shorted or open, or they may exhibit parametric changes. Separation of the termination from the ends of the metallized film causes an open circuit or a decrease in capacitance. Small short circuits may occur through pinholes in the polymeric film, and if energy available in the circuit is sufficient, but not too great, self healing will occur because of the ease with which the very thin metal film evaporates. However, if enough energy is available when a short circuit forms, self healing will not occur, and there may be obvious physical damage to the capacitor. Delaminations in the film stack occur frequently; often these delaminations are of no consequence, but in some applications in which the high leakage resistance expected of a polymeric-film capacitor is required, contamination that enters delaminations may cause circuit instability or outright malfunction.

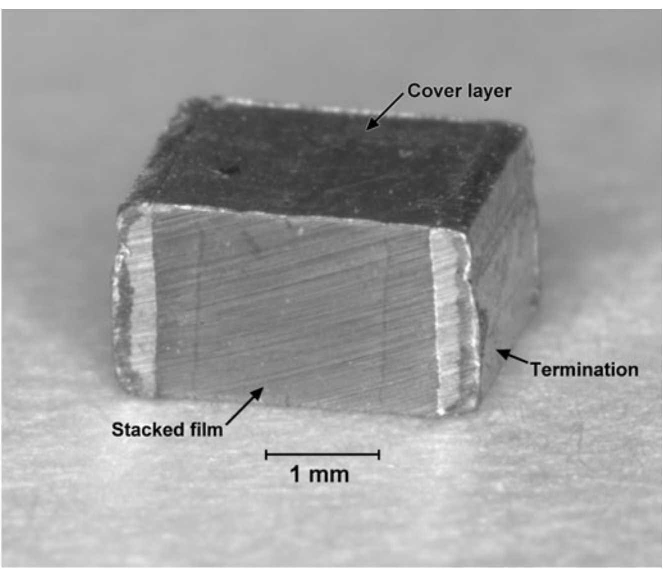


Figure 3. Typical surface-mount stacked-film capacitor

Fluid-Filled Metallized-Film Capacitor

Construction. Film capacitors intended for use in alternating-current (ac) power applications (e.g., some single-phase motors, power-factor correctors, power-line filters, etc.) are usually of the fluid-filled metallized-polymer-film type, a typical example of which is shown in Figure 4. In most such capacitors, the dielectric film is polypropylene, and the thin deposited metallization is zinc, aluminum, or a combination of these metals. Two strips of metallized film, possibly with additional strips of nonmetallized film, are rolled to form a section (Figure 5). End terminations on the section are usually sprayed-on zinc. Because the termination process produces overspray, it is customary to drill at least one end of the section core and to brush the outer surface of the section to interrupt potential short-circuit paths. The metal-film plates are very thin, so a localized pinhole short circuit through the dielectric film will self heal if energy dissipated at the failure site is not too great.

Often, a power capacitor has a built-in pressure-actuated protection device. Typically, as pressure in the capacitor housing increases during a catastrophic failure or during normal self-healing (i.e., clearing) events, bulging of the housing cover puts the leads attached to the section terminations in tension so that the associated solder joint or weld is ripped apart, thus opening the circuit.

While power capacitors are designed to have very low equivalent series resistance (ESR) values, they nonetheless can dissipate substantial heat because of the high currents usually flowing through them. To enhance heat removal, the section is frequently immersed in a dielectric fluid, commonly referred to as “oil.” The oil may be castor oil, but in modern devices, the fluid may not be oil in the conventional sense; instead it may be one of a family of synthetic organic materials that may range from freely flowing liquids to highly viscous syrups. In addition to facilitating heat transfer, the fluid has a second function; that is, it excludes oxygen from narrow spaces between metallic materials that have high potential differences, thereby preventing corona and consequent oxidation of the thin metal film.

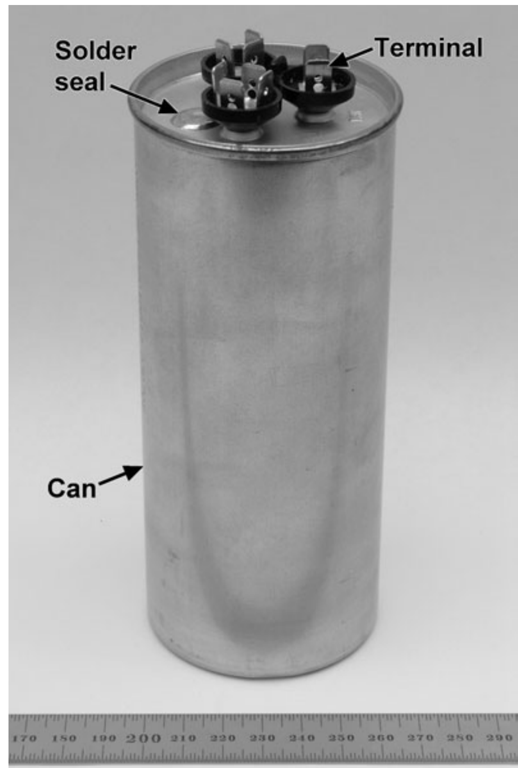


Figure 4. Typical fluid-filled metallized-film capacitor intended for alternating-current service

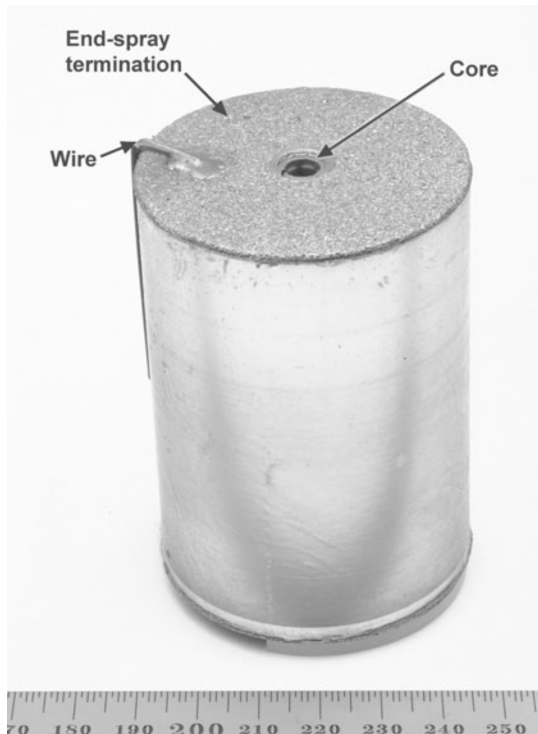


Figure 5. Typical section removed from fluid-filled metallized-film capacitor

For the fluid to perform its dual functions properly, the entire section or sections must be covered; however, there must be some ullage to provide a gas cushion that prevents false tripping of the pressure-activated protection device as ambient temperature rises and the fluid expands. Because of these requirements, fluid-filled metallized-film capacitors should be mounted only with their terminals up; if this type of capacitor is mounted on its side or, worse yet, with its terminals down, then the necessary ullage leaves part of the section uncovered by the fluid, and in such cases, failure usually occurs at or slightly above the fluid line.

There are two common processes for filling fluid-filled capacitor housings: (1) vacuum impregnation and (2) direct injection. Either process can be made to work, but vacuum impregnation, though more expensive and time consuming, is easier to control, and this process yields much more consistent fill levels. In capacitors filled by direct injection, fill level often varies widely; in fact, it is common for the sections in such capacitors to be only partially covered even when the devices are properly mounted with their terminals up. Low fluid level usually leads to premature failure, and as noted earlier, failure usually occurs at or just above the liquid level. Very low fluid level may also increase ullage to the point that the gas cushion is so soft that the pressure-activated protection device will not work.

Failure Modes and Mechanisms. There is a popular notion that because of the inherent self-healing property, fluid-filled metallized-film capacitors never develop short circuits, but this is definitely not the case. In fact, sections in fluid-filled metallized-film capacitors normally fail shorted, and when this happens, the effects are usually catastrophic. However, if an internal protection device is present, then a capacitor may appear open as viewed from its package terminals even though it is shorted internally.

If a fluid-filled metallized-film capacitor does not fail catastrophically, then its capacitance slowly decreases with passing time. This effect is caused by self-healing or clearing events that may be triggered by small latent defects (e.g., pinholes) in the polymeric film or by transient overvoltages (e.g., power-line-voltage surges or spikes). Each clearing event reduces the total plate area slightly, and, of course, reduced plate area implies decreased capacitance; decreasing capacitance will eventually cause circuit malfunction. A typical non-catastrophic cleared site is shown in Figure 6; it is not unusual to see millions of similar cleared sites in a capacitor that has been in service for a long time. The cleared area illustrated in Figure 6 is approximately centered around a small pinhole in the polymeric film; immediately surrounding the pinhole, both layers of metallization have cleared, but in a larger area, only one layer has cleared.

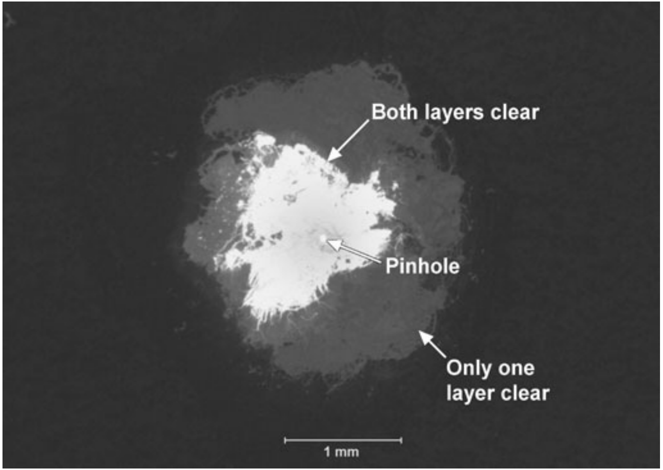


Figure 6. Backlighted view of normal cleared site in a metallized-polymeric-film capacitor

As alluded to in the foregoing discussion of construction features, fluid level in a fluid-filled metallized-film capacitor is very important. If part of a section is not covered with fluid, probability

of catastrophic failure at or just above the liquid level increases. Temperature rise in the uncovered part of the section will be higher than in the fluid-covered part of the section, and the higher hot-spot temperature may decrease both mechanical and dielectric strengths of the film; moreover, the temperature gradient at the transition between fluid-covered and uncovered parts of the section may introduce localized mechanical stresses that increase susceptibility of the film to dielectric breakdown. In some cases, it is better to have no fluid than to have a section that is only partially covered.

Failure-Analysis Tip. In analyzing a failed fluid-filled metallized-film capacitor, it is a good idea to weigh the device (if it is not leaking) before any destructive operations are performed. If multiple capacitors of the same type, hopefully including some good devices, are available, they should be similarly weighed, and weights of all devices should be compared; light-weight capacitors are most likely to have low fluid levels.

After a capacitor has been weighed, lay it on its side for several hours (overnight works well); then carefully cut a narrow longitudinal slot in the upward-facing side of the capacitor can, and immediately cover the slot with wide adhesive-backed transparent-plastic strapping tape. Following these preparatory operations, stand the capacitor with its terminal end up for two to four days to allow the viscous fluid to settle. If this process has been performed correctly, it will then be possible to see both the top of the section and the fluid level through the slot (refer to Figure 7).

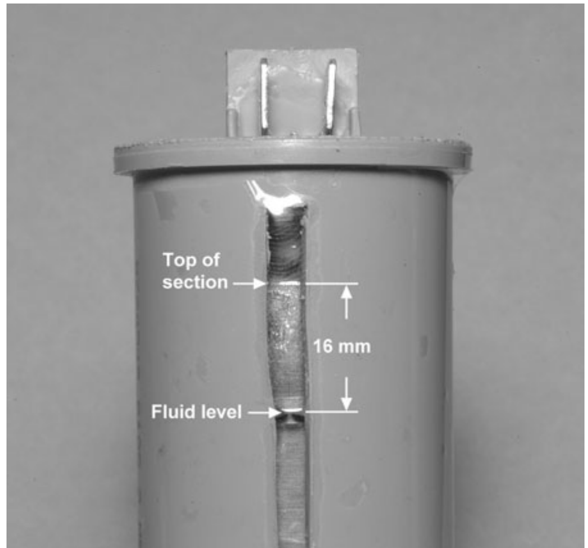


Figure 7. Top of section and fluid level visible through longitudinal slot in capacitor can wall

Aluminum Electrolytic Capacitor

References 1 and 2 provide much information on small aluminum electrolytic capacitors. Construction features of large versions of these devices (e.g., so-called "computer-grade" capacitors) are similar, but differences do exist; accordingly, details of large aluminum electrolytic capacitors will be reviewed briefly.

Plates in an aluminum electrolytic capacitor comprise two deeply etched aluminum-foil strips that are terminated with transversely oriented crimp-attached aluminum ribbons. Two such plates and two or more slightly wider strips of porous paper are formed into a roll. The paper separators are saturated with a liquid electrolyte, and the roll is placed in an deep-drawn aluminum can. In large aluminum electrolytic capacitors, the open end of the can is closed by a rigid header that is sealed with a gasket. The header is penetrated by the device terminals, which in large capacitors are usually of the screw type. After the roll has been packaged, voltage is impressed between the device terminals to form the thin aluminum oxide dielectric on the positive plate. A typical large aluminum electrolytic capacitor is illustrated in Figure 8.

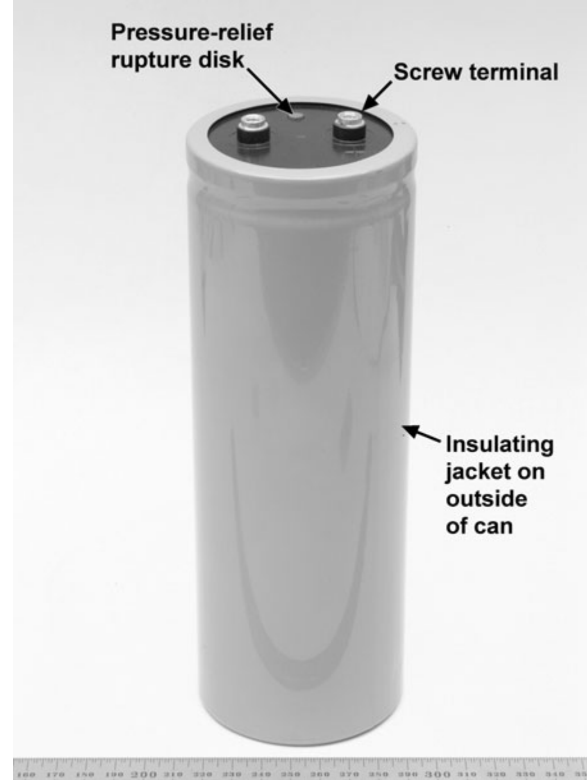


Figure 8. Typical large aluminum electrolytic capacitor

Usually, the electrolyte in an aluminum electrolytic capacitor, large or small, has a fishy odor. If such an odor is not detected just after the device has been opened, contamination of the electrolyte should be suspected.

Wound Components

As mentioned in Reference 2, wound components come in so many types, constructions, and package styles that a comprehensive discussion of these devices is far beyond the scope of a ***Desk Reference*** article.

Failure Modes and Mechanisms. Short circuits between adjacent turns within a single layer, short circuits between layers, open circuits, and short circuits between a winding and the core are the most frequently observed failure modes in wound components. Open circuits are usually caused by excessive current (i.e., fusing of the wire) or by mechanical abuse, such as overstress of a termination. Short circuits result from insulation failure, which is usually caused by dielectric breakdown that, in turn, is brought about by overheating, excessive voltage stress, or insulation aging. All types of electrical insulation, and especially the insulation systems used in wound components, are subject to aging that is accelerated by elevated temperature. Wound components that must handle power get hot, so insulation aging is a built-in wear-out mechanism in such devices.

Corrosion of the wire in a wound component can lead to failure (Reference 5). It is possible for a wound component to have turns that have been severed by corrosion while continuity still exists because of an electrically conductive path through the

corrosion products. In such cases, winding resistance and quality factor (i.e., Q) are usually out of the specification range.

Failure-Analysis Tip. Before attempting to check continuity or measure resistance of a winding in a wound component, connect a high-input-resistance dc millivoltmeter across the winding. If a voltage is present (e.g., a few millivolts or higher), then corrosion in the winding should be suspected, and subsequent failure-analysis operations should be selected accordingly. Usually, if corrosion is present, it will be in parts of the winding where the wire is closest to a grounded metallic surface (e.g., the core) or near solder joints at the winding terminations.

CASE HISTORIES

Case histories effectively illustrate the types of failures that occur in actual practice and the methods employed in analyzing these failures. Accordingly, a number of passive-component case histories are discussed in the following subsections of this article.

Surface-Mount Thick-Film Resistors

Mechanical Damage. A UHF balun began behaving erratically during a temperature-cycling test. Opening the device and inspecting the circuit board revealed that one of the matching resistors was cracked transversely (Figure 9). Additionally, there was a chip, aligned with one end of the crack, in one edge of the substrate. Detailed examination of the failed resistor showed that there was a metallic smear on the ceramic substrate adjacent to the chip, and energy-dispersive x-ray analysis of this smear detected iron, an element that did not exist elsewhere on the resistor or in nearby components on the circuit board. It was also noted that there was a twisted pair of wires near the chip; this twisted pair formed a low-value capacitor that had been used in trimming frequency response of the balun. In the trimming process, radio-frequency response of the balun is determined with a network analyzer, and if the response is off in a selected frequency band, part of the twisted pair is clipped off with wire cutters; these steps are repeated until balun response meets specifications. In one iteration, probably the last, of the trimming process, tips of the wire cutter (i.e., an iron tool) nicked the edge of the brittle ceramic substrate of the resistor causing the aforementioned chip. It is likely that the transverse crack did not immediately sever the resistive material; instead, it is probable that the resistive material was fractured during temperature cycling.

Thermal Damage. Another UHF balun failed during vibration testing. Dc electrical checks of the circuit board in this device showed that one of the matching resistors was open, and subsequent visual inspection of this resistor revealed a crack in its substrate under the termination (Figure 10). Location and orientation of this crack strongly suggested nonuniform heating of the brittle ceramic substrate, a phenomenon that occurs frequently when improper soldering techniques are used.

Nowadays, ceramic surface-mount components are usually soldered by reflow techniques that provide the necessary heating uniformity. However, during a quality audit of the balun manufacturer, it was observed that a conventional soldering iron was being used to solder surface-mount resistors and capacitors to circuit boards for this relatively low-volume product and to rework the solder joints on these components. Unless a conventional soldering iron is used with great care, there is high probability that cracking of the ceramic material will occur. Usually, thermally induced cracks lie under the device terminations where they are extremely difficult to see. If ceramic components must be manually soldered or if solder joints on such components must be manually reworked, it is imperative that a hot-air soldering tool be used.

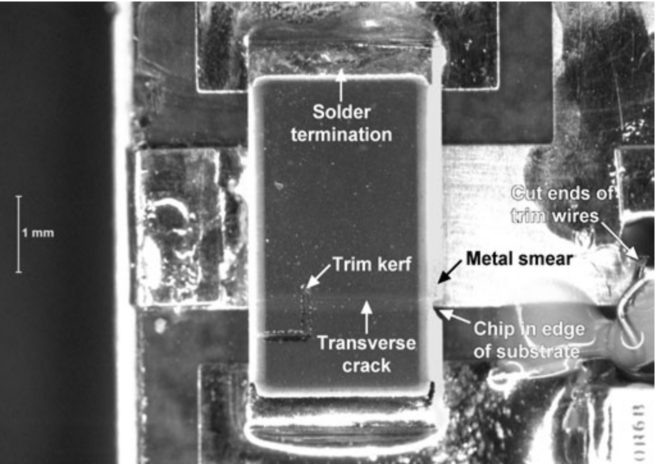


Figure 9. Area surrounding cracked surface-mount thick-film resistor

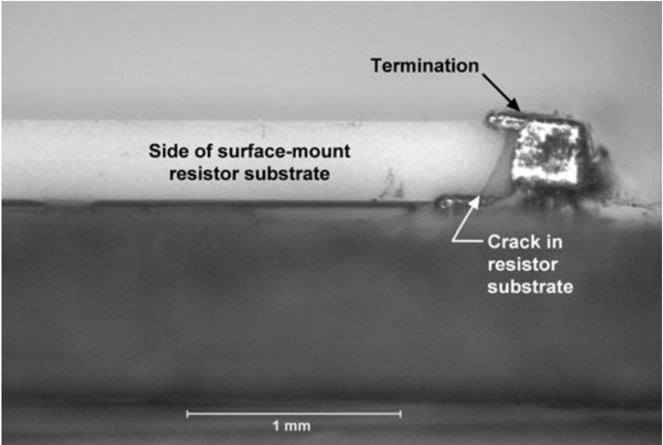


Figure 10. Side view of cracked surface-mount thick-film resistor still mounted on circuit board

Pattern Misalignment. During final test of an electronic device, an open surface-mount thick-film resistor was detected. Overall appearance of this resistor (Figure 11) was normal, except that there was a possible lateral shift in the resistive-material deposition relative to the terminations.

Two mutually exclusive next steps were possible: (1) the resistor could be delayered or (2) the resistor could be cross sectioned. Delayering usually requires some tailoring of the process, and this requires multiple samples; in the present case, only the single failed device was available for analysis, so cross sectioning was selected because of the higher probability of first-time success. Figure 12 is a high-magnification cross-sectional view of the region in which the resistive material and the termination should overlap; it is evident that these materials did not overlap, so there was an open circuit at this site. For comparison purposes, Figure 13 shows the opposite end of this resistor where there was more-than-adequate overlap. Thus, the resistive material deposition pattern had, indeed, been shifted with respect to the terminations during device manufacture.

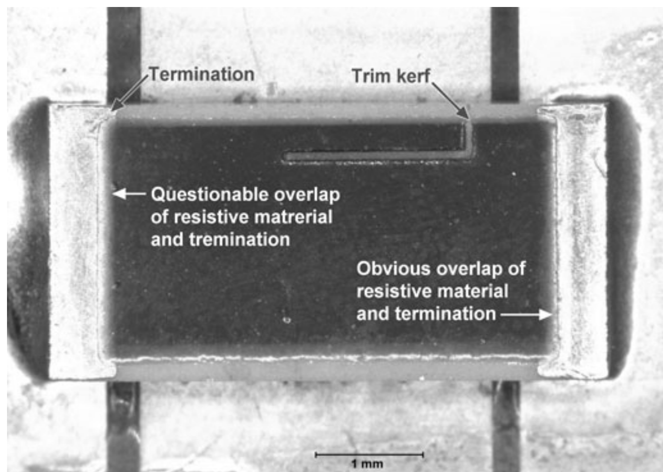


Figure 11. Top view of failed resistor still mounted on circuit board

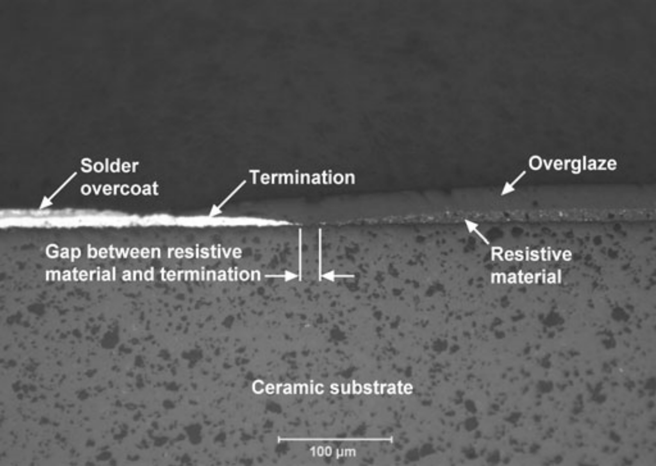


Figure 12. Cross section through open end of failed resistor

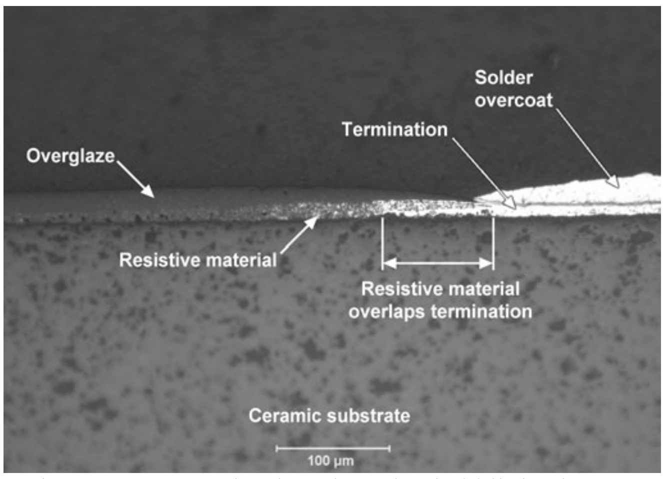


Figure 13. Cross section through good end of failed resistor

As an aside note, this resistor had been trimmed after the overglaze had been applied, and the trim kerf was L-shaped. Also, the terminations had been deposited first, and the resistive material had been deposited in a subsequent operation.

Parametric Shift. Over a few hundred cycles of equipment activation, resistances of several size-0805, 0.125-W surface-mount thick-film resistors increased to the point that the circuit would no longer function. Nominal value of the failing resistor was 3.9 Ω, but resistance after failure had been detected was in the range of hundreds to thousands of ohms. Detailed inspection of numerous failed resistors did not reveal any externally visible defects or damage. Cross sectioning of representative devices showed that the resistive material properly overlapped the terminations and that there were no cracks. However, during delayering of some of the failed devices, resistances decreased to a few ohms (typically, 8 Ω to 20 Ω), but in no case as low as the initial resistance value.

Analysis of the circuit in which the failed resistors had been used showed that a high-peak-current surge occurred each time the equipment was activated; following this transient, current through the resistor dropped to a small, safe value for the remainder of the equipment operational cycle. However, during the transient event, which lasted about 90 µs, peak power dissipated in the resistor was 48 W. Because of current crowding, discussed earlier in this article, a high fraction of this 48 W was concentrated in the narrow region of the resistive material between ends of the two trim kerfs. Damage to the resistive material in the current-pinch region was apparent in a typical delayered device (Figure 14).

The manufacturer of the failed resistor was unknown, so it was not possible to obtain specific rating data, and manufacturers of similar-size general-purpose resistors did not give pulse ratings or thermal-mass data. One manufacturer of pulse-rated surface-mount thick-film resistors did give useful data, which may or may not apply to the failed resistor, but this information revealed that even a pulse-rated resistor of the same physical size would have been overstressed.

The mechanism that causes increasing resistance was described in an earlier section of this article. Briefly, localized heating in the current-pinch region melts the under side of the overglaze; the resulting molten material infuses grain boundaries in the resistive material, and upon cooling, locally interrupts current flow by pushing adjacent grains apart. Wet-chemical delayering removes this material, thereby allowing some of the previously separated resistive-material grains to touch again; the result is a substantial, but not complete, recovery of resistance.

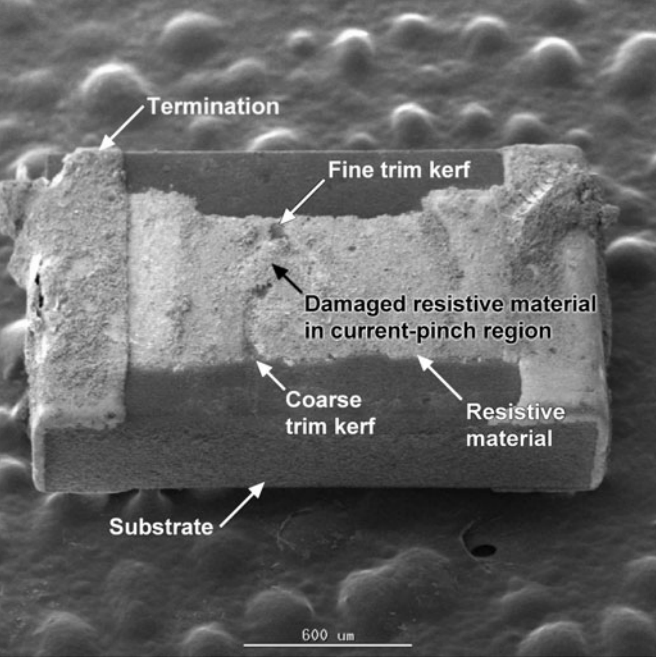


Figure 14. Scanning-electron micrograph of failed surface-mount thick-film resistor with overglaze removed

Axial-Lead Resistor

The package material on a low-value, low-inductance axial-lead resistor, which was used in a switching-power-supply current-sensing application, was cracked at one end (Figure 15). Presence of cracks in this resistor and others of the same type raised concerns about long-term reliability.

Resistances of numerous cracked samples were all within tolerance limits. X-radiography showed that the metal-foil resistance elements in the cracked devices were off center toward the cracked ends, and these defects were verified by longitudinally cross sectioning representative resistors (e.g., Figure 16).

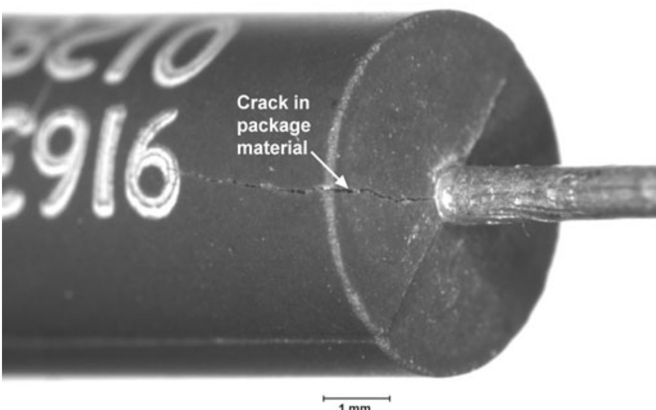


Figure 15. End of axial-lead resistor showing cracks in package material

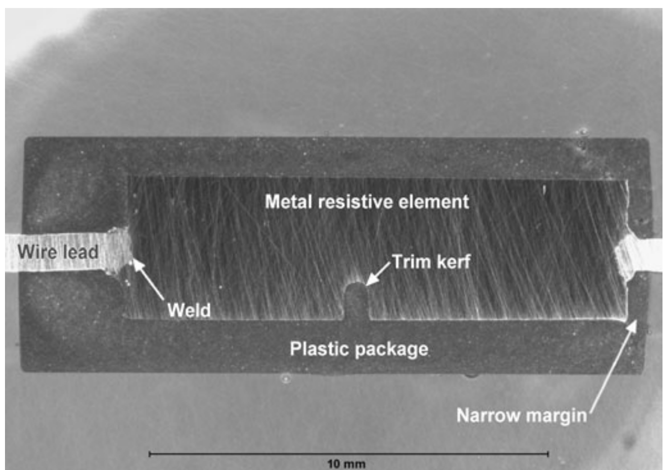


Figure 16. Longitudinal cross section through cracked axial-lead metal-foil resistor showing off-center resistive element

Because the resistive element was off center longitudinally, package material was abnormally thin, and, hence, mechanically weak, at one end. Accordingly, stresses induced during the lead-forming process caused the thin package material to crack. Of course, cracks in the package material opened paths through which potentially corrosive contaminants could reach the metal-foil resistance element, so the concern over long-term reliability was justified. It was possible to continue using the remaining stock of these resistors by employing carefully controlled lead forming. For safety, future lots of these resistors should be subjected to construction analysis until confidence in the vendor's manufacturing process has been restored.

Stacked-Film Capacitor

Normally, film capacitors, which include the stacked-film subclass, have high and stable leakage resistances. In a balanced dc-amplifier circuit that depended on these characteristics of stacked-film capacitors for stable operation, imbalance and drift were encountered frequently enough to be of concern. Meticulous cleaning of a malfunctioning circuit board would often cure the problem; however, such cleaning was labor intensive, and action of the solvent and stresses applied during brushing often damaged components on the circuit board.

Detailed inspection of a number of stacked-film capacitors mounted on failing circuit boards revealed delaminations as pointed out in Figure 17. These delaminations provided paths into which contaminants, such as solder-flux residue and moisture, could wick. These contaminants formed temperature-sensitive, moisture-sensitive, and time-variable leakage resistances in parallel with the affected capacitors, and the resulting unstable leakage resistance caused correspondingly unstable circuit operation.

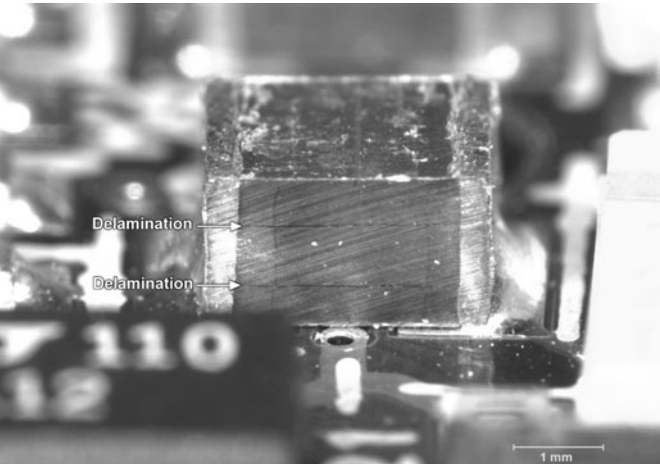


Figure 17. Delaminated stacked-film capacitor mounted on circuit board

Many new, unused stacked-film capacitors of the same type and lot also had delaminations. This observation showed that delaminations were not being caused by the pick-and-place and reflow-soldering operations of circuit-board assembly; instead, a problem in the capacitor-manufacturing process was indicated.

Fluid-Filled Metallized-Film Capacitor

Upside-Down Capacitor. A two-section fluid-filled metallized-film capacitor, which had been used with two single-phase capacitor-start, capacitor-run motors, was electrically open. A notation on the outside of the can wall stated that the capacitor was equipped with an internal protection device, and a bulged can lid indicated that this device had been actuated, which accounted for the capacitor being open. A key application factor that entered into the failure mechanism was that the capacitor had been mounted with its terminal end down.

When the capacitor was shaken, sloshing of the fluid was audible. Sloshing indicated two things: (1) the fluid had relatively low viscosity and (2) there was a lot of ullage in the can. Weighing several capacitors from the same date-code lot showed that weight of the failed device was at the lower end of the range and that the

difference between the lightest and heaviest of the available capacitors was substantial.

After the can had been opened and the sections had been withdrawn, it became immediately evident that that there was a dielectric-breakdown site near the bottom of the large section and that the bottom part of the section was swollen (Figure 18); note that the "top" of the device was the terminal end. Resistance checks showed that the large section was shorted and that the small section was not.

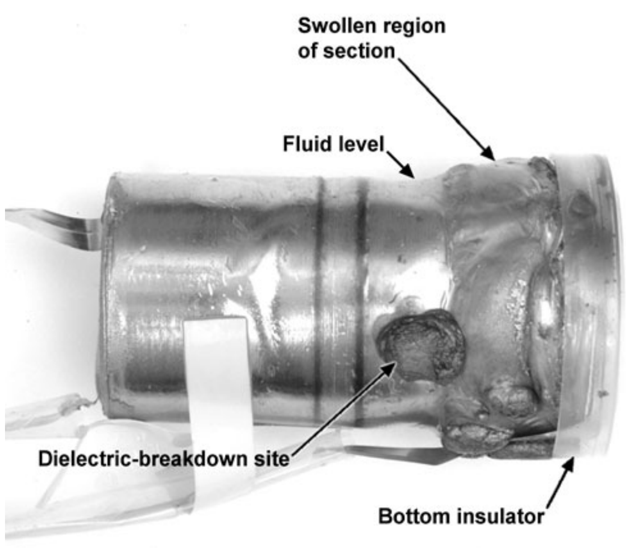


Figure 18. Failed large section from fluid-filled metallized-film capacitor that had been operated upside down

Checks of other capacitors from the same lot and same weight range indicated that fluid level in the failed capacitor had been aligned with the edge of the swollen region, as indicated in Figure 18, and that the dielectric-breakdown site straddled the fluid level. Low fluid level and upside-down mounting left a large part of the bottom section uncovered. As discussed in an earlier section of this article, partial fluid coverage causes temperature gradients that, in turn, produce stresses in the dielectric film; these stresses lead to catastrophic failure of the film. Information from the device vendor revealed that the failed capacitor had been filled by the direct-injection method, which often yields wide variations in fluid-fill level.

Horizontally Mounted Capacitor. A single-section fluid-filled metallized-film capacitor failed shorted, and the internal protection device had not opened. A hole had been burned in the wall of the plastic can, and some of the fluid had leaked out, so weighing was not applicable. Inspection of the section after it had been removed from the can revealed three dielectric-breakdown sites (Figure 19), one at the bottom-termination end and two on the side of the section.

In the application, the capacitor had been mounted with its longitudinal axis horizontal. In a similar capacitor of the same type and lot, the fluid covered only half of the section as pointed out in Figure 20. Comparison of Figures 19 and 20 shows that the three failure sites were just at the fluid surface. Again, inadequate fluid-fill level and improper mounting in the application caused failure of this direct-injection-filled capacitor.

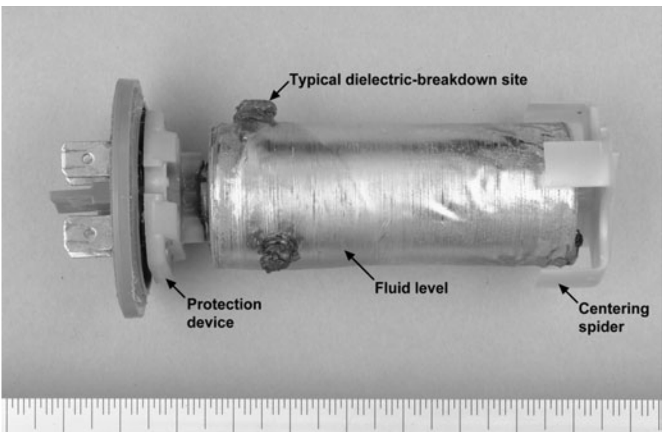


Figure 19. Failed section removed from horizontally mounted fluid-filled metallized-film capacitor

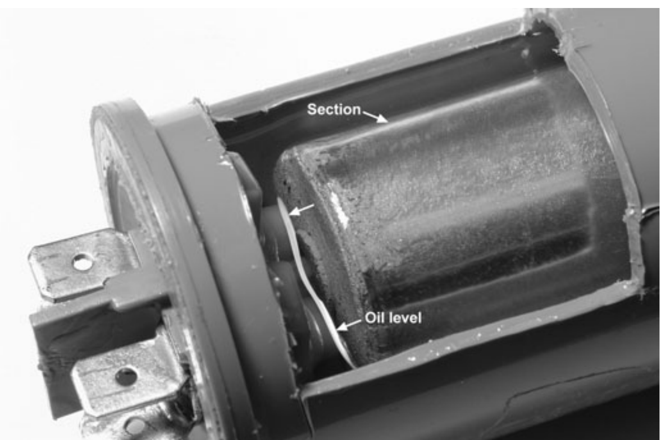


Figure 20. Fluid level in horizontally oriented metallized-film capacitor

Aluminum Electrolytic Capacitor

A large aluminum electrolytic capacitor (similar to the one illustrated in Figure 8) failed with an open circuit. There were no external physical signs of failure (e.g., bulging of the can or damage to the rupture disk). X-radiography showed that one of the termination ribbons was severed near a package terminal. Circumferentially cutting the can just below the header-support crimp and folding the header aside revealed that corrosion had consumed most of the positive-plate aluminum ribbon (Figure 21). Energy-dispersive x-ray analysis of the corrosion product did not reveal chlorine, which was originally suspected, but only a carbon-based (i.e., organic) material. The fact that corrosion affected only the positive-plate ribbon, but no other aluminum component of the capacitor, indicated that the failure mechanism was electrochemical corrosion; that is, in an electrochemical cell, metal ions move away from the positive electrode toward the negative electrode, so that the positive electrode is consumed, while the negative electrode is cathodically protected.

Normally, the electrolyte in an aluminum electrolytic capacitor does not attack aluminum, even when voltage is impressed between the device terminals; the fact that corrosion had occurred suggested presence of a foreign substance. As mentioned earlier in this article, the electrolyte usually has a fishy odor; in the present capacitor, however, the electrolyte had a very strong phenolic odor, indicating that the electrolyte had been contaminated with a corrosive derivative of phenol. The source

of this substance was the phenolic-plastic header. Apparently, composition of the header (and headers in numerous other examples of this capacitor) was not correct for use in the interior environment of an aluminum electrolytic capacitor. The electrolyte in the failed capacitor had leached a phenolic constituent from the header material, thus altering electrolyte characteristics, and electrochemical corrosion had occurred when voltage was impressed on the device. Interior inspection of unused samples of the same type of capacitor revealed evidence of beginning corrosion; thus, it is likely that corrosion started in the forming process during capacitor fabrication.

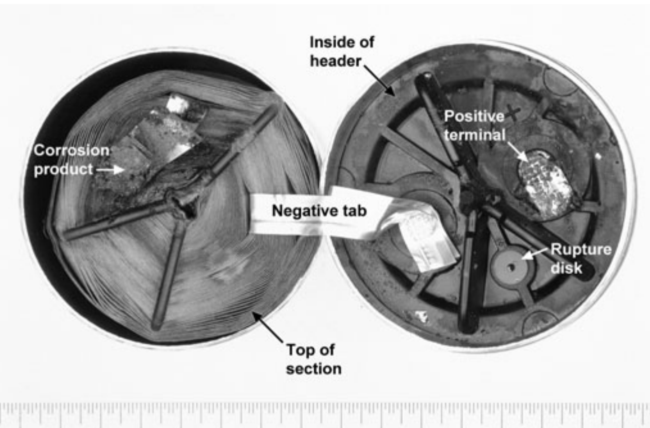


Figure 21. Opened large aluminum electrolytic capacitor showing corrosion failure of positive ribbon

Wound Components

Large Inductor. Smoke was observed coming from a physically large inductor used in a high-power application. There were signs of charring on the exterior of the winding, but inductance and quality factor of the device were normal. However, a high-voltage (i.e., hi-pot) test showed that dielectric breakdown had occurred between the winding and the core (i.e., ground). Disassembly of the inductor revealed that there were two large arc-damage sites inside the innermost layer of the winding (Figure 22).

Figure 22. Interior of winding structure on large power inductor showing arc-damage sites

Insulating systems in this inductor included the enamel on the flat wire, paper strips between layers of the winding, a paper overwrap on the winding, and a multilayer paper sleeve around the center leg of the laminated core. All of these insulating materials were brittle. The enamel on the wire would flake off when scratched with a fingernail, and the paper would crack or break when flexed; these characteristics are indicative of thermal aging of the insulation materials. The inductor had been in almost continuous service at higher-than-rated temperature for about 17 years, so thermally induced aging of the insulating materials was to be expected. Arcing, initiated by the normal high operating voltage with respect to ground, had occurred in areas where the insulation between the inside of the winding and the core had deteriorated to powder consistency. It was concluded that the inductor had reached the normal end of its life.

Power Transformer. The circuit breaker serving a 7.5-kVA transformer tripped even though the load current was normal; resetting the circuit breaker produced an immediate trip. This transformer had two primary windings that could be connected for either 240-V or 480-V service; inductance and quality factor of one of these primary windings were low, while corresponding values for the other primary winding were normal. Further, hi-pot testing showed that there were no leakage paths between windings and the core nor among windings.

To gain access to the windings, it was necessary to remove the laminated core. Under one of the arms of the core, a burned spot was observed on one of the primary windings. Removal of the carbonized insulating material from this area exposed a small metallic bridge (Figure 23) between two adjacent layers of the winding. It was evident that a layer-to-layer arc had occurred. Of course, the multiple shorted turns caused inductance and quality factor of the winding to be low. Unrolling the winding did not reveal any additional short circuits, so it was apparent that the metallic bridge pointed out in Figure 23 was the only failure site.

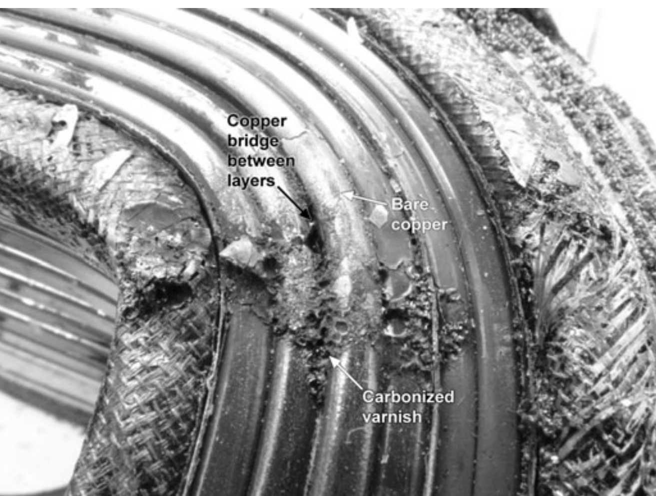


Figure 23. Metallic bridge between two adjacent layers of power-transformer primary winding

Insulation on the flat wire of the winding was brittle, and near the failure site, some of the insulation was missing as pointed out in Figure 23. Moreover, the paper insulation between winding layers was brittle, and it had crumbled at its edges (i.e., in the vicinity of the failure site). The failed transformer had been in service for many years, and its operating temperature had been fairly high; accordingly, it was concluded that the insulation system had deteriorated because of thermally induced aging and that the transformer had reached the end of its normal life. It appeared that

the arc that shorted two winding layers together was probably triggered by a power-line transient.

Switch Contact Block

One of the normally closed contacts in a modular-switch contact block failed to close when the actuator was released. Although the contact-button surfaces appeared in good condition, buttons in the failing contact set were not properly aligned. The cause of misalignment was a crack (Figure 24) in the contact-block cover.

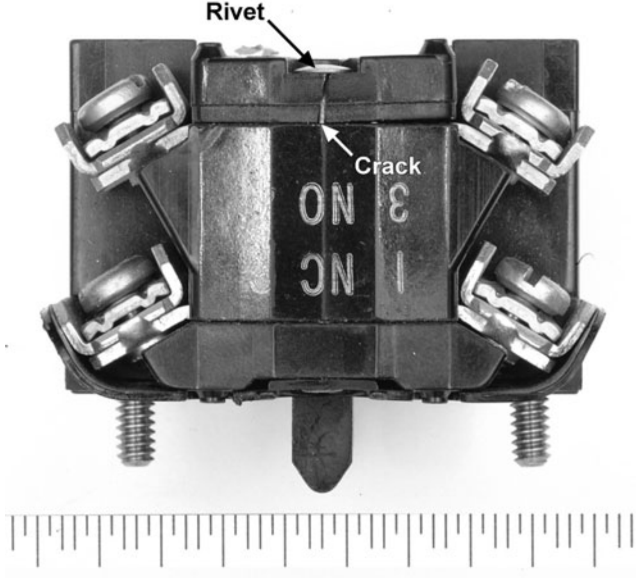


Figure 24. Cracked cover on modular contact-block assembly

Detailed examination of the failed contact block and a good one of the same type showed that the rivet in the failed block was so tight that it could not be rotated about its longitudinal axis; in contrast, the rivet in the good block was snug, but not so tight that it could not be rotated. This observation suggested that the tight rivet in the failed block was stressing the cracked cover in a mechanically weak area. To test this theory, the surface of the contact-block body against which the cracked cover had been mounted was checked for flatness by clamping it against a flat-ground steel plate and looking at the interface with back lighting. As apparent in Figure 25, the body surface had concave curvature. Because of this curvature, outside edges of the cover were supported by high points on the body, and the center of the cover was bent downward into the concave depression on the block body by tension in the rivet. As a result, the bottom surface of the cover was placed in tension, and a crack formed in the relatively brittle material to relieve the stress. It appeared that excessive tightness of the rivet was the cause of failure.

CONCLUSIONS

Passive-component failures, particularly repetitious failures, cannot be ignored. There are many consequences that may not be immediately apparent when such simple devices fail. For example, production yield may be adversely affected, or consumer injury or property damage may result from a failure. Costs (real and intangible) of these consequences can greatly exceed the cost of a failure analysis and the ensuing corrective action. Thus, it is prudent to make frequent use of failure analysis as a quality- and safety-enhancement tool that helps insure product integrity.

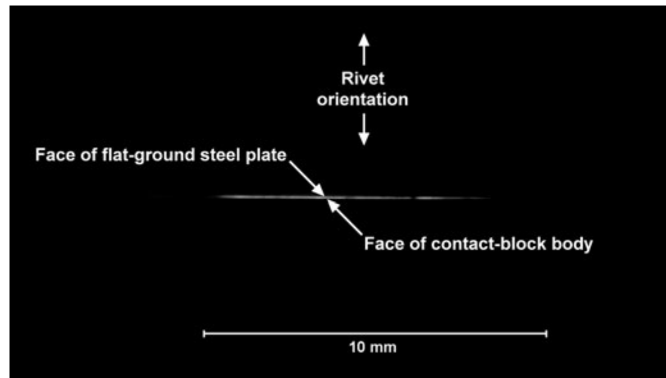


Figure 25. Back-lighted interface between contact-block body surface and flat-ground steel plate

REFERENCES

1. M. Johnson and S. Smith, "Failure Modes and Mechanisms of Non-Semiconductor Electronic Components," ***Microelectronic Failure Analysis Desk Reference***, 3rd edition, T. Lee and S. Pabbisetty, editors, ASM International, Materials Park, OH, 1993, pp. 303-320.

2. S. Silvus, "Failure Analysis of Passive Components," ***Microelectronic Failure Analysis Desk Reference***, 4th edition, R. Ross, C. Boit, and D. Staab, editors, ASM International, Materials Park, OH, 1999, pp. 379-399.

3. S. Axtell, "Failure of Thick-Film Resistors in Sulfur-Containing Environments," ***Microelectronic Failure Analysis Desk Reference 2002 Supplement***, T. Kane, editor, Electronic Device Failure Analysis Society, ASM International, Materials Park, OH, November 2002, pp. 161-173.

4. S. Silvus, "The Wonders and Wanderings of Silver," ***Electronic Device Failure Analysis***, Electronic Device Failure Analysis Society, ASM International, Materials Park, OH, vol. 3, no. 4, November 2001, pp. 37-39.

5. S. Silvus, "Failures in Relay and Solenoid-Valve Coils," ***Electronic Device Failure Analysis News***, Electronic Device Failure Analysis Society, ASM International, Materials Park, OH, vol. 3, no. 2, May 2001, pp. 4-8.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Reliability and Quality Basics for Failure Analysts

Steven Hoffman (Theoretical Materials)
Chris Henderson (Semitracks Inc.)

Abstract

Although failure analysis serves to determine the cause and impact of failures with an eye to appropriate corrective action, it is worth noting that this activity is interdependent on many other parallel disciplines such as process design, component design, and (notably) quality and reliability engineering.

In order to serve as a very brief (and arguably incomplete) guide to these disciplines, this chapter surveys both basic quality and basic reliability concepts as an introduction to the failure analysis professional.

Introduction: Quality versus Reliability

The distinction between quality and reliability can be somewhat tenuous, but, for the purposes of this summary (and context) is considered to be a matter of *when* more than *what*. Quality, then, is a measure of adherence to specification immediately following manufacture (designated time-zero or $\mathbf{T_0}$), where reliability is an equivalent measure at any time later in the product (or component) lifetime.

One example of the relationship between these two relates to defects. Quality ($\mathbf{T_0}$) defects are those that have escaped final test, where "reliability defects" may manifest themselves over time as reliability fails.

In fact, quality and reliability defect-driven measures of a group of semiconductor devices have been shown to be closely related to one another, with a high correlation between initial defects and later failures occurring during burn-in testing [1].

Quality Concepts

Although quality in a broad sense may be viewed as conformance to customer expectations (which, of course, also subsumes reliability as defined here), for the purposes of microelectronics manufacturing it may more narrowly (and practically) be considered conformance to specifications following manufacture, and is typically specified as the fraction of units shipped that are established to meet specification(s).

The quality organization in semiconductor manufacturing is tasked with monitoring product quality through the reduction of defects, and continuing to improve quality through improving the manufacturing process [1].

Key quality engineering (QE) concepts/measurements summarized in this discussion include process variation and control, Acceptable Quality Level, and Average Outgoing Quality Level. Implicit in this discipline is the employment of statistics, handmaiden to QE.

Process Variation

All processes (and measurements) exhibit some level of variation. This is abundantly true in semiconductor fabrication since it comprises numerous steps in the fabrication sequence, and meticulous control of many process parameters. Slight variations of these parameters can drastically (and negatively) affect product yield [1].

Process Control

Statistical Process Control (SPC) is a systematic way of controlling quality through the use of statistical process programs or "charts" in a manufacturing process, and is utilized to monitor the consistency of such processes. For a microelectronics fabrication facility, an extensive number of such charts are utilized covering all parts of the fabrication process [2]. Charts are maintained on a myriad of variables including water contamination levels, equipment flow rates, temperatures, deposition rate(s), etch rate(s), alignment tolerances, critical dimensions of structures, and many more. Figure 1 is an example of a statistical process control chart. Note that this example chart summarizes the deposition rate mean and standard deviation and clearly identifies process excursions.

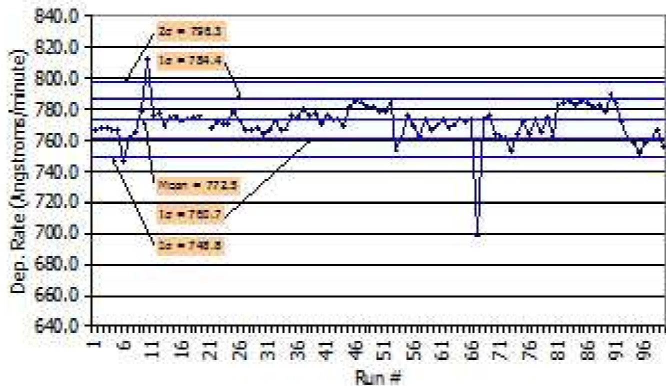


Figure 1: SPC chart example tracking TEOS deposition rate vs. run on a Lam Integrity deposition tool [1].

SPC methodology employs statistical techniques to measure and evaluate process variation with a goal of maintaining processes to specific target specifications.

Statistical Quality Control (SQC) refers to the use of statistical techniques to measure and improve product

quality, and includes SPC as one component in its toolbag. Other techniques within the umbrella of SQC include experiment design (covered in following paragraphs), process capability analysis, sampling plans, and other tools.

Acceptable Quality Level (AQL)

AQL (Acceptable Quality Level) specified by a sampling plan is designed to assess the quality level required (by the plan), and specifies the percent defective units accepted 95% of the time. Thus, lots greater than or equal to AQL will be accepted at least 95% of the time and rejected at most 5% of the time.

In the case of microelectronics fabrication, AQL is typically specified at 0.25% (or one in 400). Since semiconductor defect rates tend to be significantly less than this target, excessively large sample sizes would be required. As a result, this measure is mainly effective in detecting significant quality deviations (typically caused by systematic issues) [3].

Average Outgoing Quality Limit (AOQL)

An Average Outgoing Quality (AOQ) curve indicates the average outgoing quality in terms of a defect measure (vertical axis) as a function of incoming quality (horizontal axis) as demonstrated in Figure 2.

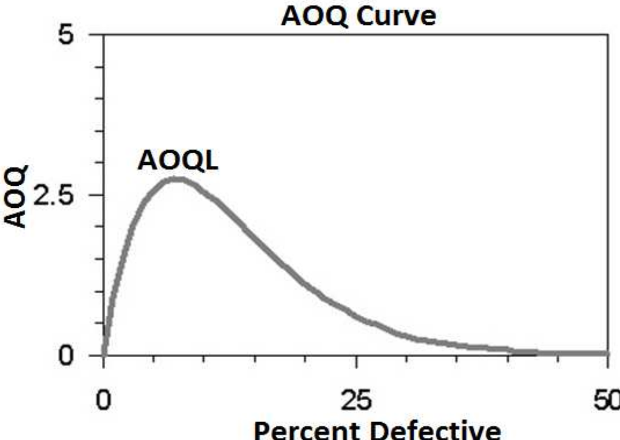


Figure 2: AOQ curve indicating AOQL at relative maximum [4].

The AOQL (Average Outgoing Quality Limit) of a sampling plan is the relative maximum (and therefore worst possible value) of the AOQ curve (assuming rejected lots are 100% inspected and that the inspection process is at least 90% effective). The outgoing (shipped) defect count should not exceed the AOQL on average, although instantaneous AOQL values may be in excess of this specification [4].

For microelectronics fabrication, typical AOQL values vary between 10 and 100ppm.

Burgess [3] notes that most product manufacturers accept and release components directly to the production floor without verification of AOQL, with few failures expected during the production process. Even a small number of failures from a shipment of a few thousand devices will far exceed a 100 ppm quality level. Whether or not such failures can be verified, analysis is require to explain the cause of the failures (such as by a change in testing, a problematic stress, sensitivity in the application, or a degradation of devices).

Experiment Design and Taguchi Methodology

In order to obtain systematic (as opposed to random) data, a set of experiments may be planned where parameters of interest are varied over specified ranges. However, the complexity of such an experimental matrix may be onerous, and such an exercise will often be prohibitive in both cost and time, motivating the modification to a subset of the "full" matrix to save on time and money.

This introduces complications, statistically, and may not even be an effective approach since effects of parameter changes may not be understood. The analysis is certainly not straightforward, and effects of various parameters on the observed data are often obscured. At best, this truncated approach is far from optimal and may bear no relevance to an effective parameter set [5].

Design of Experiments

The concept of experiment design (DoE: Design of Experiments) relates to an experimental design (in a general sense) over parameter variation. In such an experiment, the intent is to observe effects as one or more process variables are changed, DoE is a statistical methodology that efficiently allows planning experiments so that resulting data can be profitably analyzed to yield valid (and objective) conclusions.

DoE begins with developing objectives of an experiment and then selecting appropriate process factors (process models that include input factors and measurable output responses) for the evaluation. The actual experimental design constitutes a detailed experimental plan and should contain:

1. A clear statement of the experimental goal/problem,
2. A determination of the outcomes (response variables) to be studied,
3. Specification of control factors to be included/varied to determine their effects on the response variable(s),
4. Specification of control factor levels to be tested,
5. Identification of uncontrollable factors that might affect the response variable(s), and that will be recorded or measured for each experimental unit,
6. Identification of controllable factors not to be included in the experiment, and the level (or constant value) to which they are to be held fixed for the experiment,
7. A statement of how to conduct the experiment,
8. Description of data analysis plans.

An effective experimental design serves to maximize the amount of information obtained from the experimental effort expended. Also, the experiment may have to account for uncontrolled factors such as different operators or environment changes [2].

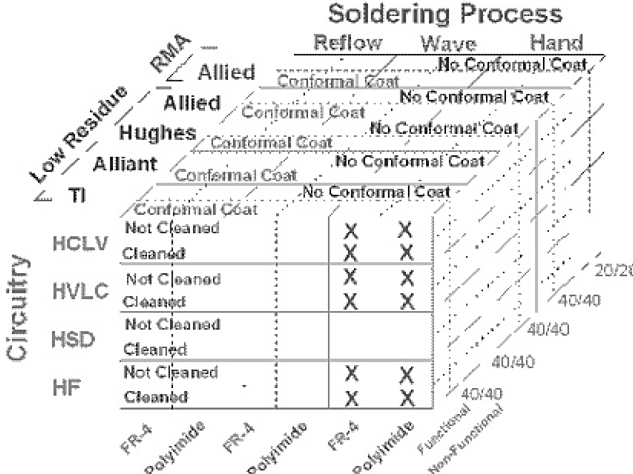


Figure 3: Example experimental design to evaluate and qualify low-residue soldering for military and commercial applications [2].

Taguchi Methodology

The Taguchi method is one DoE example, and was developed by Dr. Taguchi of Nippon Telephones and Telegraph Company (Japan). This approach is based upon Orthogonal Array experiments yielding a much reduced variance for the experiment with optimum settings of control parameters. The effect is a marriage of DoE with control parameter optimization to obtain optimum results using a set of "well balanced" (minimum) experiments [5]. This method is not without its critics, but has proved to be effective in many applications.

Reliability Engineering and Accelerated Life Testing

One of the key challenges in the semiconductor industry is determining the expected lifetime of a component. Numerous failure mechanisms can be involved, complicating the overall calculation, and many failure mechanisms can take months or years to manifest themselves in an end-user application. Furthermore, most applications require very low failure rates. This combination of low failure rates and long times requires a calculated approach to reliability estimation [1].

Purpose(s) of Reliability Engineering

Reliability engineering fundamentally addresses the question of how long will a given component (or system) function correctly in the target application. Answering this question is not a trivial matter since dozens of failure mechanisms must be accounted for, some possibly inconsequential, but others greatly affecting reliability. In addition, there is the possibility that a new mechanism might appear, also affecting reliability.

It is the reliability engineer's job to understand the application environment, understand potential failure mechanisms, devise a plan to evaluate the mechanisms, determine if any new mechanisms might exist, and develop an overall reliability model for the component so that decisions (and tradeoffs) can be evaluated with regard to warranty periods, failure rates, mitigation at the system level, and use conditions. These tradeoffs can also impact the component design and package. In general, the accuracy of the solution to this problem is limited by classic constraints of cost and time, but the engineer can make tradeoffs to provide meaningful guidance to the system developers within a given budget and timeframe. This requires knowledge of statistics, failure mechanisms (and models), and accelerated testing.

Typically, the reliability engineer will begin by gathering input from the customer on use conditions of the component, taking into account all environmental conditions. This may include factors such as operating temperature and voltage, stability of temperature and humidity, potential shock environments, electrical transients, and noise. In addition, individual applications may well have unique profiles. For example, an integrated circuit in a cell phone may need to operate in a moderately wide temperature and humidity environment, but the critical concern could be shock from accidental drops. Another example might be a military aircraft application where the component is expected to work at very low temperatures in the Arctic, but also very high temperatures in the Middle East deserts. Still another might be the stable environment of a server farm for a cloud computing application.

The engineer must be familiar with failure mechanisms that are known to impact the component, such as Negative Bias Temperature Instability, electromigration, package delamination, and corrosion. Since the manifestation and effect of these mechanisms vary from one technology to the next, the mechanism will need to be understood in detail for a particular process technology, and characterized for the specific component application.

These mechanisms may be studied using accelerated testing techniques including thermal cycling, shock, temperature/humidity/voltage, and other environmental parameters. Failures are then characterized and mapped into use condition scenarios to determine their projected frequency of occurrence in the anticipated real-world application. For yet-to-be-discovered phenomena, the engineer may use a life test to look for other failure mechanisms. If any such are found, they can be analyzed and characterized, and then a model of the failure mechanism developed to project from accelerated stress conditions to use conditions in order to determine failure rates and times to a particular percent failure.

Once individual failure mechanisms are characterized, the results can be combined using various approaches to determine an overall reliability/failure rate for the component. These results may then be presented to the customer, and supplier and customer can then work together to develop an approach to ensure desired reliability for the intended application [1].

Statistical Distributions

A brief introduction to statistical distributions is required to understand failure rates. More extensive discussions of statistical distributions are described by Nelson [6], Lawless [7], McPherson [8] and Tobias and Trinidade [9].

Four basic distribution functions useful in reliability engineering include: Normal, Log-Normal, Exponential, and Weibull.

The Normal Distribution

The Normal distribution takes on the classic bell-shaped curve which is defined over all values of x. The equation for the probability density function (PDF) can be applied to wafer-level or package-level cases and is given by,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

where μ is the mean and σ is the standard deviation.

This distribution is symmetrical about μ, and σ is a scale parameter that indicates how closely packed the density function is. Figure 4 depicts the shape of the curve.

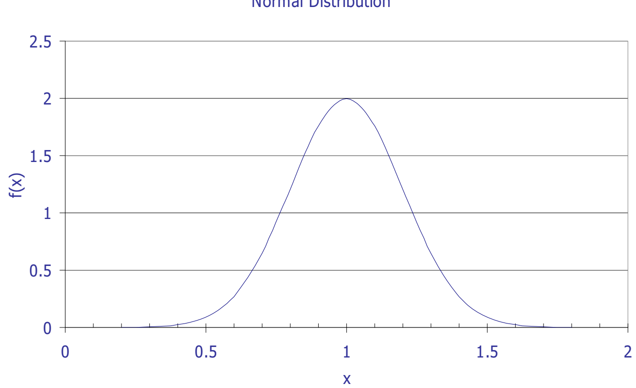


Figure 4: Normal distribution curve for μ = 1 and σ = 0.2.

Once identified, a series of tests utilizing this distribution can be designed to characterize these mechanisms of interest. In order to produce a statistically significant number of failures in a reasonable length of time, accelerated test methods involving high temperatures, high-humidity levels, higher than normal voltage levels and thermal conditions are utilized,

The Lognormal Distribution

The Lognormal distribution is similar to the Normal distribution except that it uses a logarithm value in place of x and m in the equation. The PDF is given by,

$$f(x) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-(\ln(t)-\ln T_{50})^2/2\sigma^2},$$

where T_{50} is the median time to fail of a given population (viz., the 50% point). In the Lognormal distribution, σ is better thought of as a shape parameter rather than a standard deviation. The Lognormal distribution is used widely in component reliability, and is used extensively to characterize long-term wearout due to an individual failure mechanism. Figure 5 depicts the shape of the curve.

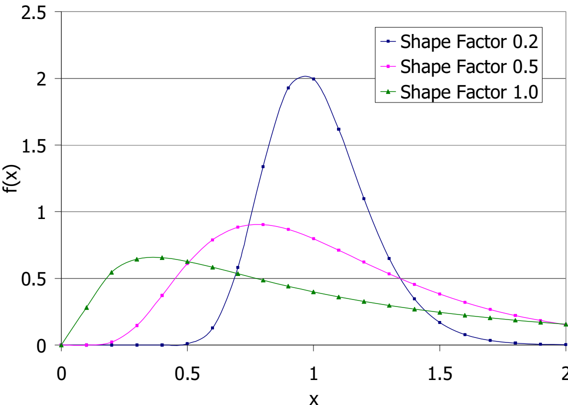


Figure 5: The Lognormal distribution curve for several shape factors.

The Exponential Distribution

The Exponential Distribution is commonly used for life expectancy. The equation is,

$$f(t) = \lambda e^{-\lambda t}.$$

In the equation above, λ is the single unknown parameter that defines the Exponential Distribution. The equation is a simple, exponentially decaying slope (Figure 6).

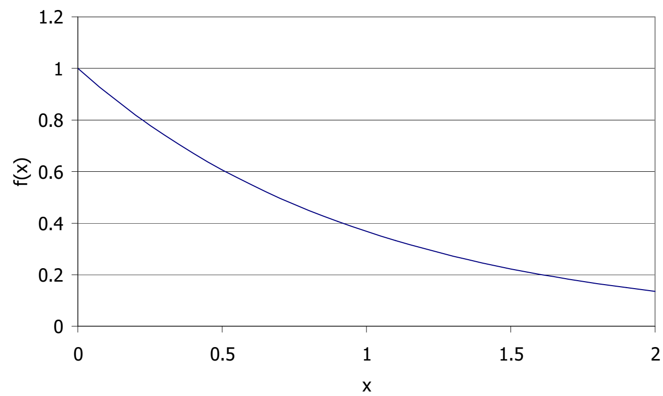


Figure 6: Exponential Distribution where λ = 1.

The Weibull Distribution

The Weibull Distribution [10] is one of the most versatile distributions for life testing. It was developed by a Swedish scientist named Waloddi Weibull to describe the breaking strength and life properties of ball bearings. The Weibull distribution is characterized by three parameters:

m: shape parameter or Weibull slope,
c: characteristic lifetime,
g: location parameter or minimum life.

Normally, a two parameter version of the Weibull distribution (where *g* is eliminated) is used. This collapses to the following probability density function for the Weibull distribution:

$$f(t)=\frac{m}{t}\left(\frac{t}{c}\right)^{m}e^{-\left(t/c\right)^{m}}.$$

The shape parameter m plays a major role in appearance of the curve. The Weibull distribution is illustrated in Figure 7.

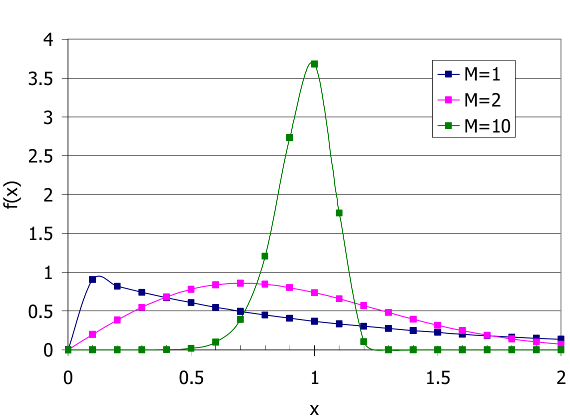


Figure 7: Weibull Distribution Probability Density Function.

The Extreme Value Distribution (EVD) is a derivative version of the Weibull distribution that is sometimes used in reliability calculations. Gumbel [11] provides more information on the EVD.

Acceleration Models

In order to bridge to a failure rate under use conditions using experimental data from accelerated conditions, an acceleration model is required. Reliability engineers use three fundamental models for these calculations: the Arrhenius model, the Eyring model, and the Power Law model [8].

Arrhenius Model

The Arrhenius Model is a pure thermal acceleration model derived from basic thermodynamics and developed by Svante Arrhenius in the late 1800s. The basic equation for time-to-failure of a given component is,

$$TTF=C_{o}e^{\frac{-E_{A}}{kT}},$$

where C_o is a constant, E_A is the activation energy, k is Boltzmann's constant, and T is the temperature (degrees Kelvin).

Some failure mechanisms exhibiting a purely thermal behavior include intermetallic formations such as gold-aluminum and copper-tin, and data retention in non-volatile memories that are not cycled. When E_A is large, the mechanism has a strong dependence on temperature; when E_A is small (< 0.5eV), the mechanism has a weak dependence on temperature.

Eyring Model

The Eyring Model, developed by Henry Eyring in the 1930s, is sometimes referred to as transition state theory, and takes the form:

$$TTF=C_{o}F^{-N}e^{\frac{-E_{A}}{kT}}, \text{ or } TTF=C_{o}e^{-N}e^{\frac{-E_{A}}{kT}},$$

where C_o is a constant, F is some driving force, N is an exponent associated with the driving force, E_A is the activation energy, k is Boltzmann's constant, and T is the temperature (Kelvin). A hallmark of the Eyring model is that it includes temperature and one or more other driving factors such as voltage, current, or humidity. Sometimes the Eyring equation will be written with ΔH in place of E_A, where ΔH indicates the enthalpy of activation.

The Eyring Model is used to model a wide variety of failure mechanisms including dielectric breakdown of thick oxides (>50Å), electromigration, ion migration, and corrosion. Several common variants of the Eyring model include the Peck Model (which adds a term for humidity and sometimes a term for voltage), the E-model (which includes a term for electric field strength), and Black's model (which includes a term for current).

Power Law Model

The Power Law Model is a generic model that works well for natural phenomena not easily modeled by other means. Usually, the Power Law Model works best when a number of underlying factors contribute to the behavior of the system. The Power Law Model is given by:

$$TTF=C_{0}F^{-N},$$

where C_o is a constant, F is some driving force, and N is an exponent associated with the driving force. Two mechanisms that lend themselves well to the Power Law model are crack propagation and thin dielectric (>50Å) breakdown. In fact, the crack propagation model is a specific form of the Power Law Model called the Coffin-Manson model, where the driving force F is a temperature range associated with thermal cycling.

Failure Rates/Mechanisms

The typical shape of a failure rate curve is known as the "bathtub" curve (Figure 8) and is broken into three separate phases or periods.

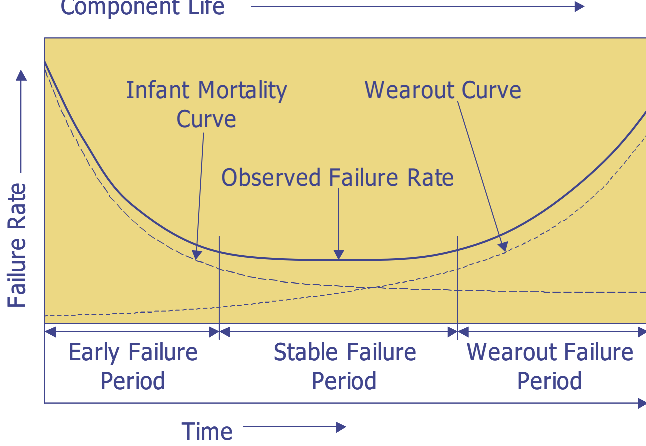


Figure 8: Classic Bathtub curve for failure rates.

The first (Early Failure) phase of this curve is approximately an exponentially decaying function with respect to time, and is dominated by infant mortality

failures (typically failures due to defects on the semiconductor device such as contamination or misprocessing). These issues are normally best addressed by yield improvement activities within the chip fabrication and assembly processes.

The final or third (Wearout Failure or End-Of-Life) phase of the life of a group of components is dominated by wearout failure mechanisms. This phase of the curve may have an Exponential, Weibull, Normal or Log-Normal distribution (The peak and decline of the wearout phase are not shown on the bathtub curve). As noted earlier, the wearout curves of contemporary IC technologies are moving toward the left because of limited margins.

The Early Failure and Wearout Failure curves can be combined to give an indication of the total failure rate. The resulting curve thus has a shape similar to that of a bathtub viewed in cross section, and sums to represent the second (Stable Failure) phase of the curve.

Incidental to this discussion, one possible component of this curve is the freak failure rate. Usually, so-called "freak" failure rates refer to failure modes that are instigated from outside of the component itself (such as electrical overstress, electrostatic discharge, mechanical shock, or single event upset). These are random in nature, but still need to be accounted for, and are often modeled as a constant failure rate throughout the component lifetime.

Burn-In

Burn-in has been one of the most common reliability tests. Historically, burn-in was used as a screen to eliminate potential reliability failures from the population since, for many years, semiconductor manufacturing was plagued with defects. These defects could sometimes be removed from the general population by stressing the devices at high temperature and voltage for some length of time [12-19].

While burn-in can be an effective method for discovering potential reliability defects, it is expensive and may consume a substantial portion of the component's useful life, especially since many contemporary devices tend to have very few defects. If this is the case, then a burn-in screen will not weed out defective components, but will instead use up some of the useful life of the component.

This might be analogous to an auto dealership driving a car around for 100,000 miles, finding no problems, and then selling the car as a new car free from defects. It may be free from defects, but a significant portion of the life has been used up.

Burn-in is now more often employed to understand the behavior of various mechanisms. Most burn-in testing involves elevated temperatures since many failure mechanisms occur more quickly at elevated temperatures. A partial list includes: TDDB, stress voiding, electromigration, ionic contamination, thermo-mechanical stress, chemical corrosion, and moisture. The rate of these mechanisms is exponentially dependent on temperature, indicating that elevated temperatures can produce failure in a fraction of the time than would occur at room temperatures. There are several limitations to how high the temperature can go and therefore limits on the acceleration [1, 20-23].

Conclusion/Afterword

The ever-increasing complexity of the semiconductor fabrication process demands a commensurate increase in product team member collaboration, in turn requiring an understanding of related disciplines. To this end, it is important that failure analysis personnel understand basic concepts of other activities supporting the microelectronics development and fabrication processes. It is the intent of the authors that this discussion may facilitate that collaboration.

References

[1] C. Henderson, "Quality and Reliability", *21st Century Product Analysis (Chapter 17)*, 1-20 (2001)

[2] Staff, "Statistical Process Control/Design of Experiments", *Center for System Reliability (CSR)/Sandia National Laboratories*, (2010) http://reliability.sandia.gov/Manuf_Statistics/Statistical_Process_Control/statistical_process_control.html

[3] D. Burgess, "Reliability and Quality Concepts for Failure Analysts," *Microelectronics Failure analysis: Desk Reference Fifth Edition,* 672-675 (2004)

[4] Taylor Enterprises Staff, "AOQL - Average Outgoing Quality Limit", *Sampling Plan Analyzer Help System*, http://www.variation.com/spa/help/hs103.htm

[5] P. Ross, *Taguchi Techniques for Quality Engineering, 2nd Ed*., McGraw Hill, (1996)

[6] W. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*, Wiley Series in Probability and Statistics, (1990)

[7] J. Lawless, *Statistical Models and Methods for Lifetime Da*ta, Wiley Series in Probability and Mathematical Statistics (1982)

[8] J. McPherson, *Reliability Physics and Engineering – Time-To-Failure Modeling*, Springer, (2010).

[9] P. Tobias and D. Trinidade, *Applied Reliability: Second Edition*, Chapman and Hall/CRC, (1995).

[10] W. Weibull, "A Statistical Distribution Function of Wide Applicability," Journal of Applied Mechanics, Vol. 18, pp. 293-297, (1951).

[11] E. Gumbel, *Statistical Theory of Extreme Values and some Practical Applications*, U.S. Govt. Printing Office, Washington D.C. (1954)

[12] J. Black, "Electromigration Failure Modes in Aluminum Metallization for Semiconductor Devices," *Proc. IEEE, Vol. 57, No. 9*, pp. 1587-1593 (1969)

[13] D. Danielson, et al., "HAST Applications: Acceleration Factors and Results for VLSI Components," *Proceedings of the International Reliability Physics Symposium*, pp. 114 – 121, (1989).

[14] H. Huston and C. Clarke, "Reliability Defect Detection and Screening During Processing - Theory and Implementation," *Proceedings of the International Reliability Physics Symposium*, pp. 268-275, (1992).

[15] M. Ruprecht, et al., "Is Product Screen Enough to Guarantee Low Failure Rate for the Customer?" *Proceedings of the International Reliability Physics Symposium*, pp. 12-16, (2001).

[16] D. Trinidade, "Can Burn-In Screen Wearout Mechanisms?: Reliability Modeling of Defective Subpopulations - A Case Study," *Proceedings of the Reliability Physics Symposium*, pp. 260-263, (1991)

[17] T. Turner, "A Step-By-Step Method for Elimination of Burn-In as a Necessary Screen," *Proceedings of the Integrated Reliability Workshop*, pp. 82-86, (1996).

[18] J. van der Pol, et al., "Impact of Screening of Latent Defects at Electrical Test on the Yield-Reliability Relation and Application to Burn-In Elimination," *Proceedings of the International Reliability Physics Symposium*, pp. 370-377, (1998).

[19] R. Vollertsen, "Burn-In," *Proceedings of the Integrated Reliability Workshop*, pp. 167-173, (1999).

[20] A. Christou, *Integrating Reliability into Microelectronics Manufacturing*, John Wiley and Sons, Inc., pp. 4-11, (1994).

[21] G. Gottlieb, "An Accelerated Testing Technique for Plastic Packaged Devices Using a Sequential Combination of Pressure Cooker and 85/85 (PCTH)," *Proceedings of the International Test Conference*, pp. 287-298 (1982).

[22] F. Jensen and N. Petersen, *Burn-in: An Engineering Approach to the Design and Analysis of Burn-in Procedures*, John Wiley and Sons Inc., (1999).

[23] D. Peck, "Comprehensive Model of Humidity Testing Correlation," *Proceedings of the International Reliability Physics Symposium*, pp. 44-50, (1986).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Electronics and Failure Analysis

Jerry Soden, Jaume Segura†, and Charles F. Hawkins††

Sandia National Labs, Albuquerque, New Mexico USA

†University of Balearic Islands, Spain

††University of New Mexico

Albuquerque, New Mexico USA

Sodenjm@sandia.gov

1. Introduction

Five years ago, we described our knowledge of CMOS IC defects, and how they manifested as electronic failures. The emphasis was on bridge and open defects with a short description of parametric failures. Now, with technology scaling entering the nanometer regime, parametric failures are a major concern often presenting difficult detection and fail site location. We have thus expanded the section on parametric failures showing the reasons for their increased significance in advanced process technologies, and the failure analysis tools available to face the problem.

In the previous edition of the Microelectronics Desk Reference, we introduced the basic operation of the MOSFET transistor using model equations from long channel transistors. We repeat that work, but expand a section on short channel transistor modeling, and the difficulties in its manual problem solving. Our descriptions emphasize this manual approach with its accompanying quick insights into transistor circuit behavior.

Electronics spans a number of devices, their configurations, and properties. A challenge is to identify those electronic subjects essential for failure analysis. Transistor circuits with and without resistors are one object of our study. A theme is that failure analysts deal with relatively simple circuits, but go deep in understanding their operation. Logic circuits must be understood, but when a defective circuit gives the final electronic clues to a defect, they are typically analog. An expanded treatment of all these topics can be found in reference [1].

Root cause corrective action requires localization of the failure mechanism and electrical characterization of the circuit failure mode. Therefore, we need a reflexive knowledge of how CMOS elements respond to defects. Often circuit depth is lacking so this paper also targets failure analysts who need electronic knowledge, but who may have degree backgrounds in physics, chemistry, chemical engineering, materials science, or biology. It cannot replace a course in electronics, but it puts that knowledge in perspective and provides direction.

This article reviews normal transistor operation and then relates it to electronic behavior in the presence of defects, such as bridges and opens. These electronic principles are then applied to an inexpensive CMOS failure analysis technique using a power supply signature analysis.

2. MOSFET Operation and Three Bias States

Figure 1 shows an *n*-channel transistor cross-section with heavily n^+ doped drain and source regions and a *p*-well. When the gate voltage (V_G) is zero and the source and drain are grounded, only a few thermally generated free carriers exist in the *p*-well and the transistor is in the *off-state*. Effectively, no drain current exists even when a voltage is across the drain-source.

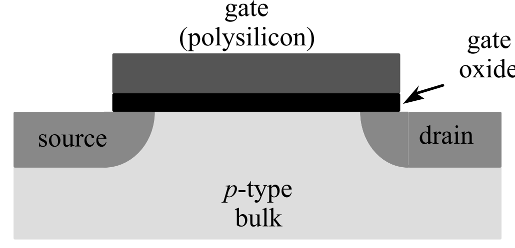


Fig. 1. *n*-channel transistor cross section.

When V_G is much larger than the drain (V_D) source (V_S), and bulk (V_B) voltages, then electrons (minority carriers) from the *p*-well are attracted to the thin oxide interface (hatched area, Fig. 2a). This second important bias state, called the *non-saturated state,* exists when a continuous electron inversion region connects the source and drain. The non-saturated state is also called the ohmic or linear state. If the drain has a positive voltage, then the inverted free electrons drift due to the electric field forming drain current. V_{GS} (gate to source voltage) has a minimum value necessary to sustain minority carrier inversion and that V_{GS} is defined as the gate threshold voltage V_t. The non-saturated and off-states exist in CMOS logic circuits when the clock pulse is off, and all voltages have settled to their quiescent values.

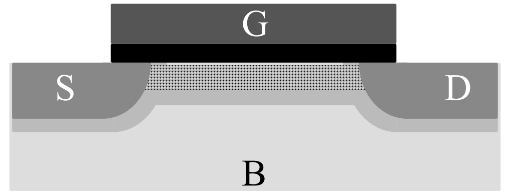


Fig. 2a. Carrier inversion in non-saturated bias state.

The *saturated state* occurs when V_D approaches V_G. If the drain voltage in Fig. 2a is increased, then the difference between V_G and V_D diminishes in the oxide region near the drain. At some point, the local gate to drain voltage across the oxide drops below V_t and no inverted carriers can exist in this local channel region. Since the n^+-doped drain has a positive voltage with respect to the *p*-well, a reversed bias *pn* junction exists here between the drain and *p*-well. Fig. 2b shows this state when $V_{GS} > V_t$, but $V_D > V_G - V_t$

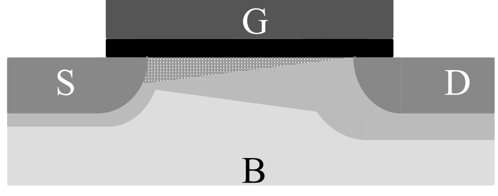


Fig. 2b. Carrier inversion in saturated bias state

The location where the drain to *p*-well *pn* junction touches the inversion layer is called the pinchoff point (L_P in Fig. 2b) and to the right of L_p is the pinchoff or depletion region. The existence of a pinchoff region puts the transistor in the saturated bias state. A high electric field exists between L_P and the drain with fixed positive charges in the drain and fixed negative charges in the substrate (*p*-well).

Surprisingly, the reverse biased drain-substrate *pn* junction does not prevent charge flow. Many inverted electrons in Fig. 2b are caught in the high electric depletion field and accelerated toward the drain. These electrons exit the drain terminal from the high impedance of a reverse biased *pn* junction.

p-channel transistors use gate voltages that are negative with respect to the source and bulk to generate minority carrier (holes) inversion in an *n*-well. The negative gate voltage draws holes from the *n*-well to the thin oxide surface if the gate voltage is greater than the threshold voltage. The *n*- and *p*-channel transistors have opposite terminal polarities for normal operation. The terminology and polarity conventions of both transistors are important.

Failure Analysis and Bias State: Mobile electrons (or holes) that are accelerated across the saturated state depletion region cause impact ionization and photon emission that are readily seen in the drain region by photoemission microscopy [2]. Normal transistors are in the saturated state during most of their logic state transition (shown later) and emit photons. Transistors in the quiescent portion of the logic cycle are either in the nonsaturated-state or off-state so that no photons are emitted. However, bridge and open defects may hold one or more transistors in the saturated state where the presence of a light emitting transistor indicates proximity of a defect. The defect may be a bridge external to the transistor that sets up a saturated state bias states. Or IC defects may lie in the transistor itself. Examples include gate shorts, soft *pn* junctions, or an open circuit on one of the terminals.

Review: MOSFETs have three bias or operating states; off, saturated, and non-saturated (or ohmic). CMOS ICs use all three states. The off and non-saturated states exist when logic gates are in their quiescent period. These

states set the high and low logic voltage levels through the power rails. The saturated state dominates when the transistors are switching. This state has analog voltage gain >> 1, which is beneficial since it shortens the switching time. The gate and drain voltage difference determines if a transistor is in the saturated or nonsaturated state. *p*- and *n*-channel transistors have similar operating mechanisms except the *n*-channel uses electrons as carriers, and the *p*-channel uses holes.

3. MOSFET Terminal Characteristics

MOSFETs have four terminals: gate, drain, source, and substrate (or bulk) (Fig. 1). The terminal current and voltage relations are well defined. Although these equations are a bit clumsy, they must be learned (Table 1). Eq. (1) relates I_D to V_{GS} when the transistor is in the saturated state. μ is the carrier mobility, ε is the dielectric constant of silicon, Tox is gate oxide thickness, W is gate width, and L is gate (channel) length. Eq. (2) relates I_D to V_{GS} and V_{DS} for the nonsaturated state. Eq. (3) states that $I_D = 0$ if $V_{GS} < V_t$. Eqs. (1) to (3) were developed for "long channel" transistors that may be approximated for L ≥ 0.5 um. Models for short channel devices (L ≤ 0.5 um) are more complex to account for electric field and charge interactions at the smaller dimensions. However, we still use Eq. (1) and (2) in modern failure analysis since the parameters give physical intuition and good estimations without resort to computer calculations. *p*MOSFETs have similar equations, but the voltage and current polarities are negative. Short channel transistor models are described later.

We can combine Eq. (1) and (2) to model the terminal characteristics over a range of gate and drain voltages. Figure 3 shows a typical I_D versus V_{DS} family of curves measured on a 0.18 um *n*-channel transistor. Each curve merges half of a parabola (non-saturated region) with a flat portion (saturated region). Eq. (2) applies to the parabola and Eq. (1) to the flat line. A pinchoff point exists at the intersection of the two curves and exists for each saturated, flat line. We need the relation that defines the boundary of these two bias states. Without knowledge of the correct bias state, we cannot pick the proper model (Eq. 1-3).

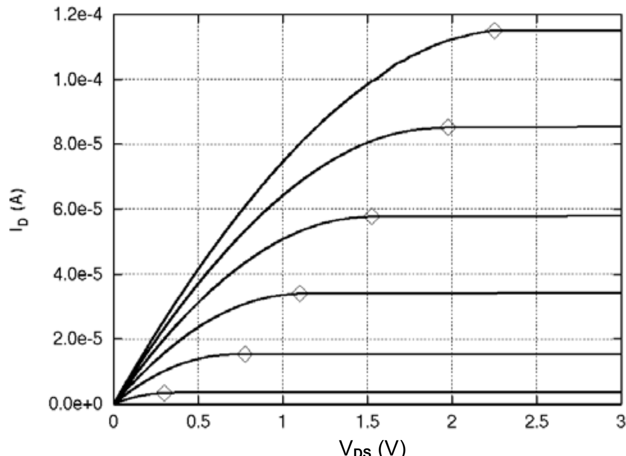


Fig. 3. I_D versus V_{DS} for a family of V_{GS} curves in an *n*MOSFET.

Eq. (2) is a parabola peaking at the boundary between the saturated and non-saturated regions. The left hand side of the parabola is the true operating region of a transistor in the non-saturation state while the right hand side is a non-defined state, and values here are ignored. The bias condition at the boundary is found by differentiating I_D with respect to V_{DS} in Eq. (2), setting the result to zero, and solving for V_{GS}. We then get Eq. (4a) that defines the voltage conditions that put the transistor at the boundary of the two states.

Table 1. *n*MOSFET Long Channel Bias State Models.

State	Eq.	
Saturated	$I_D = \frac{\mu \varepsilon}{2Tox} \frac{W}{L} (V_{GS} - V_t)^2$	(1)
Non-Saturated	$I_D = \frac{\mu \varepsilon}{2Tox} \frac{W}{L} [2(V_{GS} - V_t) V_{DS} - V_{DS}^2]$	(2)
Off-state	$I_D = 0$ for $V_{GS} < V_t$	(3)

$$V_{GS} = V_{DS} + V_t \quad (4a)$$

The most common use of Eq. (4a) is to define saturated and non-saturated states by

Saturated State: $V_{GS} < V_{DS} + V_t$ (4b)

Non-Saturated State: $V_{GS} > V_{DS} + V_t$ (4c)

Bias State Examples: Three transistors are shown in Fig. 4a-c with terminal voltages. An exercise is to use Eq. (4) and give the correct bias state.

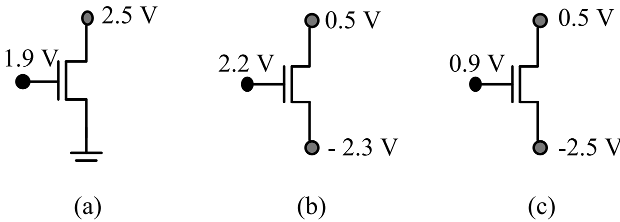


Fig. 4a-c. Transistor bias examples. ($V_t = 0.4$ V)

Fig. 4a is in the saturated state since $V_{GS} = 1.9 < 2.5 + 0.4$. Fig. 4b is in non-saturated state since $V_{GS} = (0 - (-2.5)) > (-1.0 - (-2.5)) + 0.4$. Fig. 4c lies on the boundary of the saturated and non-saturated states, therefore Eq. (1) or (2) may be used.

Analysis of a Complete Circuit

Transistors don't operate in isolation. They require a power supply (V_{DD}), ground (V_{SS}), an input signal, and a load. The load can be a resistor (Fig. 5) or another transistor. Eqs. (1-3) allow analysis of node voltages and the drain current (I_D).

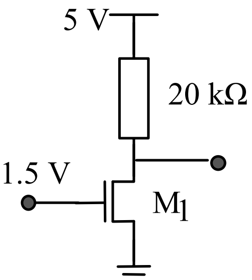


Fig. 5. *n*MOSFET 20 kΩ load circuit.

Example (Fig. 5): Calculate I_D, I_S, V_{DS}, and power dissipated by the circuit (P_{DD}) and transistor (P_T) if $V_t = 0.4$ V, $V_{IN} = 1.0$ V, and the conduction constant

$$K' = \frac{\mu \varepsilon}{2Tox} \frac{W}{L} = 60 \ \mu A/V^2$$

Solution: We don't know which bias state exists so we will try the saturated state Eq. (1) and see if the solution is consistent with Eq. 4b.

$$I_D = K'[V_{GS} - V_t]^2$$

$$= 60 \frac{\mu A}{V^2} [1 - 0.4]^2 = 21.6 \ \mu A$$

Then from Kirchhoff's Voltage Law,

$$V_{DD} = I_D R_D + V_{DS} \quad (5)$$

$$V_{DS} = 3 - (21.6 \ \mu A)(20 \ k\Omega) = 2.57 \ V$$

from Eq. (4b), $V_{GS} < V_{DS} + V_t$

$$1.0 < 2.57 + 0.4$$

The transistor is in saturation so the first guess was correct and $I_S = I_D = 21.6$ uA. The power to the circuit is $P_{DD} = V_{DD} \cdot I_{DD} = (3 \ V)(21.6 \ \mu A) = 64.8 \ \mu W$. The transistor power is $P_T = V_{DS} \cdot I_D = (2.57 \ V)(21.6 \ uA) = 55.5 \ \mu W$.

What if the transistor was in the non-saturated state? We repeat the procedure assuming that the transistor is in saturation. Set $V_{GS} = 2.5$ V and use

$$I_D = K'[V_{GS} - V_t]^2$$

$$= 60 \frac{\mu A}{V^2} [2.5 - 0.4]^2 = 264.6 \ \mu A$$

$$V_{DS} = 3 - (264.6 \ \mu A)(20 \ k\Omega) = -2.29 \ V$$

This is an error since $V_{GS} > V_{DS} + V_T$ is not consistent with our initial saturated state assumption. Also, a negative potential is not possible in this circuit. We then start over again using the non-saturated Eq. (2). Since Eq. (2) has two unknowns (I_D, V_{DS}), we need another equation. Kirchhoff's Voltage Law states that the sum of the voltage drops in a loop in zero. Therefore,

$$V_{DD} = I_D R_D + V_{DS}$$

and $I_D = [V_{DD} - V_{DS}]/R_D$ (6)

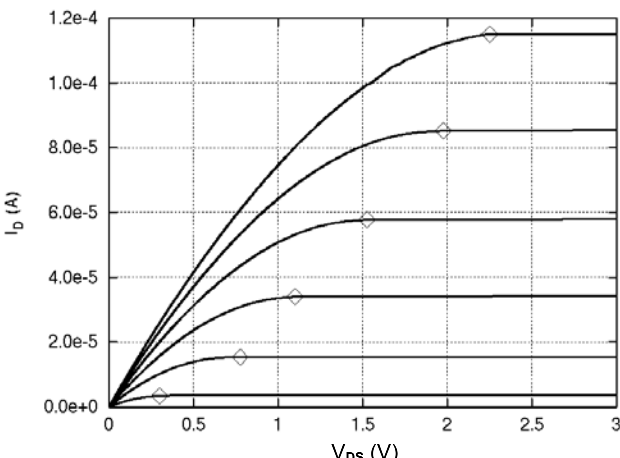
Combine Eq. (6) with Eq. (2) below

$$I_D = K'[2(V_{GS} - V_t) V_{DS} - V_{DS}^2] \quad (2)$$

and solve the quadratic equation for two values V_{DS} = 4.47 V, 0.559 V. The solution is V_{DS} = 0.559 V since 4.47 V violates our non-saturated state assumption, and it is also larger than V_{DD}. Eq. 4c is satisfied: 2.5 V > 0.559 V + 0.4 V confirming the non-saturated state assumption. I_D is calculated by either Eq. (6) or (2) above as I_D = 122.1 µA.

Short Channel Transistors

When transistor channel length shrunk below about 0.5 µm, and especially to 0.35 µm and smaller, then two phenomena appeared that changed MOSFET transistor properties. The first was that field strengths in the channel became large enough to cause charge carrier velocity saturation. The second was that the nearness of the transistor structures caused interactions leading to modulation of threshold voltage and effective channel length. The characteristics were altered as shown in Fig. 6 that compares long channel transistor I_D versus V_{DS} characteristics on the top with short channel transistor characteristics on the bottom. We observe that the velocity saturation effect converts the I_D relation to V_{GS} from a square law to a linear one. There are equally spaced current lines as V_{GS} increases in short channel transistors. In addition, the pinchoff point defining the boundary between the saturated and ohmic regions occurs at lower V_{DS} values on the short channel transistor than for the long channel transistor, and the slope of the saturated current lines is larger for the short channel transistor.



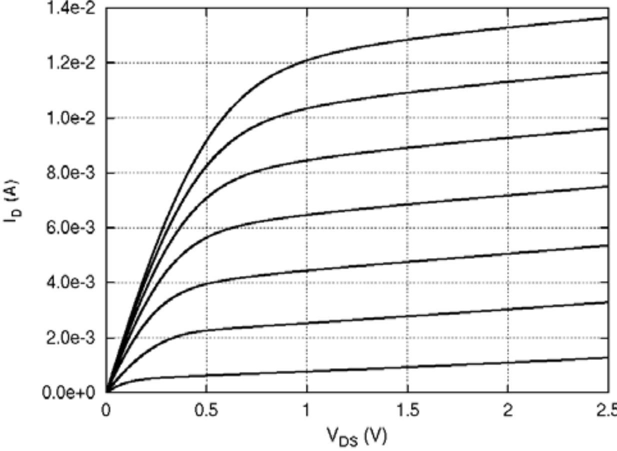


Fig. 6. Comparison of I_D versus V_{DS} for long channel transistor (top) with short channel transistor (bottom), from [1].

Naturally, the long channel device modeling equations (Eq. 1,2) change and become more complex as shown in Table 2. Eq. (7) and (8) describe the short channel transistor saturated and non-saturated drain current respectively. New parameters are included: E_{crit} is the electric field strength at which velocity saturation occurs, δ is a charge related constant, and θ_1 is a gate bias mobility constant.

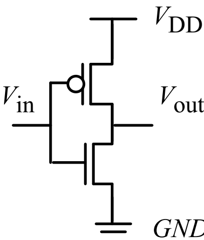
Table 2. MOSFET Short Channel Bias State Models

State	Eq.
$I_D = \mu_0 C_{ox} \frac{W}{L_{eff}} \frac{\left[(V_{GS}-V_t)V_{Dsat} - \frac{V_{Dsat}^2}{2}(1+\delta)\right]}{\left[1+\theta_1(V_{GS}-V_t)\right]\left[1+\frac{V_{Dsat}}{L_{eff}E_{crit}}\right]} \times \left[1+\lambda(V_{DS}-V_{Dsat})\right]$	(7)
Non-Saturated	
$I_D = \mu_0 C_{ox} \frac{W}{L_{eff}} \frac{\left[(V_{GS}-V_t)V_{DS} - \frac{V_{DS}^2}{2}(1+\delta)\right]}{\left[1+\theta_1(V_{GS}-V_t)\right]\left[1+\frac{V_{DS}}{L_{eff}E_{crit}}\right]}$	(8)
Off-state $I_D = 0$ for $V_{GS} < V_t$	(9)
Simplified Saturated	
$I_D \approx W C_{ox}(V_{GS}-V_t)\mu E_{crit}$	(10)

A simpler relation exists for the short channel saturated region. If we assume that the (L_{eff} E_{crit}) term is small, then we get Eq. (10). This simplification shows the linear relation between I_D to (V_{GS} - V_t), the independence of channel length, and the influence of channel electric field. This is a popular equation to model short channel transistors, but it is too simplistic to use in manual analysis. The accuracy is poor, and the pinchoff point defining the boundary between saturation and non-saturation doesn't match for both bias equations. Therefore, these equations define short channel models, but leave us without a back-of-the-envelope technique. This is a fact of life. Accuracy is extremely important in short channel transistors to model pico-second timing, but that process is opaque to the designer as detailed models and parameters are fed to the computer. These limitations led us to conclude that the long channel transistor relations (Eq. 1-2) are the only ones that allow reasonable, fast manual analysis. They teach us the intuition necessary to understand saturated and non-saturated states, the effect of loads, and insights needed to understand logic circuits, such as the simple inverter, NAND and NOR gates, and transmission gates.

4. CMOS Inverter

The smallest CMOS logic gate is the inverter having one *n*- and one *p*-channel transistor (Fig. 6a). The transfer curves in Fig. 6b illustrate the complementary action of the two transistors. When V_{IN} = 0 V, the *n*-channel is off and the *p*-channel with 1 V across the source to gate is driven hard into its non-saturated state. It is a low resistance switch so V_{OUT} = 1 V (logic high). As V_{IN} rises, the *n*-channel will turn on at V_{IN} = V_{tn} going from off to the saturated bias state. The gate voltage is above threshold, but much less than the drain voltage. The *p*-channel is in its non-saturated state and drain current is drawn through both transistors (dotted line). As V_{IN} rises, the drive voltage to the *p*MOSFET gate (V_{SG}) drops, and the *p*MOSFET current drive strength weakens. Near the midpoint, a maximum current exists that recedes as V_{IN} rises, and the *p*-channel drive strength diminishes. When (V_{DD} – V_{IN}) < -V_{tp}, the *p*-channel is off, I_{DD} and V_{OUT} are zero (logic low), and the *n*-MOSFET is driven hard into non-saturation. The point on the curve where V_{OUT} = V_{IN} is called the logic threshold voltage V_{TL}. The logic state is defined as changed when V_{IN} moves above or below this point. Figure 6b also shows a transfer curve at V_{DD} = 0.5 V. The slope in transition is steeper, and for this low V_{DD} value, the transient through current is gone.



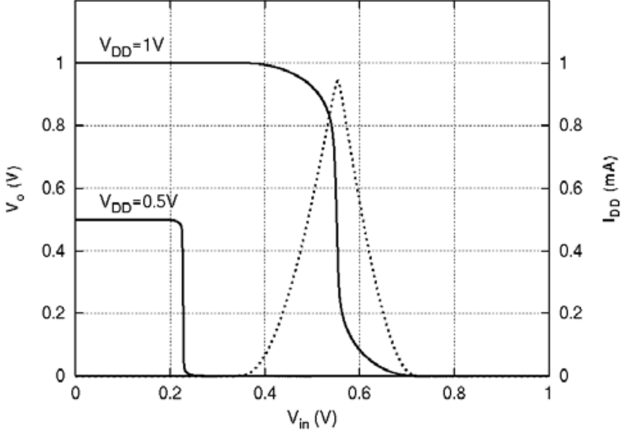


Fig. 6a. Inverter circuit.

Fig. 6b. Inverter voltage transfer curves at two V_{DD} values with voltage and current transfer curve for higher V_{DD} [1].

A detailed analysis of the transfer curve using Eq. 1,2 shows that both transistors are in saturation for about the middle 50% of the transition. Each transistor is in saturation for about 75% of the transition. You can verify this by plotting Eq. 4a overlaid on Fig. 6b for both *n*-channel (V_{tn} = 0.8 V) and *p*-channel (V_{tp} = -0.8 V) transistors. The analog voltage gain in the transition region can be estimated by measuring $\Delta V_{OUT} / \Delta V_{IN} \approx$ -10-15. Photon emission occurs during all of the transition region where one or both transistors emit photons during the transition.

The power supply current (I_{DD}) and the voltage (V_{OUT}) shapes should be committed to memory (Fig. 6b). Defects alter the shape of I_{DD} or V_{OUT} versus V_{IN}. The power supply current I_{DD} is near zero at the quiescent logic states because the drain current is junction leakage that is in picoamps (pA) for small, single transistors. Deep submicron transistors have additional leakage mechanisms that complicate test and failure analysis [3].

5. Test Methods

The electronic response of a defective IC implies knowledge of the test stimulus that led to a failure detection. There are several test methods [1].

1. Functional testing has at least three meanings: (a) stimulation of the IC in a way that replicates how the customer will use the part, (b) applying the logic truth table to the circuit, (c) using the primary inputs pins to an IC in contrast to using a scan chain test port.

   Functional testing cannot replicate customer use. For example, a 1K SRAM has 21024 possible states. If these patterns are run at 100 MHz then the test of full function will take about 10293 years. Complete functional testing for virtually any IC is numerically impossible.

2. Stuck-at-fault testing (SAF) uses a test abstraction that assumes all failures behave as one of the signal nodes is clamped to either the power or ground rail. SAF patterns apply voltages that represent Boolean logic patterns. A correct response for each pattern is stored in the tester and then compared with the result measured on the IC.

3. At-speed testing is the third major voltage-based test method. The F_{MAX} test applies many functional patterns to the IC and finds the maximum functional clock rate. Another test generates delay fault patterns that target either logic gates or signal paths, and measures the propagation delay.

4. I_{DDQ} testing measures the power supply current of a CMOS circuit during the quiescent period of the clock cycle. Most defects in CMOS ICs elevate I_{DDQ} above its normal value.

5. Other test types include I/O pin current and voltage levels, set up and hold time measurements, measurement of power supply current during the transient period of the clock, and testing at low power supply voltages.

Failure analysts work with ICs that fail one or more of the tests described above. These assorted tests fall into three categories: (1) voltage-based tests of logic function at

slow clock rate (scan stuck-at fault testing), (2) at-speed voltage-based testing, (3) parametric tests (I_{DDQ}, pin I/O, etc.). Modern ICs have parameter variations that can make test limit setting quite difficult. That issue will be addressed later, but for now we will assume that these tests have limits that are easily defined.

6. Electronics Properties of CMOS Bridge and Open Defects

CMOS IC defects have several types with different electronic patterns. Defects are classed as bridges, opens, or parametric delay types [4]. Their properties and response to voltage and current-based testing are described.

6.1 Bridging Defects

Bridging defects are unintentional shorts between interconnect or power lines. Bridges can be tiny connections (Fig. 7) or may cover several interconnect lines. They also occur within the structure of transistors, such as with gate oxide shorts or soft *pn* junction breakdowns. Bridges can also occur between metal layers stacked on top of each other (vertical shorts).

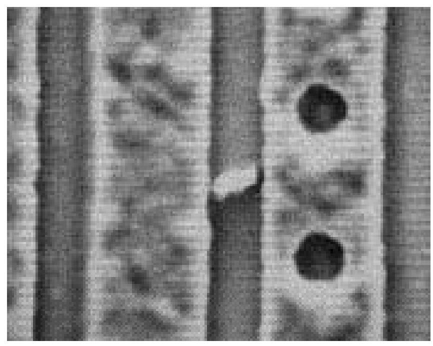


Fig. 7. SEM photo of bridge defect between two metal bus lines.

Figure 8 sketches a bridge defect between two logic gates. The effect of a bridge on functionality depends on bridge location and its resistance. If the resistance is large (≥ 5 kΩ), then little effect is seen on circuit functionality. A circuit will fail only when the resistance is sufficiently small, the nodes of the bridge are driven to opposite logic states, and the failing signal node can be propagated to an output pin. When failure occurs, one bridge node is correct and the other faulty.

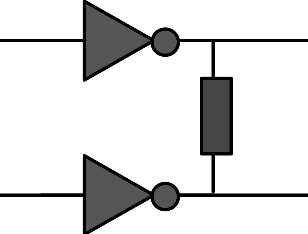


Fig. 8. Bridge defect across output signal nodes of two inverters.

The dominating property of bridge defects is called *critical resistance* [5]. Critical resistance (R_{crit}) is the minimum value above which the circuit functionally passes. Imagine a 1 GΩ bridge between the output of two logic gates (Fig. 8). The effect on the signal node voltage levels is insignificant and if this resistance was reduced to 1 MΩ or 10 kΩ, then the effect is still small. We typically do not see circuit failures until most bridges get below 2 kΩ. This is surprisingly low, but verified with simulations and failure analysis.

Figure 9 shows a buffered 2NAND gate in which variable gate to drain defect resistances (dotted line) were simulated for the *n*-channel transistor. Figure 10 shows the DC transfer curves for several gate-drain bridge resistances in an *n*-channel transistor showing that the circuit is functionally correct above about 1 kΩ (R_{crit} ≈ 1 kΩ). The exact value of R_{crit} depends on the relative current driving power of the pull-up and pull-down transistors that contend at each end of the bridge. A stronger current drive transistor dominates a weak one, and the bridge defect node tied to the weaker transistor will fail first. Failure is defined as the logic threshold voltage where the logic state changes. R_{crit} goes to zero when the pull-up and pull-down strengths are equal.

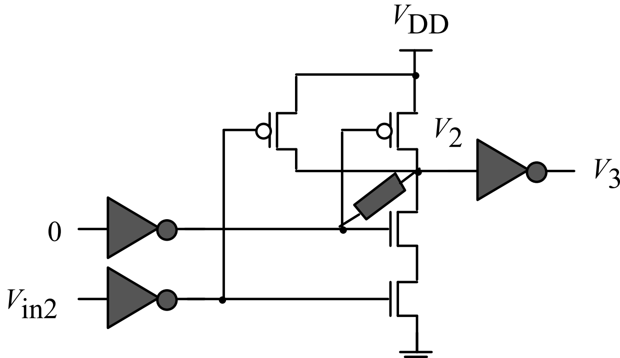


Fig. 9. Buffered 2NAND test circuit showing logic response to a gate-drain defect resistance (dotted line) [4].

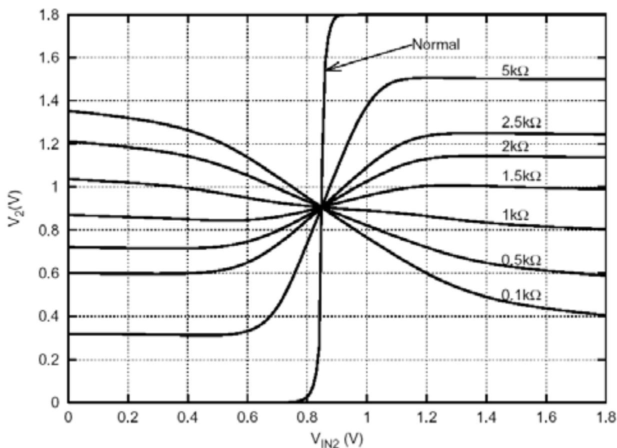


Fig. 10. Transfer curves of 2NAND when R_{def} across gate to drain terminal is varied (Fig. 9) [1].

Simulations of the effect of bridge resistance on timing showed an effect similar to the static transfer curves. The affect on timing was negligible until the bridge defect resistance became less than about 1 kΩ [1]. This R_{crit} effect explains why many quite dramatic bridge defects do not cause circuit failure. It is a reason why voltage-based tests such as, functional, stuck-at, or delay fault are weak in bridge defect detection. The I_{DDQ} test does however have sensitive detection capability for bridging defects. Figure 11 shows mA I_{DDQ} elevation for the circuit shown in Fig. 9.

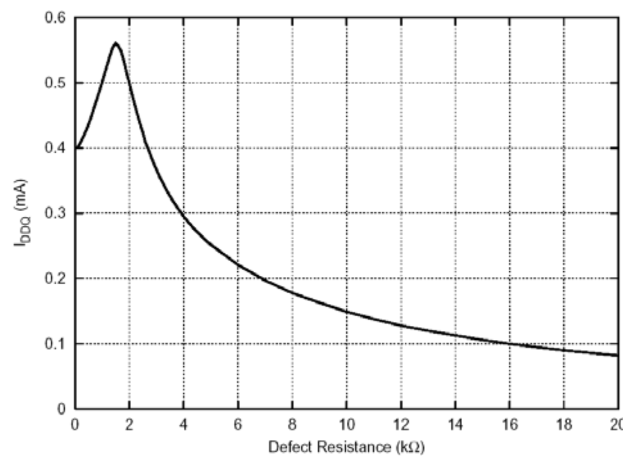


Fig. 11. I_{DDQ} response to gate-drain bridge defect (Fig. 9) [4].

Any bridge defect with at least one node tied to a signal line shows the critical resistance effect. Rail-to-rail bridges do not, but show a constant I_{DDQ} elevation for all test vectors. The critical resistance can be calculated for a given circuit using Eqs. (1-2), but space prevents it being done here [1].

Gate oxide shorts are defective electrical paths between the gate material and anything under the thin oxide [6,7]. A gate short is not a zero ohm structure. They can have linear or nonlinear I-V properties. Gate shorts can be severe hard breakdowns, or lightly stressed causing soft breakdowns. Older technologies had larger transistor gate capacitance (C_g) driven by larger 5 V power supplies. This led to large energy discharges (0.5 C_g V^2) such that the oxides ruptured under high stress causing what are called hard breakdowns. Recent technologies use V_{DD} on the order of 1 V, and the transistor gate areas are several times less than older generations. Therefore, gate capacitive energy discharges are smaller, and failure analysis labs don't see the catastrophic gate short as often.

The inverter has several forms of hard gate shorts. A connection in the *n*-channel transistor from an *n*-doped gate to an n^+drain or n^+source creates a parasitic linear resistor. An *n*-doped gate to *p*-well connection creates a diode that when powered up with positive logic on the gate forms a parasitic *n*MOSFET whose gate and drain are connected. A connection between the source or drain of the *p*-channel to their gate also creates a diode. The *p*-

doped gate to *n*-well connection creates a diode that forms a parasitic *p*MOSFET when power is applied. Test detection of hard gate shorts is similar to linear bridges. Voltage based tests are weak and I_{DDQ} is the only test that guarantees detection [7].

Ultrathin oxides in the 15 – 25 Å range show a soft breakdown in addition to the hard breakdowns of thicker transistor oxides. Soft breakdown in these ultrathin oxides is an irreversible damage to the oxide whose most significant effect is the increase in noise of the gate voltage. I_{DDQ} is not elevated for the soft gate-substrate ruptures of ultrathin oxides. The noise can show up to four orders of magnitude increase after soft breakdown, and this is the only certain evidence of the irreversible damage to ultrathin oxides. Hard oxide breakdowns appear in the advanced technology ultrathin transistor oxides if the gate is highly stressed, but the soft breakdowns are the object of reliability studies.

Recent studies have turned to the effect of ultrathin gate shorts on IC functionality. This issue is important for failure analysis. Degraeve, et al., found that uniformly stressed 2.4 nm gates oxides had a uniform area breakdown [8]. Breakdown over *n*MOSFET transistor channels had a high resistance from 10^5 – 10^9 Ω, while breakdowns over the drain and source had resistances of 10^2 – 10^4 Ω. Since the drain and source gate overlap area was much smaller than the area over the channel, most of the breakdowns occurred over the channel. By measuring breakdowns in 150, 180, and 200 nm transistors, they found that the percentage of gate to drain and source breakdowns increased as the transistors became smaller. The drain and source gate overlap area becomes a larger fraction of the transistor gate area. The gate to drain or source breakdowns were likely to cause hard failure, while the gate to channel breakdown showed no failure effects. A 41-stage ring oscillator with seven transistor gate ruptures continued to function with a 15% decrease in oscillator frequency. Part of the frequency reduction was due to hot carrier damage during a pre-stress. Significantly, the ring oscillator did not fail despite having several gate oxide breakdowns present.

Detection of softly ruptured ultrathin oxides does not appear possible at this time, nor is the reliability status clear. The normal functioning of the transistor with an ultrathin oxide is not as effected as were the killer ruptures of the thicker oxides. The recent ultrathin oxide experiments indicate test escapes are likely, but subsequent reliabilities may not be as risky as for breakdown in older technologies [8-12]. These different properties of the transistor oxide demand more studies at the circuit level to assess the implications of test escapes.

Summary of Bridge Defect Electronics: A bridge defect with impedance above critical does not cause functional failure. It weakens node voltages, elevates I_{DDQ} , worsens noise margins, and is often a reliability risk. The affect on propagation delay is negligible until the bridge resistance approaches the critical value. Bridge defects < R_{crit} will show at least one signal node in error.

6.2 Open Defects

Open defects have several forms and their electronic behavior is much more diverse than bridges. Defect location and physical dimensions are important open variables. Six open defect behavior classes are [1,4]

1. Transistor with missing gate contact. Fig. 12a shows a test structure and its transfer function (Fig. 12b). The circuit is functional, but has a weak high voltage and an even weaker low voltage. I_{DDQ} is elevated in one logic state. The capacitive coupling between drain-gate and gate-source forms a capacitive voltage divider. When the drain voltage is sufficiently high, then the divider allows the gate voltage to exceed threshold and the transistor conducts. When the input voltage is high, the *p*-channel shuts off. The *n*-channel transistor stays on draining charge from the load capacitance until V_G is no longer above threshold. Then the output node is a weak logic low level with a high impedance or floating state as shown by the horizontal lines in Fig. 12b.

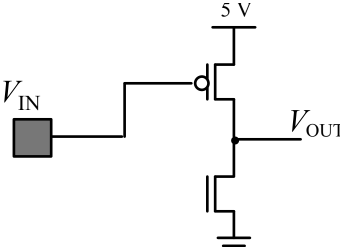


Fig. 12a. An open defect (missing contact) at *n*-MOSFET gate [4].

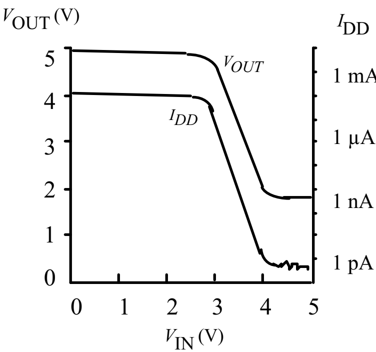


Fig. 12b. Transfer curves for open gate contact [4].

2. <u>An interconnect break affecting both complementary transistors</u> (Fig. 13). This open circuit creates a high impedance or floating node that acquires a steady state voltage. That voltage is dependent upon local topography, especially parasitic coupling capacitance. The isolated gate can float to either rail potential or any voltage in between allowing *two* possible behavior modes. If it floats to a rail, then that input node has a stuck-at behavior, and I_{DDQ} is not elevated. If the floating node acquires an intermediate voltage greater than the *p*- and *n*-channel transistor thresholds, then both transistors are permanently on, I_{DDQ} is elevated, and the output is a weak stuck-at voltage. That form of open is detectable with a stuck-at fault test or an I_{DDQ} test.

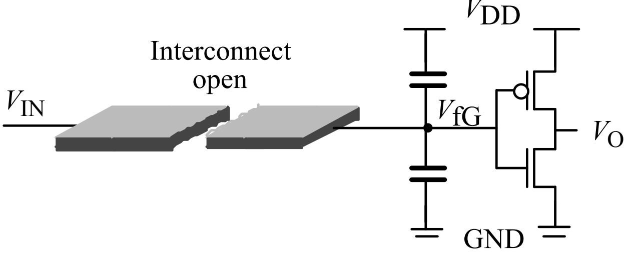


Fig. 13. An open defect to the logic gate input.

3. <u>Open defect behavior in sequential circuits</u>. Fig. 14 shows a master-slave flip-flop with several possible open defect sites. Analysis shows that any of these open defects will generally cause behavior consistent with the three classes described above. An open in one of the CMOS transmission gate signal lines still allows functionality as if the gate were a single pass transistor, but signal strength is degraded. The responses of these sequential open defects are either I_{DDQ} elevation only, functional fail only, or both.

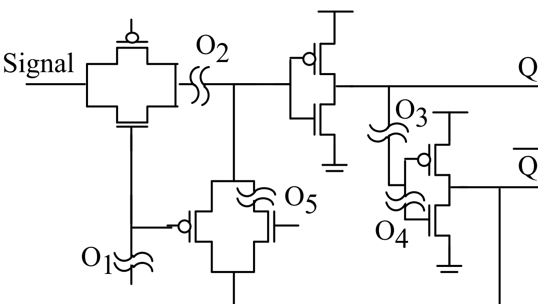
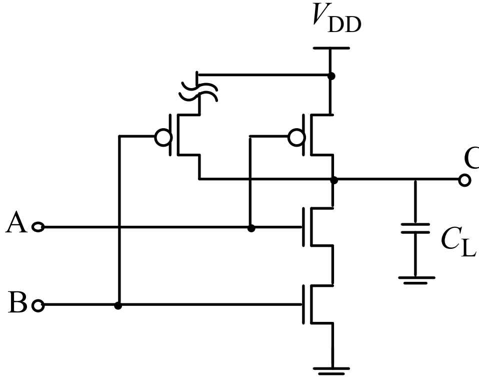


Fig. 14. Open defects in a sequential circuit.

4. <u>CMOS Stuck-Open (Memory) Defect.</u> The fifth open defect class is peculiar to CMOS and occurs when an open circuit happens in the drain or source of a transistor [13]. This also called the CMOS memory defect. Fig. 15 shows such a defect and its truth table response for a 2NAND gate. The first vector (AB = 00) sets the output to a logic high by turning on one *p*-channel and turning off both *n*-channel transistors. The second vector (AB = 01) drives the output high through the good *p*-channel pull-up. The third test vector (AB = 10) attempts to turn on the defective *p*-channel pull-up, but no charge can pass since the drain has an open circuit defect. The output node

goes into a high impedance state and might be indeterminate except that the previous logic state was a high. The load capacitance holds this high value, and the third test vector is read correctly. The fourth vector pulls the output node low and the result is dramatic. No error in the truth table occurred for such a flagrant defect! What then is the functionality problem?



A	B	C
0	0	1
0	1	1
1	0	1
1	1	0
⋮	⋮	⋮
1	1	0
1	0	0

Fig. 15. Open memory defect in a 2NAND gate and its truth table.

The defect is detected when the sequence of vectors is changed. If AB = 11 is followed by AB = 10, then the AB = 10 vector is read at the output as a fail state zero (the previous logic state). Stuck-open defects occur and are a difficult failure analysis challenge. The main reason is the conflicting information that the IC sends due to the memory response of that defective node. I_{DDQ} is often elevated due to design contentions caused by the unintended logic state, or the node drifts with a 1-3 s time constant at room temperature and can turn on pairs of load transistors [8].

5. Crack in an interconnect line (Fig. 16). This defect supports circuit functionality, especially for very narrow cracks [14]. A narrow crack allows electron tunneling across the barrier, and ICs with this defect type can operate in the hundreds of MHz. They are called tunneling opens. These defects occur more often in vias and contacts. Cold temperature lowers F_{MAX} since metal contraction widens the crack and reduces the tunneling. Conversely, circuits with these defects run faster at hot temperatures. These unusual frequency-temperature properties are symptomatic of an interconnect crack.

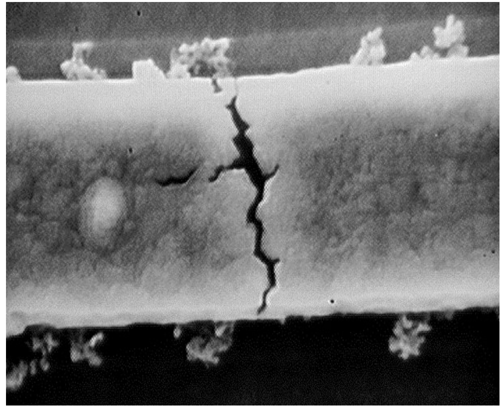


Fig. 16. Metal crack [14].

Summary of Open Defect Electronics: Six different behavior patterns were described that are dependent on the open defect location and size. A comprehensive test for open defects includes both voltage and current-based tests. Inattention to these details can frustrate failure analysis.

7. Parametric Failures

Parametric failures are the third and most difficult class of defects to detect [1]. Parametric failures have been with us since the beginning of CMOS technology, but their significance is now more serious and growing. Parametric failures have unusual properties that form broad behavioral patterns. Parametric timing failures fall into two classes: (1) intrinsic ICs (free of defects), and (2) extrinsic ICs (presence of defects). Intrinsic ICs can fail due to an unfortunate distribution of circuit statistics

and/or environmental conditions. An extrinsic IC may have subtle imperfections in vias or contacts that cause its speed characteristics to worsen at lower temperatures. We will discuss each class with supporting behavioral data, and then discuss the challenge of detection.

7.1 Intrinsic Parametric Failures:

Two factors cause intrinsic parameter variation: environmental and physical. Environmental factors include variation of the power supply levels within the die, on the board, or during switching activity and also from temperature variation across the circuit. Physical variation comes from the inherent weaknesses in IC manufacturing control that allows transistor and interconnect variations. These deviations from targeted values are limitations or imperfections in process and mask steps. The random and uncontrollable nature of parameter variations may cause these failures to be non-systematic from die-to-die, wafer-to-wafer, and even transistor-to-transistor property variation within a die (intra-die variation) [15-18]. For example, a drive transistor with a current strength slightly above nominal may compensate a slightly high interconnect resistance in a signal path. This same high resistance may lead to an unacceptable signal delay if the transistor has a driving strength below or at the nominal conductance value. This leads to inaccuracies that impact circuit quality, and can provoke erroneous behaviors that occur at very specific circuit states or environmental conditions. Recent technologies show die-to-die reordering of critical paths [16].

These types of failures can be difficult to detect and locate. Several failure analysis experiences with parametric delay defects showed that one to three months of effort may be necessary to locate one defect on an IC. This is intolerable, and research advances with scanning thermal lasers have reduced this fault location times to well under one hour. [19-21].

Present design technology is unable to characterize the whole complexity of parameter combinations, so present strategies often check only the corner parameters. Therefore, unfortunate parameter combinations can be very difficult to detect and screen. For example, an IC that is normally on the fast edge of the distribution could have a delay defect that puts it now at the slow, but acceptable range and the part passes. Defective parts that pass function are an increased reliability risk [22]. Test limits can no longer rely strictly on single limit approaches, but rather statistical techniques that improve the test signal-to-noise ratio.

Variations in IC Critical Parameters

Statistical variations in primary IC parameters is a concern not only for functionality, but also for setting test limits. It is a primary concern in advanced technology ICs. A bad die can be within the statistical spread of the normal population. Figure 17 shows V_t variation from 8% at the 180 nm die to more than 12% for the 130 nm die [23]. Keshavarzi, et. al., also presented L_{eff} variation data and its impact on I_{Dsat} which is the transistor speed parameter [3]. Transistor L_{eff} intrinsic variation is the primary variable that influences F_{max} [24].

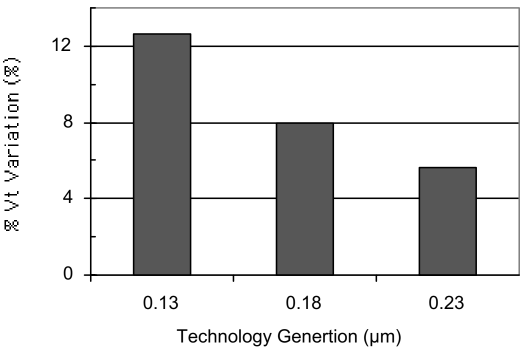


Fig. 17. Threshold voltage variation within the die [23].

Defect-free ICs also show increased random failures due to interconnect properties [24]. These properties include crosstalk, ground bounce, power line IR droop influence on timing or even Boolean failure [25]. Crosstalk errors arise from poor design rule implementation (a difficult problem) or statistical fluctuations in metal line spacing (and metal width) [26-27]. These properties are aggravated by the present geometric structure of metal cross-sections in which their height to width ratios (aspect

ratio) are now >2. Metal interconnect aspect ratios of 2-3 make the dominant metal capacitance from sidewall to sidewall instead of to the power rails as in older technologies.

Although crosstalk and switching noise are traditionally categorized as design–related problems, technology-scaling complexity has diffused the boundaries between design and functionality. Designers must take statistical variation into account in the designs, accepting that some circuits from the same fabrication lot fail, while others do not. An Intel speaker at the 2003 Design Automation and Test in Europe Conference cited that 67% of F_{MAX} field returns on a product were due to cross-talk noise and 20% were due to IR droop on the power lines. These failures occur in defect-free parts, but they are just a subset of a broader class of parametric failures.

Another intrinsic concern is the rate of change in power supply currents that is now on the order of tens of Amps per ns. Low V_{DD} with higher di/dt increases the statistical risk of parametric failures due to inductive power rail bounce and power supply IR drop, combined with lower signal margin.

Each technology node has a minimum metal width of about that technology node number. A 100 nm node has a minimum metal width of about 100 nm [28]. The minimum dielectric spacing between minimum metal geometry is also about 100 nm. The vias and contacts may have diameters on the order of the metal widths, and since the intermetal dielectric vertical spacing scales more slowly, the via and contact aspect ratios tend to get larger. ICs have hundreds of millions and billions of these structures and defect-free vias cannot be guaranteed. Vias and contacts are a difficult challenge for fabrication, test engineering, and failure analysis.

Figure 18 shows the statistical timing variation on a signal path measured on twenty-five 0.25 μm technology wafers with cumulative propagation delays measured on 910 die in each wafer [29]. The initial rise in the plots shows a near straight line relation to the normal distribution. The difference in propagation delay between fastest and slowest die at V_{DD} = 2.5 V is about 1.7 ns representing a difference of about 24% with respect to the faster IC. When the power supply was reduced to V_{DD} = 1.2 V (Fig. 18b), the propagation delays increased almost three times. The break in the curve at the 95% cumulative point is more distinct. These inherent statistical deviations can lead to timing failures, especially when deviations exist within the die. A slow data path signal combined with a fast clock path to a flip-flop can violate setup times and cause functional failure.

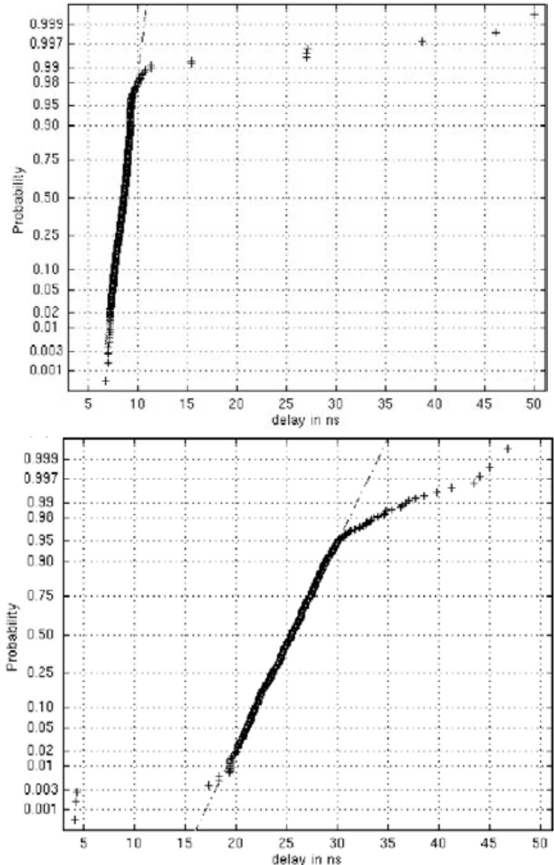


Fig. 18. Cumulative distributions for propagation delay from 25 wafers with data including 910 die per wafer. (a) V_{DD} = 2.5 V, (b) V_{DD} = 1.2 V [29].

There is a physical difference between parts in the normal and outlier distributions. The outliers have defects causing delay in addition to the part's otherwise normal variation. An important point is that real tests that target delay faults do not have the capability and resolution to measure to this fine degree for each delay path. Each 2-vector timing test pattern does not have an individual test limit, so that delay fault testing typically uses the period of the system

clock, and then adds an amount to account for tester noise and parameter variance. This can lead to test escapes and reliability returns from the field. Delay fault testing is a gross test for timing errors. A test limit must take into account the slowest parameter measurement. Slower outliers from the fastest die overlap normal data in the slower die. This is part of the challenge in detecting these statistical variance failures.

Figure 19 shows the sensitivity of F_{MAX} to V_{DD} of an Intel 1 GHz, 6.6 million transistors, router chip built in 150 nm technology. The chip has an approximate 1.8 MHz change in F_{MAX} per mV of power supply voltage at V_{DD} = 1.3 V. Figure 19 shows about a 14% increase in F_{MAX} when V_{DD} increases 10% from V_{DD} = 1.3 V. The sensitivity is higher at the low end V_{DD} and saturates above V_{DD} = 1.6 V. Bernstein, et al., reported a change in F_{MAX} with V_{DD} of 200 kHz/mV for a 180 nm microprocessor, and gave a performance rule of thumb that chips vary by 7-9% when the power supply varies by about 10% [24]. Vangal, et al., showed F_{MAX} versus V_{DD} plots for a 130 nm dual-V_t 5 GHz technology with a sensitivity of 11.3 MHz/mV at V_{DD} = 1.0 V [30].

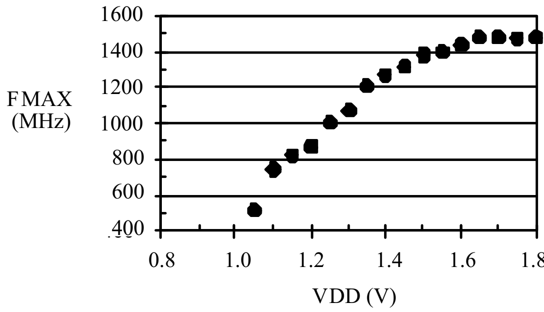


Fig. 19. F_{MAX} versus V_{DD} for a 150 nm IC [15].

Figure 20 emphasizes that mV noise changes in V_{DD} induced on a critical signal node in normal operation can measurably affect IC speed. A related system problem occurs when board power supplies have a ± 5% accuracy. This could move a test verified 1 GHz IC performance in Fig. 20 to an operational range from about 883 MHz to 1.18 GHz on the board. One protection used by manufacturers is to guardband the shipped product anticipating power supply influences, but this is wasteful of IC capability and also a yield loss. Again, these are statistical variations that were not so important in older technologies, but they are now. An IC apparently failing may be due to environmental influences beyond its control.

Temperature also affects the die speed performance. This is especially so for thermal gradients on the die. Simulations and measurements on high performance chips show gradients of 130°C to 50°C on the same die. Temperature data taken from a small, 20,000 transistors test circuit showed temperature-F_{MAX} slopes of –13.1 kHz/°C and –12.3 kHz/°C for the fast and slow ICs [15]. The major effect that slows an IC with temperature rise is the decrease in carrier mobility. A compensating speed effect is that the absolute values of the *n*- and *p*-channel transistor threshold voltages decrease as temperature rises.

7.2 Extrinsic Parametric Failures:

We will describe four extrinsic IC mechanisms associated with parametric failures: (1) resistive vias and contacts, (2) metal mousebites, (3) metal slivers, and (4) gate oxide shorts in ultrathin technologies. Typically the most common extrinsic parametric failures in recent technologies are the metal vias and contacts, and the presence of metal slivers aggravated by Chemical Mechanical Polishing (CMP). Slivers are a common defect in which a metal particle lies between two metal conductors and barely contacts the signal lines. Mousebites occur when sections of metal are missing from an interconnect line. Gate shorts shows a variety of responses. Some gate shorts show timing and power supply dependent failures, and the recent ultrathin oxides have a unique failure mode questioning the implied reliability risk for some gate oxide shorts.

Resistive Vias

Modern ICs can have billions of transistors and perhaps ten times that number of vias. Modern via aspect ratios are now > 5, so that defective vias with elevated resistance are not surprising. Vias and contacts have different sizes depending upon the metal level. The vias and contacts at the lowest metal level are the smallest usually close to

minimum feature size. Cracks in flat metal lines also show properties similar to resistive vias, but are a less common failure mechanism, particularly with the shunting barrier metals deposited around the Al or Cu metal interconnect.

Resistive vias have a unique, but expensive signature. A speed parameter (F_{MAX} or prop delay) will show degradation as the temperature is reduced. The problem was severe enough even at 0.25 um technologies to require two temperature testing [31].

Metal Mousebites

Missing regions of interconnect metal are called mousebites [1]. They can be due to particle defects, electromigration, or stress voids. Mousebites have a minor electrical effect, but are a major reliability risk [15]. Figure 20 (top) sketches a defect-free and a defective (mousebite) section of interconnect. If sheet resistance is $R_{\square} \approx 70$ mΩ/□ and a 90% bite is taken out of the middle section, then that line resistance changes from 210 mΩ to 840 mΩ. This major defect in the line would not elevate resistance sufficiently to be detected by a speed test.

Figure 20(bottom) plots the increase in resistance of the stripe versus fraction of mousebite. The resistance increases dramatically beyond about 95% voiding. This observation is useful when visually looking at mousebites and predicting the impact on failure. These small resistance increases due to mousebites have a negligible affect on RC time constant of that line. Mousebites are difficult to detect, but pose an electromigration failure risk due to the increased current density in the stripe.

These conclusions extend to voided vias and contacts that also show this resistance dependency on volume voiding. The via or contact must be well voided to cause an RC delay failure sensitive to temperature.

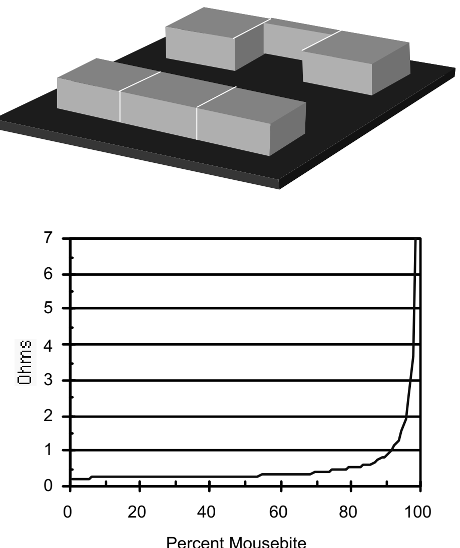


Fig. 20. (Top) Control metal line and normal metal line and one with mousebite, (bottom) Voided metal resistance (mΩ) versus percent metal voiding using R = 70 mΩ/□ [15].

Metal Slivers

Metal slivers increased their presence with Chemical Mechanical Polishing (CMP). A small metal sliver lies between two interconnect lines barely or not even touching them. These slivers can be from any of the metals used in IC fab, such Al, Cu, W, or stainless steel. When temperature rises, metals expand, and the sliver now touches the signal lines. Higher voltages at burn-in can promote the rupture of the high resistance oxide surface of the metals, bonding the three metal elements [22]. The bridge resistance is now permanent and low enough to reduce noise margins or even cause functional failure.

Metal slivers resistance can appear low but deceiving. Figure 21 sketches two bus lines with a small bridge sliver connecting them. The size and connectivity of the defect will affect the critical resistance.

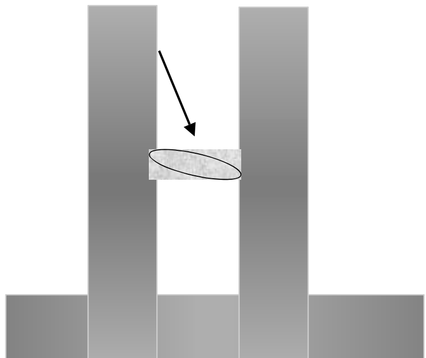


Fig. 21. Resistance of metal sliver.

The resistance of the sliver can be estimated. If we assume an aluminum sliver then the resistivity $\rho = 3.0\ u\Omega$ cm. If the dimensions of the sliver are: W = 0.2 um, L = 0.4 um, and thickness t = 0.4 um, then the resistance is

$$R = \frac{\rho L}{W t} = 150\ m\Omega \qquad (11)$$

This is a small value well below all critical resistances. However, failure analysis shows that these types of defects have from tens to hundreds of ohms [22]. Possible reasons are that the bonding of the sliver to the bus lines may not be strong or the material may not be Al, but chrome or other foreign particulate having higher resistivity. Also, all metals form an oxide at their surface that drives up the resistance when two metals touch.

Gate Oxide Breakdown

Hard and soft gate oxide breakdown properties were described earlier. Detection of hard breakdowns requires a current-based test, while soft breakdowns are not detectable in an IC. The soft breakdowns in the advanced technology ultrathin oxides may be a lessened reliability risk.

How to Detect Parametric Failure

The source of the parametric failure can be an actual defect, or it may be due to a particular transistor or interconnect line. The failure is probably sensitive to temperature, V_{DD}, or clock frequency. Traditional approaches would examine test data and simulation results. This approach is inefficient, and failure analysis with these techniques have been known to take months or never.

PICA and laser voltage probing (LVP) achieved some success over the older approaches [34,35]. PICA can ferret out subtle picosecond timing delays that cause functional failures while LVP can measure picosecond waveforms when it knows the particular signal nodes affecting the failure. Recently PICA sensitivity has decreased with lower power supply values for nanotechnology ICs.

Another approach has reported fail site location of parametric failures using scanning thermal and electron-hole-pair laser beams [19-21]. This technique powers the chip, and then drives it with a digital tester that repetitively cycles a test pattern. The output signal of the pin that shows failure is monitored and adjusted to a pass-fail margin by adjusting V_{DD}, clock frequency, or temperature. A laser then scans the die with either an 1140 nm or 1064 nm wavelength while the die image is put as background on the monitor. When the laser energy strikes the fail site, then the signal output pin will show a change of logic state from the margin. A latch driven output of the tester pin modulates the e-beam of the monitor indicating precise location of the fail site. This technique was successfully reported for resistive via defects [20] and for defect-free timing failures [19,21]. When the equipment is set up and the part is mounted in the fixture, then diagnosis of fail site can be done in minutes.

Signature Analysis with I_{DDQ} versus V_{DD}

CMOS IC designs with their low quiescent currents present a unique opportunity for failure analysis. Sandia National Labs has used I_{DDQ} versus V_{DD} signature analysis curves since the late 1970's to perform a rapid assessment of failure modes. Figure 22 shows the technique for sweeping the power supply pin voltage, V_{DD}, and monitoring I_{DDQ}. The IC is first put in a logic state that draws abnormal I_{DDQ}. The input pins at logic 0 are tied to ground and the logic high pins are tied to V_{DD}. It is important that input high pins are at the same voltage as V_{DD} or there is a risk that input protection circuits can be

damaged. The measurement setup is simple and conducted essentially under DC conditions.

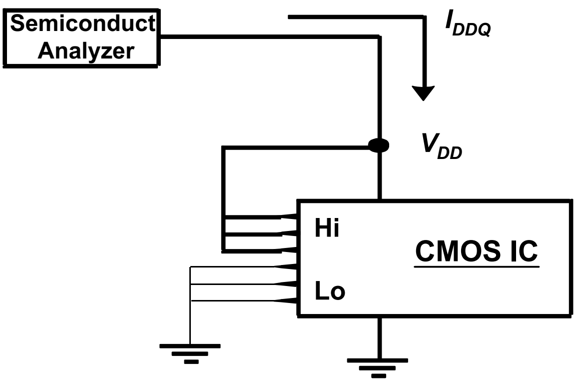


Fig. 22. Measurement setup for I_{DDQ} versus V_{DD} defect signatures.

I_{DDQ} versus V_{DD} signatures can rapidly indicate clues to the nature of the defect. The technique doesn't pinpoint defect location, but can often distinguish bridges from opens and power rail shorts from signal node shorts. A bridge defect with at least one end tied to a signal node will not draw current until V_{DD} rises above transistor threshold, V_T. In contrast, rail to rail bridge defects show conduction from the initiation of power. Figure 23 and 24 illustrates these two types of bridging defect signatures found in a microprocessor [32]. Figure 23 shows a 1.9 MΩ rail to rail short whose I-V curve reflects conduction of the defect without contributions from driving transistors. This defect will also show a line with same slope if V_{DD} is driven slightly negative.

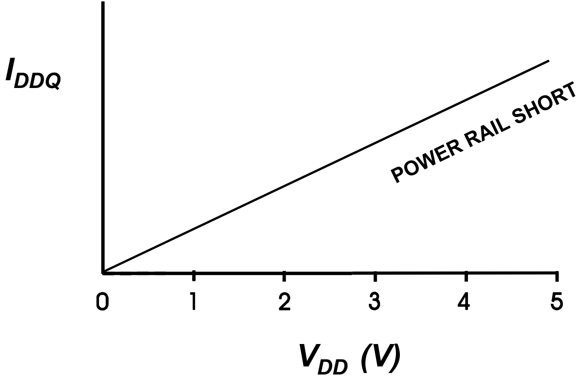


Fig. 23. A bridge defect between V_{DD} and V_{SS}.

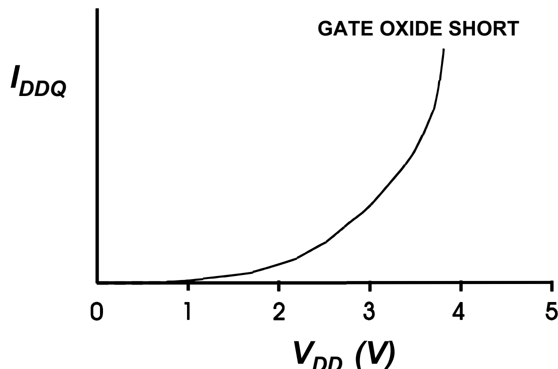


Fig. 24. Signature of *n*-channel gate oxide short.

Figure 24 shows an *n*-channel transistor gate oxide short I-V signature [11]. Conduction doesn't begin until V_{DD} exceeds V_t. The signature now includes contributions from the defect, the defective transistor, and the driving transistor. Gate to substrate shorts in *n*-channel transistors typically show a pronounced parabolic curve [6]. The gate short forms a parasitic MOSFET transistor whose gate and drain are connected. This connection places the device in its saturated state whose I_D and V_{GS} response is defined by a square law (Eq. (1)).

Several classes of open defects were described, and their signatures can also take different forms. Often, an open defect signature shows a time instability when the defective IC is powered. Figure 25 shows this type of behavior. Capacitive dividers may delay turn-on of the transistor as shown at ≈ 1.5 V. There is need for more research analyzing open defect signatures.

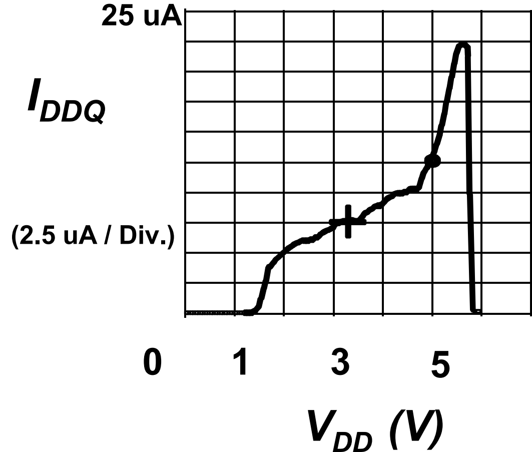


Fig. 25. An open defect response.

Soft breakdown of *pn* junctions also has a distinct signature (Fig. 26). This defect and most gate oxide shorts are strongly photon emitting.

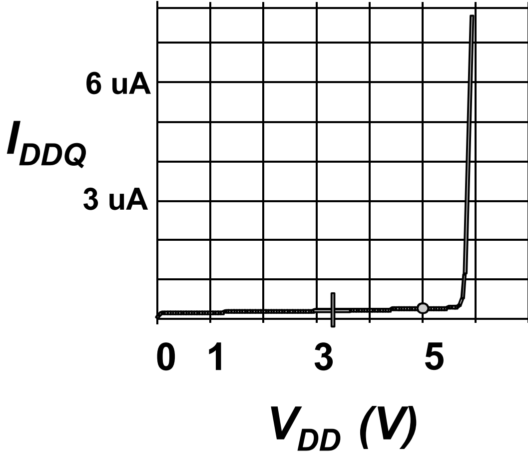


Fig. 26. Signature of *pn* junction early breakdown.

The curvature of the signature can direct a course of action. Curved signatures usually have light emission occurring either at the defect or in the transistor driving the defect. Straight-line responses have not been found to emit light [2]. A transistor in its linear (non-saturated) state will not emit photons. MOSFETs in their saturated state emit considerable light in the drain depletion region as do parasitic bipolar transistors in linear bias, and certain defective structures in the protection circuits. Wilson describes in his text a similar signature technique using curve tracer responses on I/O pins as opposed to the V_{DD} pin applications described here [33].

Summary

Failure analysts need skills in relating currents and node voltages to the three bias states of the transistor: off, saturated, and non-saturated. The CMOS inverter is the most fundamental logic gate, and its current and voltage transfer curve properties must be understood in detail. The electronic behavior of opens, shorts, and parametric delay defects was described, and it is essential for understanding the symptoms of a failing IC. A final application was given using I_{DD} versus V_{DD} signatures to identify types of defects.

Acknowledgements:

1. Sandia National Labs is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract number DE-AC04-94AL85000.
2. Jaume Segura acknowledges partial support from the Spanish Ministry of Science and Technology and the Regional European Development Funds (FEDER) from the European union (CICYT- TIC02-01238).

References

1. J. Segura and C. hawkins, *CMOS Electronics; How it Works, How it Fails,* IEEE Press & John Wiley & Sons, Inc, 2004.
2. C. Hawkins, J. Soden, E. Cole Jr., and E. Snyder, "The use of light emission in failure analysis of CMOS ICs," *Int. Symp. for Testing and Failure Analysis (ISTFA)*, Oct. 1990.
3. A. Keshavarzi, K. Roy, and C. Hawkins, “Intrinsic leakage in low power deep submicron CMOS ICs,” *International Test Conference (ITC),* pp. 146-155, Nov. 1997.
4. C. Hawkins, J. Soden, A. Righter, and J. Ferguson, “Defect Classes – An overdue paradigm for CMOS IC testing,” *Int. Test Conf.,* pp. 413-424, Oct. 1994.
5. R. Rodriguez-Montanes, J. Segura, V. Champac, J. Figueras, and A. Rubio, “Current vs. logic testing of gate oxide short, floating gate, and bridging failures in CMOS,” *International Test Conference (ITC),* pp. 510-519, Oct. 1991.
6. C. Hawkins and J. Soden, "Electrical characteristics and testing considerations for gate oxide shorts in CMOS ICs," *International Test Conference (ITC).*, pp. 544-555, Nov. 1985.
7. J. Segura, C. DeBenito, A. Rubio, and C.F. Hawkins, A detailed analysis and electrical modeling of gate oxide shorts in MOS transistors,“ *J. of Electronic Testing: Theory and Applications (JETTA),* pp. 229-239, (1996).

8. R. Degraeve, B. Kaczer, A. De Keersgieter, and G. Groeseneken, “Relation between breakdown mode and breakdown location in short channel NMOSFETs,” *International Reliability Physics Symposium (IRPS),* pp. 360-366, May 2001.

9. M. Alam and K. Smith, “A phenomenological theory of correlated multiple soft-breakdown events in ultra-thin gate dielectrics,” *International Reliability Physics Symposium (IRPS),* pp. 406-411, April 2003.

10. R. Rodriguez, J. Stathis, and B. Linder, “Modeling and experimental verification of the effect of gate oxide breakdown on CMOS inverters,” *International Reliability Physics Symposium (IRPS),* pp. 11-16, April 2003.

11. J. Suehle, “Utrathin gate oxide breakdown: A failure that we an live with?,” ASM Electron Device Failure Analysis Magazine, pp. 6-11, February 2004.

12. B. Weir, et al., “Ultra-thin gate dielectrics: they break down, but do they fail?,” *International Electron Device Meeting (IEDM),* pp*. 73-76,* 1997.

13. J. Soden, R. Treece, M. Taylor, and C. Hawkins, "CMOS IC stuck-open fault electrical effects and design considerations," *International Test Conference (ITC).*, pp. 423-430, August, 1989.

14. C. Henderson, J. Soden, and C. Hawkins, "The behavior and testing implications of CMOS IC open circuits *International Test Conference (ITC)*., pp. 302-310, Nashville, TN, Oct. 1991.

15. J. Segura, A. Keshavarzi, J. Soden, and C. Hawkins, “Parametric failures in CMOS ICs – A defect-based analysis,” *International Test Conference (ITC),* October 2002.

16. M. Orshansky, L. Millor, P. Chen, K. Keutzer, and C. Hu, “Impact of spatial intrachip gate length variability on the performance of hi-speed digital circuits,” *IEEE Transactions on Computer-Aided Design,* Vol. 21, No. 5, May 2002.

17. C. Visweswariah, “Death, taxes, and failing chips,” *Design Automation Conference*., pp. 343-347, June 2003.

18. M. Weber, “My head hurts, my timing stinks, and I don’t love on-chip variation,” *Synopsys User Group Meeting, SNUGBoston02*, 2002.

19. M. Bruce, V. Bruce, D. Epps, J. Wilcox, E. Cole, P. Tangyungyong, and C. Hawkins, “Soft defect localization (SDL) on ICs,” *International Symposium on Test and Failure Analysis (ISTFA),* November 2002.

20. E. Cole, P. Tangyungyong, C. Hawkins, M. Bruce, V. Bruce, R. Ring, and W-L. Chong, “Resistive interconnection localization,” *International Symposium on Test and Failure Analysis (ISTFA)*, pp 43-51, November 2001.

21. J. Rowlette and T. Eiles, “critical timing analysis in microprocessors using near-IR laser assisted device alteration (LADA),” *International Test Conference (ITC)*, pp. 264-273, Charlotte, NC, october 2003

22. W. Righter, C. Hawkins, J. Soden, and P. Maxwell, “CMOS IC reliability indicators and burn-in economics,” *International Test Conference (ITC)*, Washington D.C., November 1998.

23. S. Narendra, D. Antoniadis, and V. De, “Impact of using adaptive body bias to compensate die-to-die Vt variation on within-die Vt variation,” *International Symposium on Low Power Electronics and Design*, pp. 229-232, 1999.

24. K. Bernstein, K. Carrig, C. Durham, P. Hansen, D. Hogenmiller, E. Nowak, and N. Rohrer, *High Speed CMOS Design Styles*, Kluwer Academic Publishers, 1998.

25. M. Breuer and S. Gupta, “Process aggravated noise (PAN): New validation and test problems,” *International Test Conference (ITC),* pp. 914-923, October 1996.

26. A. Deutsch, et al., “On-chip wiring design challenges for gigahertz operation,” *Proceeding of the IEEE*, Vol. 89, No. 4, April 2001.

27. D. Sylvester and K. Keutzer, “Impact of small process geometries on microarchitectures in systems on a chip,” *Proceedings of the IEEE,* Vol. 89, No 4., pp. 467-489, April 2001.

28. International Technology Roadmap for Semiconductors 2001 (http://public.itrs.net/).

29. K. Baker, G. Grounthoud, M. Lousberg, I. Schanstra, and C. Hawkins, “Defect-based delay testing of resistive vias-contacts – a critical evaluation,” *International Test Conference (ITC),* pp. 467-476, October 1999.

30. S. Vangal, et al., "5-GHz 32-bit integer execution core in 130-nm dual-V_T CMOS," *IEEE Journal of Solid-State* Circuits, vol.37, no.11, p.1421-32, Nov. 2002.

31. W. Needham, C. Prunty, and E. Yeah, "High volume microprocessor escapes; An analysis of defects our tests are missing," *International Test Conference (ITC)*, pp. 5-34, Oct. 1998.

32. T. Miller, J.M. Soden, and C.F. Hawkins, "Diagnosis, Analysis, and Comparison of 80386EX I_{DDQ} and Functional Test Failures," IEEE I_{DDQ} *Workshop*, pp. 66-68, Oct. 1995

33. D. Wilson, "Curve Tracer Data Interpretation for Failure Analysis," pp. 45-57, in ISTFA Electronic Failure Analysis. Desk reference, Editors S. Pabbisetty and R. Ross, ASM Int. Pub., 1998.

34. T. Tsang, J. Kash, and D. Vallett, "Time-resolved optical characterization of electrical activity in integrated circuits," *Proceeding IEEE*, pp. 1440-1459, Nov. 2000.

35. W. Lo, S. Kasapi, and K. Wilsher, "Comparison of laser and emission based optical probe techniques," *International Symposium on Test and Failure Analysis (ISTFA)*, pp 33 42, November 2001.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Submicron CMOS Devices

Theodore A. Dellin, © 2004, Dellin
Quick Start Micro Training, Tijeras, NM, www.quickstartmicro.com

Introduction

The transistor and the integrated circuit are probably the most significant technical inventions of the 20th Century. A transistor is an electrically controlled current source. An integrated circuit (IC) is a collection of transistors, along with their electrical wiring, that is designed to perform a specific function.

Although there are several types of semiconductors and transistors, silicon-based Metal Oxide Semiconductor (MOS) transistors dominate high-density Integrated Circuits (ICs). CMOS (Complementary MOS) ICs contain both n and p channel MOS transistors. CMOS is the mainstream IC technology because it does a better job of managing power consumption.

One of the most important features of MOS transistors is that as their dimensions and voltages are scaled down, the transistors are capable of higher clock rates and take up less area, i.e., become cheaper to produce. The IC industry is focused on introducing new IC manufacturing technologies every 2 to 3 years that reduce feature sizes on an IC by approximately 70%. This scaling, along with other improvements, allows each new generation of ICs to contain twice the number of transistors, running at twice the operating frequency, but at only half the cost per transistor.

After decades of scaling, the performance of CMOS ICs is being limited by fundamental physical and material limitations. There are also increasingly difficult trade-offs, most notably power versus performance. Sustaining historic rates of improvement literally requires reinventing the IC. The materials and device geometries historically used to make CMOS ICs need to be changed.

The purpose of this article is to provide an overview of how MOS transistors work, the major limitations to scaling, and the changes to transistors that have and will be used to meet these challenges. Readers are referred to the references at the end of the article for more details.

MOS transistor

An overview of the MOS transistor will be presented in this section. Figure 1 shows a schematic cross section of an n-channel MOS transistor. The transistor is formed in a region of the silicon semiconductor that is p-type. Using 'counter-doping', an n-type source and drain are formed on either side of the gate region. The gate region consists of a thin silicon dioxide layer thermally grown on the silicon surface. On top of this gate insulator is the gate electrode made from heavily doped ("metal-like") n-type poly-crystalline silicon.

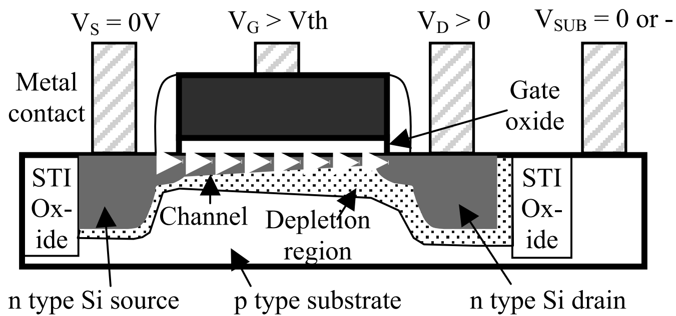


Figure 1: A schematic diagram of the cross section of an n channel MOS transistor in the 'on' (electron current conducting) state is shown.

The four primary dimensions of the transistor are the channel length 'L' between the source and drain depletion regions, the oxide thickness 't_{OX}', the depth of the source and drain at the channel edge, and the width of the transistor 'W' (the depth into the page in Figure 1).

The transistor consists of two pn junctions between source/substrate and drain/substrate, a MOS capacitor in the channel, and several metal/semiconductor junctions. The transistor is a four terminal device having electrical connections to the source, drain, gate and substrate. By convention, the source is usually taken as ground (0V) and thus all voltages mentioned are measured relative to the source. All pn junctions in the transistor are always at 0 bias or reversed biased (n side more positive than p side).

The MOS transistor uses the gate voltage to conduct or not conduct an electron current between the source and drain. When the gate voltage, V_G, is greater than a "threshold" voltage, V_{TH}, a thin surface layer of the p type substrate is inverted into an n type, electron-conducting channel. When the gate voltage is less than threshold the channel does not exist and there is essentially no current flow between source and drain. The drain voltage, V_D, causes current to flow from the source to the drain once the channel is created.

This article will only consider the n channel transistor. The other type of MOS transistor is the hole conducting p channel transistor. The p channel transistor is essentially equal and opposite of the n channel transistor. Relative to the n channel, the p channel transistor has opposite types of doping and the opposite polarity of the voltages. The one asymmetry is in the mobility and velocity saturation described below. Electrons move more easily across the channel than holes and this means that, everything else being equal, a p channel transistor will have one half to one third the current of a similar n channel transistor.

Semiconductors

Semiconductors have special electrical properties. Most importantly, we can modify those electrical properties both during manufacturing (by doping) and during operation (the field effect). Silicon is the dominant semiconductor material for CMOS. Silicon has many desirable attributes as a microelectronics material. Most importantly, silicon dioxide (or 'oxide') is an excellent insulator for CMOS transistors.

In solids, electrons are confined to ranges of energies, called bands. In semiconductors we need to be concerned with the almost full, lower energy valance band and the almost empty, higher energy conduction band. The top of the valance band is separated from the bottom of the conduction band by a bandgap energy, E_G. The band gap energy decreases slightly with increasing temperature.

Current is carried in the conduction band by the motion of mobile, negatively charged electrons. In the valance band, current can be considered to be conducted by the motion of mobile, positively charged holes. (A hole is a missing electron in the valence band.) For the rest of the article when we talk about electrons we are referring to electrons in the conduction band and when we talk about holes we are referring to holes in the valence band.

As pictured in Figure 2 we can have intrinsic, n type or p type semiconductors. In a pure (intrinsic) semiconductor the conduction-band electrons and valance band holes are thermally generated and are equal in concentration. The concentration of electrons and holes depends on the temperature and the semiconductor's bandgap.

$$n_i = p_i = \sqrt{N_C N_V} e^{-E_G/2kT} \quad (1)$$

where n and p are electron and hole concentrations, the i subscript indicates an intrinsic material, N_C and N_V are the effective density of states in the conduction and valence bands, E_G is the bandgap energy, k is Boltzmann's constant and T is the absolute temperature. At room temperature in Si $n_i \sim 10^{10}$ carriers/cm^3.

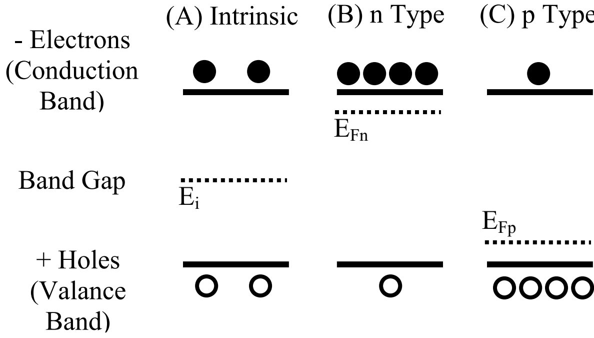


Figure 2. A band diagram of (A) Intrinsic, (B) n Type and (C) p Type semiconductors indicating the relative concentration of negative electrons in the conduction band and positive holes in the valence band is shown. The position of the Fermi Level(Ex) is also indicated in each case.

It is important to know the location of the Fermi level (E_{Fi}). In an intrinsic semiconductor:

$$E_{Fi} = E_i = \frac{E_V + E_C}{2} - \frac{kT}{2} \ln\left(\frac{N_C}{N_V}\right) \quad (2)$$

where E_V is the energy at the top of the valence band and E_C is the energy at the bottom of the valence band.

To make electron-rich "n type" silicon we replace a Si atom in the crystal with its 4 valence electrons with a phosphorus (P) or arsenic (As) "donor" atom with 5 valence electrons. Each "donor" atom donates one extra electron to the conduction band. In practical devices the concentration of electrons due to donor atoms, N_D, is much greater than the concentration due to thermal generation. Thus the concentration of electrons, n_n, is determined by the donor concentration.

$$n_n \approx N_D \quad (3)$$

The increased number of electrons due to the donor atoms results in increased recombination of electrons in the conduction band with holes in the valance band. The number of holes is reduced by this increased recombination. At equilibrium (e.g., no current flowing) the Law of Mass Action applies.

$$np = n_i^2 \quad (4)$$

This holds for both intrinsic and extrinsic semiconductors. It follows that at equilibrium the hole concentration in the n type material, p_n, is

$$p_n = n_i^2 / n_n = n_i^2 / N_D \quad (5)$$

Thus in a n type semiconductor there are many more electrons and much fewer holes than would be found in the corresponding intrinsic semiconductor as shown in Figure 2(B).

The Fermi level measured from the intrinsic level is

$$E_{Fn} - E_i = \frac{kT}{q}\ln\left(\frac{N_D}{n_i}\right) \quad (6)$$

In n type semiconductors the Fermi Level is above midgap and increases as the log of the doping concentration. We can derive similar equations for p type silicon that is formed by doping with boron (B). The Fermi level in the p type material is below midgap, nearer the valence band.

Extrinsic semiconductors are charge neutral. In n type material the negative charge of the extra electrons is compensated by the positive charge of the ionized donor ions. The electrons are mobile and can conduct an electrical current. The ionized donor ions are locked in the silicon lattice and thus cannot produce a current.

There are two physical processes that can produce currents. First, drift currents result from the net motion of electrons or holes in an electric field. Drift is the major conduction mechanism in the on state of mos transistors. At low fields, the electron and hole drift currents are given by

$$\mathbf{J}_{n,drift} = qn\mu_n\mathbf{E} \quad (7)$$

$$\mathbf{J}_{p,drift} = qp\mu_p\mathbf{E} \quad (8)$$

where n (p) is the electron (hole) concentration, q is 1.6E-19 coulombs, **E** is the electric field (V/cm), μ_n is the electron mobility (cm^2/V), and μ_p is the hole mobility (cm^2/Vs). The mobility in the channel is less than the mobility in the bulk due to surface scattering. In addition, the electron mobility is about 3 times the hole mobility. Mobility will be discussed in more detail below. Over the operating ranges of interest the mobility of electrons and holes decreases with temperature. This results in a decrease in the maximum operating frequency with temperature and, thus, can lead to speed failures at high temperatures.

The second transport mechanism we want to consider is diffusion. Diffusion results when there is a concentration gradient (i.e., when there is more of one species at a point than there is in an adjacent point). The diffusion currents of electrons and holes are given by

$$\mathbf{J}_{n,diff} = qD_n\frac{dn}{dx} \quad (9)$$

and

$$\mathbf{J}_{p,diff} = -qD_p\frac{dp}{dx} \quad (10)$$

where D_n and D_p are the electron and hole diffusion coefficients (cm^2/s).

The pn junction

A pn junction is formed around the interface between a p and n type semiconductor. The pn junction with an external voltage, V_{EXT}, applied across it is the basic building block of semiconductor devices. The MOS transistor has "one-sided" junctions that are formed between the heavily doped n type source (and drain) and a more lightly doped p type substrate. In this section the abrupt n+/p junction formed between two uniformly, but unequally, doped semiconductors will be considered.

As shown in Figure 3, there are three special features of the pn junction (barrier, depletion width and electric field) that each has an influence on the MOS transistor. First, an energy barrier exists across a pn junction with the p side being at higher electron energy. A voltage applied across the junction can raise or lower the barrier. The barrier height, E_J, is

$$E_J = (E_{Fn} - E_{Fp}) - qV_{APP} = qV_{bi} - qV_{APP} \quad (11)$$

where q is the magnitude of the charge on an electron (1.6E-19 C) and V_{APP} is the external voltage applied across the junction. V_{APP} is taken as positive for forward bias (p side connected to + side of voltage source). V_{bi} is the built in voltage drop across the junction that would occur with 0 external bias.

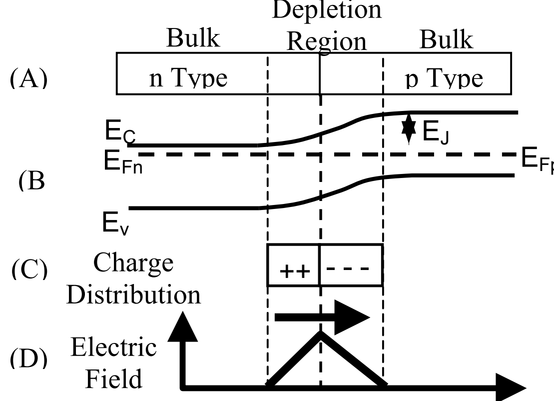


Figure 3: A pn junction showing the energy barrier, depletion region and electric field is diagrammed.

The ability of an external voltage to change the energy barrier height leads to the diode current voltage characteristics shown in Figure 4. Under forward bias the barrier is lowered and large diffusion currents flow. Under reverse bias the barrier is raised, the diffusion current is essentially zero, and only a small reverse leakage current flows. In general, both holes and electrons contribute to the diode current. However, for the n+/p one-sided junctions in an n channel transistor the

current is essentially all due to the motion of electrons. The current in this case is given by

$$J_n = \frac{qD_n n_i^2}{N_A L_n}\left[e^{(qV_{APP}/kT)} - 1\right] - \frac{qn_i W_d}{\tau_g} \qquad (12)$$

where N_A is the acceptor concentration in the p material, L_n is the electron diffusion length and τ_g is the generation lifetime (decreases as the generation rate increases). The first term on the right is due to diffusion of carriers entering the junction from the bulk regions and the second term on the right is due to carriers generated within the depletion region.

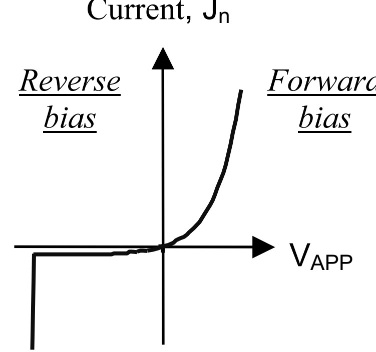


Figure 4: The current versus. voltage characteristics of a pn junction. Forward bias occurs when the p side of the junction is biased positive and the n side negative.

In the MOS transistor we want the reverse leakage currents to be as low as possible. Factors that increase the leakage current (and thus should be minimized) include increasing temperature and increasing the generation rate. The generation rate can be increased by temperature, radiation and/or by impurities (like Cu or Fe) that form midgap electron states.

Second, a depletion region exists. The depletion region consists of a dipole of charge (+ on the n side and – on the p side). For a one-sided n+/p junction the depletion width is essentially all inside the more lightly doped p side. The width of the depletion region for the one-sided junction is

$$W_d = \sqrt{\frac{2\varepsilon_{Si}(V_{bi} - V_{APP})}{qN_A}} \qquad (13)$$

where ε_{SI} is the permittivity of Silicon. Note that the width increases with increasing reverse bias and decreases with increasing doping density.

In the MOS transistor there are two back-to-back junctions. It is important that the depletions regions of the two junctions never meet. If they touch there will be an increase in leakage current called "punch through." Thus, as we scale down the dimensions of the MOS transistor we need to also scale down the depletion widths associated with the source and drain junctions. This requires increasing the doping level in the substrate.

Third, an electric field exists across the junction with a peak value

$$\mathbf{E}_{\mathbf{MAX}} = \frac{qN_A W}{\varepsilon_{Si}} \qquad (14)$$

The electric field points from the n to the p side of the junction. The electric field grows with reverse bias (since W grows with reverse bias). At large reverse biases the pn junction can breakdown and large reverse currents will flow as shown on the left side of the current curve in Figure 5.

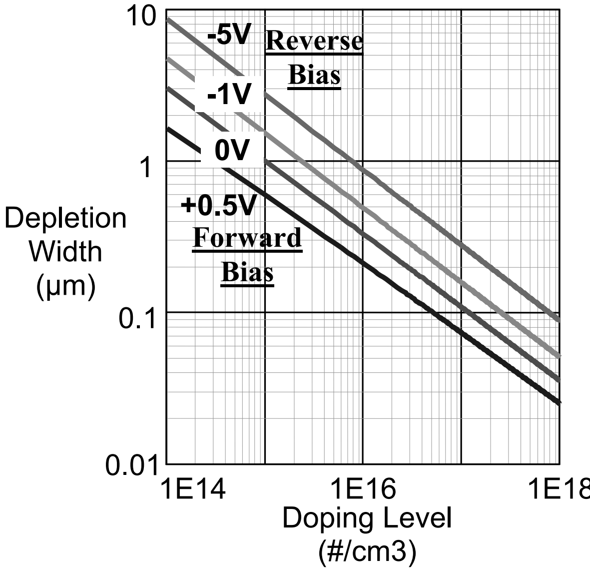


Figure 5: Calculated values of the depletion width in a one-side junction as a function of the doping on the lighter side of the junction.

The MOS capacitor

The next thing we need to investigate is the MOS (Metal Oxide Semiconductor) Capacitor. The capacitor consists of a top metal gate, an insulator, and an extrinsic semiconductor. There are two electrical connections to the gate and to the bottom semiconductor substrate. The substrate is grounded.

A voltage between the gate and the substrate can be used to control the charge near the surface of the semiconductor as shown in Figure 6 for a capacitor formed on p type silicon (as in the n channel MOS transistor). Assuming that the flat band voltage (discussed below) is 0, there is zero net charge when V_G = 0V as shown in (A). When a small positive gate voltage is applied, the holes are repelled from the silicon interface. The result is a depletion region with a negative charge due to the presence of the negative acceptor ions (that made it p type) as shown in (B). However, when the gate voltage is at or above a value called the threshold voltage a thin layer at the silicon surface becomes inverted from p to n type. This inversion layer of electrons is the channel in the n channel MOS transistor.

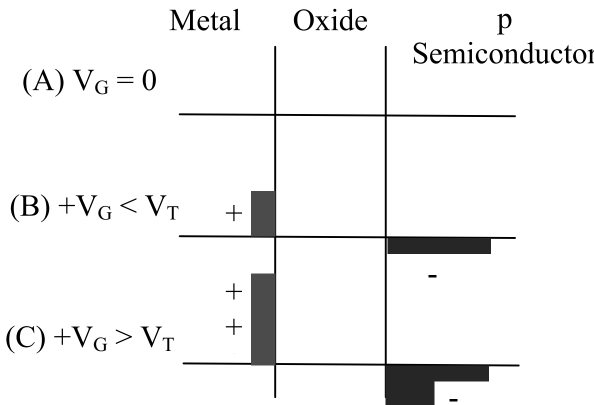


Figure 6. Charges in a MOS capacitor for different gate voltages.

Threshold voltage
One of the most important parameters of a transistor is its threshold voltage. The transistor turns on gradually and, thus, there is no precise threshold voltage. When *measuring* threshold voltage it is usual to plot source-to-drain current versus gate voltage and extrapolate the measured curve to give the gate voltage at zero current. In *calculations*, it is typical to use the definition that threshold occurs when the electron concentration in the channel at the surface is equal in magnitude to the hole concentration in the bulk of the p silicon. Using this definition the threshold is given by

$$V_{TH} = V_{FB} + 2\phi_b + \frac{t_{OX}\sqrt{2\varepsilon_{Si}N_A 2\phi_b}}{\varepsilon_{OX}} \quad (15)$$

where ϕ_B is the displacement of the Fermi level from midgap in the p type material

$$\phi_b = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) \quad (16)$$

and V_{FB} is the flat band voltage

$$V_{FB} = \Phi_{ms} - \frac{Q_{OX} + Q_{it}}{C_{OX}} \quad (17)$$

where Φ_{ms} is the difference between the work function of the gate electrode and the silicon, Q_{OX} is the equivalent positive oxide charge (from all sources) at the oxide/silicon interface and Q_{it} is the net charge due to interface states. Q_{it} depends upon the gate voltage. During manufacturing the threshold voltage is set by using a threshold voltage implant, which changes the value of Q_{OX} and thus changes V_{TH}. Figure 7 shows some calculated threshold voltages for an n channel (p substrate) with an n doped polysilicon gate and no oxide charges or interface states.

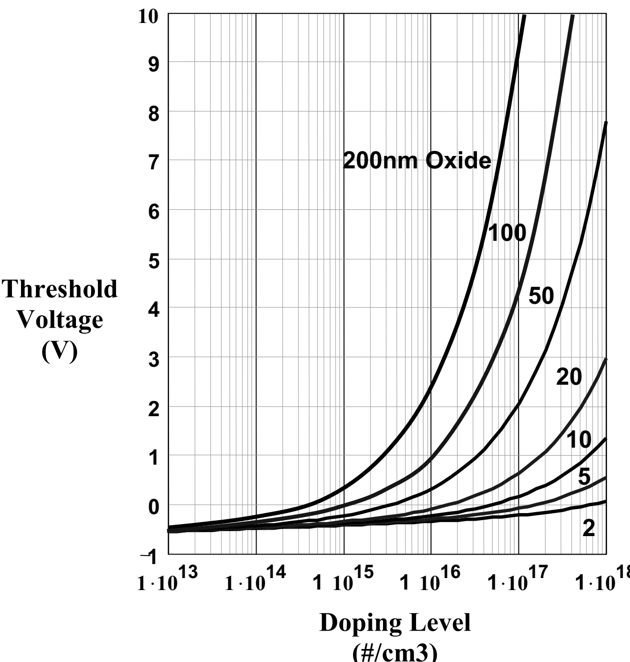


Figure 7: Calculated threshold voltages for an n channel MOS transistor with n+ poly silicon gate as a function of uniform substrate doping level for oxide thicknesses from 2 to 200nm.

The magnitude of the threshold voltage *increases* with:
(a) Increasing oxide thickness
(b) Increasing substrate doping
(c) Decreasing temperature ($dV_{TH}/dT \sim -1mV/K$)
Positive charges (from processing, threshold adjust implants with donor atoms or total dose irradiation) in the oxide or at the silicon surface make the threshold more negative. Negative charges (including acceptor implants) make the threshold more positive. Transistors made with p+ doped polysilicon gates have about a 1.1V (approximately equal to the band gap) more positive threshold voltage than n+ doped polysilicon gates.

The depletion region in the silicon under the gate grows with increasing gate voltage until the threshold voltage is reached and then grows very slowly with increasing gate voltage. The maximum width (thickness) of the depletion region is given by

$$W_{dm} = \sqrt{\frac{4\varepsilon_{Si}kT\ln(N_A/n_i)}{q^2 N_A}} \quad (18)$$

The long channel MOS transistor
The simplest MOS transistor is the "long channel" transistor. It would probably be better to call this the "relatively long channel" transistor since the key requirement is that the channel length, L, be much greater than either the maximum depth of the source/drain depletion regions at the edge of the channel and much

greater than the oxide thickness. Transistors with submicron channel lengths can be "long" if their channel length is long compared to the depletion depth. Figure 8 illustrates qualitatively how the transistor source to drain current, I_D, depends on the gate voltage and on the drain voltage.

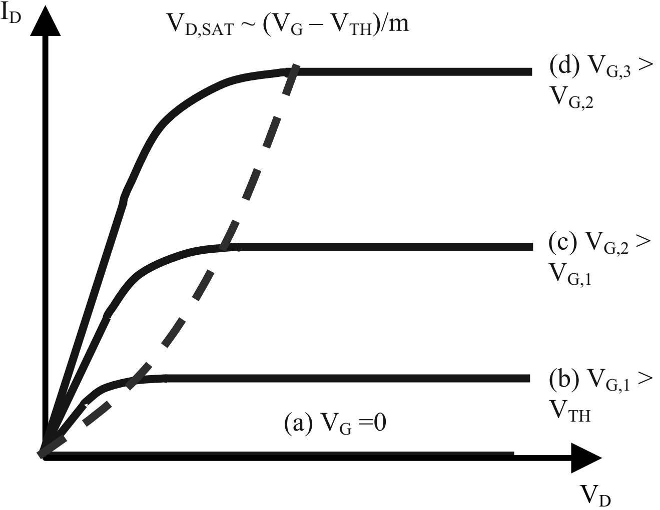


Figure 8. The qualitative shape of the source to drain current in an n channel MOS transistor versus drain voltage for several values of gate voltage.

Linear region

For small values of the drain voltage the current increases linearly with drain voltage (assuming the gate voltage is above threshold). In the linear region, the inversion layer will exist all the way across the channel from the edge of the source depletion region to the edge of the drain depletion region. The channel of the MOS transistor behaves like a resistor, with the resistance determined by the inversion layer. The formula for the current in the linear region is

$$I_{D,LIN} = \frac{\mu_{n,eff} W \varepsilon_{OX}}{L t_{OX}} (V_G - V_{TH}) V_D \qquad (19)$$

for $V_D << (V_G - V_{TH})$ and where $\mu_{n,eff}$ is the effective mobility of electrons in the inversion channel. (V_G-V_{TH}) is called the gate overdrive.

Saturation region

As the drain voltage is increased, the drain current vs. drain bias curve begins to saturate, i.e., I_D does not linearly increase with increasing V_D. This results from the effect of the drain voltage on the inversion layer near the drain end of the channel. The drain voltage reduces the size of the channel at the drain end. If the drain voltage exceeds a value V_{DSAT} the channel disappears (is "pinched off") at the drain end. This does not cause the flow of current to stop. Rather, what it does is to fix the voltage at the drain end of the channel at $V_{DSAT.}$ This prevents increasing drain voltage from increasing the transistor current. The extra drain voltage above V_{DSAT} is dropped between the end of the channel and the drain creating a region of very high electric fields. V_{DSAT} is given by

$$V_{D,SAT} = (V_G - V_{TH}) / m \qquad (20)$$

where

$$m = 1 + \frac{3t_{ox}}{W_{dm}} \qquad (21)$$

m typically has a value between 1.1 and 1.3.

The saturation current is given by

$$I_{D,SAT} = \frac{\mu_{n,eff} W \varepsilon_{OX}}{2mL t_{OX}} (V_G - V_{TH})^2 \qquad (22)$$

Notice that the saturation current of the long channel transistor depends on the square of gate overdrive.

Integrated circuits

Simple models can reveal how transistor properties affect the speed and power consumption of an integrated circuit. For speed, consider a p channel transistor whose drain drives a wire at the end of which is the gate of an n channel transistor. If it is assumed that at t=0 the p channel starts supplying a current I and that it takes a voltage V_G (>V_{TH}) on the gate of the n channel to turn it "on" then the delay time before the n channel transistor is turned on is

$$\tau = \frac{(C_{WIRE} + C_{GATE}) V_G}{I} \qquad (23)$$

From this two limiting cases can be considered. If the transistors are next to each other this is called "local interconnect." In this case, generally $C_{WIRE} << C_{GATE}$ giving

$$\tau_{LOCAL} = \frac{C_{GATE} V_G}{I} \qquad (24)$$

On the other hand, if the transistors are widely separated this is "global interconnect" where $C_{GATE} << C_{WIRE}$ giving

$$\tau_{GLOBAL} = \frac{C_{WIRE} V_G}{I} \qquad (25)$$

The frequency, f, of the IC is given by $1/\tau$.

For power considerations we treat the IC like a giant capacitor. "Active" power results from the movement of data through an IC (i.e., from charging and discharging the gates of transistors). Active power is given by

$$P_{ACTIVE} = fC_{ACTIVE}V^2 \quad (26)$$

where C_{ACTIVE} is the total capacitance being switched.

"Standby power" is consumed when the IC is not processing data

$$P_{STANDBY} = I_{LEAK}V \quad (27)$$

Where I_{LEAK} is the total leakage current from all sources. Note that standby power consumption is especially important in battery-powered applications.

Scaling
The scaling down of transistor dimensions and voltages leads to faster, better and cheaper ICs. Figure 9 illustrates "ideal" or constant (electric) field scaling. All dimensions and voltages are reduced by the same factor S (which is typically 70% for each new manufacturing technology generation). Relative to the previous generation scaling down by 70% reduces transistor area by half, increases the frequency for locally transmitted signals by 40% and keeps the active power per chip constant.

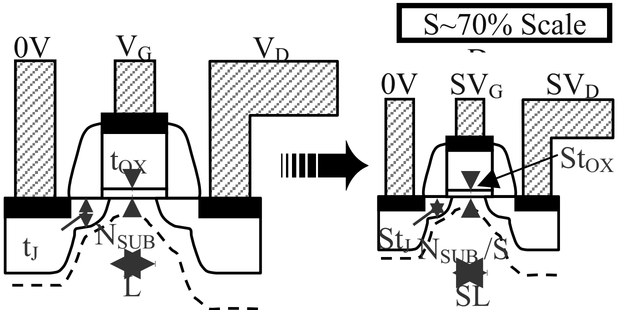


Figure 9. Constant Field Scaling.

Not everything gets better with scaling. The lowering of the threshold voltage increases the subthreshold leakage current and the thinning of the gate dielectric increases gate tunneling currents (described below). These increased leakage currents increase standby power consumption. Scaling also requires reducing the cross sectional area of, and spacing between, the metal wiring on the IC. This increases the resistance and capacitance of the wiring and slows down the global interconnect. In addition, the current density in the wiring increases which is bad for reliability.

Finally, ideal scaling is not always used. Often the voltage is scaled down more slowly than the dimensions. This results in higher transistor currents, which helps the speed of global interconnect. However, scaling the voltage more slowly is bad for both active and standby power consumption. Furthermore, in this scaling scenario the electric fields (voltage/dimension) increase which is bad for reliability.

Non-ideal behavior: leakage currents

So far the only current considered is the source to drain current that flows when $V_G > V_{TH}$. Currents that do not flow between the source and drain and the source to drain current when $V_G=0$ are parasitic leakage currents. Leakage currents lead to standby power consumption. Several types of leakage currents are possible as described in the next sections.

Subthreshold current
As the value of the gate voltage becomes less than the threshold, the carriers in the channel do not instantly go to the zero. Rather there is a region of "weak inversion" in which the carriers decrease approximately exponentially with decreasing gate voltage. The current is exponentially reduced as the gate voltage drops below threshold. Unfortunately, there is a fundamental physical limit to how fast the current can fall as shown in Figure 10. At room temperature it takes at least a 60mV reduction in gate voltage to reduce the drain current by a factor of 10. The leakage current when $V_G=0$ is given by

$$I_D(V_G = 0) = I_D(V_G = V_{TH})e^{-V_{TH}/kT} \quad (28)$$

As the threshold voltage is scaled down the subthreshold leakage current increases exponentially. Also note that the leakage current grows exponentially as the temperature is increased.

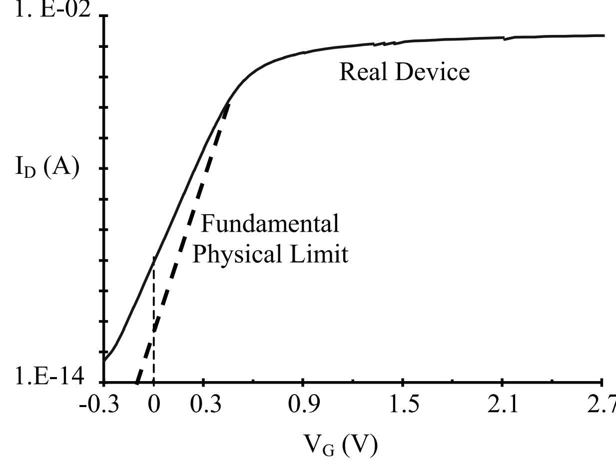


Figure 10: Drain versus gate voltage.

Short channel effect
In the ideal transistor the threshold voltage does not depend on the channel length or on the drain voltage. However, if the channel length is scaled down without scaling down the source, drain and gate insulator dimensions then the short channel effect shown in Figure 11 occurs. The threshold voltage is reduced from its ideal value by an amount that is approximately [1]

$$\Delta V_{TH} = \frac{24t_{OX}}{W_{dm}}\sqrt{V_{bi}(V_{bi}+V_{DS})}e^{-\pi L/(2W_{dm}+3t_{OX})} \quad (29)$$

Furthermore, the threshold voltage reduction increases with increasing drain voltage. This is called Drain Induced Barrier Lowering (DIBL).

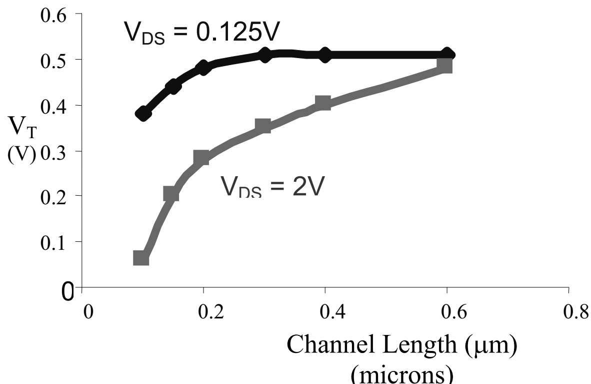


Figure 11. The decrease in threshold voltage when the channel length is reduced without scaling down the source, drain and gate insulator (short channel effect).

Gate insulator tunneling currents

At the silicon/silicon dioxide interface there is a large barrier (~3.2 eV) that prevents electrons from moving between the silicon and the gate. However, as the gates are thinned (oxides with less than 2nm thicknesses are in production) it is possible for electrons to quantum mechanically tunnel through the oxide. The gate leakage current increases exponentially as the oxide is thinned and results in increasing standby power. Figure 12 shows the 2003 ITRS projections for scaling of the effective oxide thickness and the resulting exponential increase in gate tunneling current. (The tunneling current for an oxynitride insulator is ~ 1/30 of that for a pure oxide).

Non-ideal behavior: reducing drive current

Source resistance

To control short channel effects the source and drain need to be shallow where they meet the channel. However, this leads to an increase in resistance and thus an increase in the voltage drop across the source and drain. The voltage drop at the drain end is not as serious since the devices are in saturation and the output current changes more slowly with drain voltage. However, the voltage drop across the source reduces the gate to source voltage and this has a strong impact on drive current. The effective gate to source voltage difference at the source/channel junction is

$$V_{GS,EFF} = V_G - I_D R_S \quad (30)$$

where R_S is the source resistance. This reduction in gate to source voltage reduces the drive current of the transistor which, in turn, reduces the frequency of the IC. Using selective epitaxial growth of silicon to produce raised sources and drains is one way to reduce the resistance.

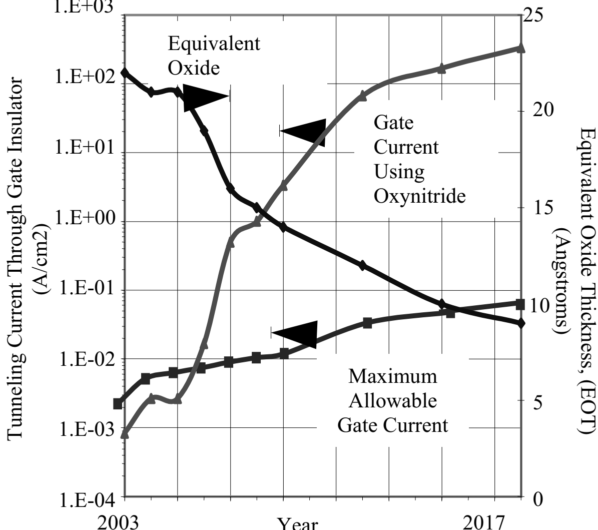


Figure 12. Projected values of the gate oxide thickness, the tunnel current using an oxynitride gate insulator and the maximum allowable gate current for low power applications from the 2003 ITRS Roadmap. (Source: Peter Zeitzoff, Sematech).

Mobility reduction and velocity saturation

The voltage is often scaled down more slowly than the dimensions. One consequence of this is to increase the electric fields (=voltage/dimension) within the transistor. The mobility is found to be dependent on the effective vertical electric field in the channel which can be approximated as

$$\mathbf{E_{EFF}} = \frac{V_{TH}+V_G+0.4}{6t_{OX}} \quad (31)$$

$$\mu_{n,EFF} = A\mathbf{E}^{-n} \quad (32)$$

where n is ~ .33 for E_{EFF} < 1 MV/cm and n is ~2 for E_{EFF} > 1 MV/cm.

At low horizontal electric fields along the channel the velocity of electrons is proportional to the electric field as discussed above. However, at large electric fields the velocity saturates. A simple treatment of velocity saturation gives the approximate expression

$$I_{DSAT} = C_{OX} W v_{SAT}(V_G - V_{TH}) \quad (33)$$

If we compare this equation with (22) we see two changes. First, with velocity saturation the current

depends on (V_G-V_{TH}) to the first power, not squared. Second, there is no 1/L dependence that causes I_{DSAT} to increase as the channel length is decreased.

Poly depletion and channel quantization

Thinning of the gate oxide is done to increase the gate to channel capacitance and thus to increase the transistor's current. Thinning the gate becomes relatively less effective in increasing the capacitance due to the effect of the small depletion region in the heavily doped polysilicon gate and to the fact that quantum mechanical effects require that the channel be below the silicon/silicon dioxide interface. These effects make the effective thickness of the oxide ~0.7nm greater than the physical thickness of the oxide. With the physical thickness of the oxide below 2.0 nm in leading-edge devices, this is a significant effect.

Transistor evolution

To mitigate the non-ideal effects described above a number of changes are being made to the geometry and materials used to make MOS transistors. Figure 13 shows what an advanced single gate MOS transistor might look like.

To reduce gate tunneling currents the transistor has a high k (dielectric constant) insulator which replaces the tradition silicon dioxide insulator. Note in Figure 13 the gate current through oxynitride exceeds the allowed limit for low power applications starting ~2006/2007 [2]. The higher dielectric constant allows a thicker dielectric while maintain the same gate to channel capacitance. The thicker dielectric (assuming it has sufficient barrier height to prevent electron and hole injection) can significantly reduce gate tunneling currents.

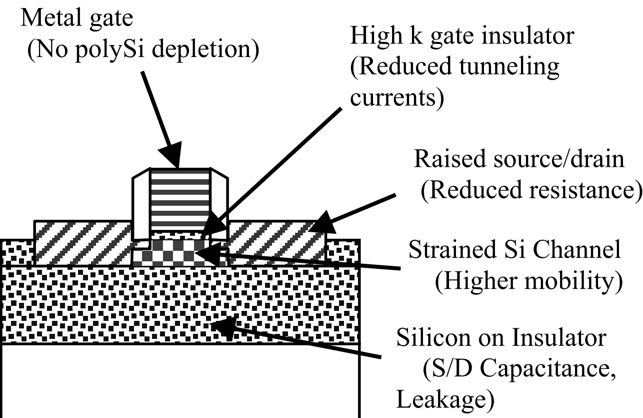


Figure 13. A cross section of an advanced MOS transistor.

A metal gate is used in place of a heavily doped polysilicon gate. The metal gate eliminates the capacitance reduction associated with the small depletion region in a polysilicon gate. In addition, a metal gate may be needed for material compatibility with certain high k insulators.

Selective epitaxial growth of silicon allows creation of raised sources and drains. The source and drain are still shallow next to the channel to minimize short channel effects. However, they are thicker away from the channel and this reduces the parasitic resistances of the source and drain.

The silicon in the channel is strained (displaced from its normal lattice spacing). This improves both the mobility and saturation velocity resulting in higher saturation currents. Higher currents lead to faster ICs. Straining can be accomplished by having SiGe under or adjacent to the channel. Alternatively, a highly strained silicon nitride layer above the gate can also strain the channel.

The transistor is built on a Silicon on Insulator (SOI) wafer. The bottoms of the source and drain touch the top of the buried oxide layer of the SOI wafer. This reduces the source (and drain) to substrate capacitance and also reduces junction leakage currents between the source/drain and substrate. The shallow trench isolation oxide between transistors also meets the buried oxide layer eliminating the possibility of latch up.

If the thickness of the silicon layer in which the MOS transistor is formed is less than the depletion width under the gate, the transistor is said to be fully depleted (otherwise it is partially depleted). In a fully depleted transistor a more lightly doped substrate can be used. This improves mobility. Most importantly, it reduces the subthreshold current to a value closer to its theoretical minimum thus reducing leakage currents and standby power. However, this requires being able to adjust the work function of the metal used for the gate electrode in order to get the desired threshold voltage.

Another alternative (not shown) for reducing the subthreshold slope would be to fabricate a transistor with two or more gates. A possible implementation of such a structure is the FINFET in which a gate is wrapped around a thin, vertical fin of silicon.

The text and figures in this article were adapted from material in the 21st Century Semiconductor Technology Handbook by Ted and Arlene Dellin, privately published, dellin@ieee.org, ©2000-04, all rights reserved

Suggested references

Due to space limitations a complete set of references has not been provided. More information, including detailed references, can be found in {note URL's may change):

1. ***An up-to-date reference with in-depth treatment of the subtleties and tradeoffs in deep submicron transistors.*** Taur, Yuan and Ning, Tak H., *Fundamentals of Modern VLSI Devices*, New York: Cambridge University Press, 1998.
2. The 2003 International Technology Roadmap for Semiconductors is online at http://public.itrs.net/.
3. Streetman, Ben G., *Solid State Electronic Devices, 5th Ed*, New Jersey: Prentice Hall, 1999.
4. ***Recent special journal issues of scaling challenges:***
   IBM Journal of R&D, Special Issue "Scaling CMOS to the Limit", Vol 46, No. 2/3, 2002, http://www.research.ibm.com/journal/rd46-23.html
5. Intel Technology Journal, Special Issue "Semiconductor Technology and Manufacturing – Current and Future", Vol 6, Issue 2, May 2002, http://developer.intel.com/technology/itj/2002/volume06issue02/index.htm
6. The author has an introductory, intuitive text on CMOS devices and processing, T. Dellin and A. Dellin, **21st Century Semiconductor Technology Handbook,** privately published, 2004, www.quickstartmicro.com.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Analog Device and Circuit Characterization

Steve Frank
Texas Instruments, Incorporated
Dallas, TX USA

INTRODUCTION

Characterization of microelectronic circuits for fault isolation and characterization of semiconductor devices for failure mechanism identification is an important part of a failure analyst's job. This paper covers the basics of analog device and circuit characterization with an emphasis on topics useful for failure analysis. The breadth of the subject matter does not allow for an in-depth treatment of every subject; however it is hoped that this paper gives a good introduction for the failure analyst new to the subject. The paper is divided into two sections. The first section gives a quick overview of characterization of semiconductor devices – namely PN junction diodes, bipolar junction transistors (BJTs), and MOSFETs (MOSFET stands for Metal-Oxide-Semiconductor Field Effect Transistor but modern MOS devices have polysilicon gates.) The second section discusses the characterization of analog circuits. These circuits are divided into five basic analog circuit blocks:

- Simple transistor circuits
- Current sources
- Voltage references
- Voltage regulators
- Op-amps

For each of these analog building blocks a simple description of its operation will be given, the DC characterization of each is outlined, and for many, possible failure modes are discussed.

ANALOG DEVICE CHARACTERIZATION

In this section we briefly look at the characterization of semiconductor devices used in analog circuits. It will by no means be a comprehensive treatment of semiconductor characterization, but focuses on the type of characterization commonly used in microelectronic failure analysis. With this in mind, the discussion will be limited exclusively to the DC characterization.

PN Junction Diode

The PN junction diode is the simplest semiconductor device that is encountered by a failure analyst; it consists of just one PN junction. PN junctions are found everywhere on integrated circuits – in diodes, BJTs, MOS transistors, diffused resistors, and ESD protection circuits. Characterization techniques used for PN junctions can be applied to these other areas as well.

Theory of Operation. Figure 1 shows the schematic symbol for a diode. A PN junction diode has three regions of operation: forward bias, reverse bias, and reverse breakdown. Figure 2 shows a measured I-V curve for a PN junction diode. The three regions of operation are marked. The forward bias region is where $V > 0$ and the current through the diode increases exponentially. The reverse bias region is where $V < 0$ and there is a small (essentially zero) current flowing through the diode. The reverse breakdown region is where the reverse bias voltage has exceeded the breakdown voltage of the device and the current through the diode rises sharply due to avalanche multiplication.

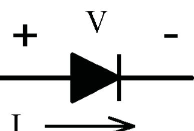


Figure 1. Schematic symbol of a diode showing the voltage and current conventions.

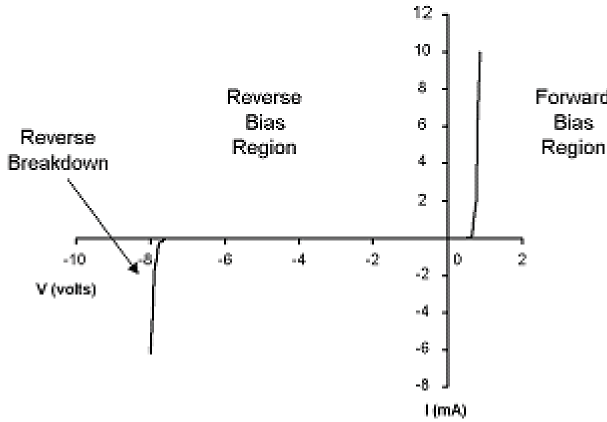


Figure 2. Measured I-V characteristics of a PN junction showing the three regions of operation.

Diode Characterization. The first characterization technique to be used on a PN junction should be a standard

curve trace of the junction. A standard curve trace is made using a curve tracer or parameter analyzer and results in an I-V plot similar to that found in Figure 2. A curve trace is a quick and easy technique and it uncovers many types of defects that result in current leakage across the junction. The reverse breakdown voltage is also easily measured [1, 2]. Figure 3 shows an example of a standard curve trace plot for a damaged junction. The I-V plot on the upper left portion of the figure shows that this PN junction is electrically behaving like a PN junction with a parallel resistive path to ground. After physical analysis was performed, the cause for the leakage was found. The lower right SEM photo shows that the damage to the junction was a result of ESD induced contact punch through.

In addition to the standard curve trace of the junction, the two diode parameters commonly measured are the saturation current and the ideality factor (n) [3, 4]. The current through a diode is related to the voltage across it by the following equation

$$I = I_S\left[\exp\left(\frac{V}{nkT}\right) - 1\right] \quad (1.)$$

where I_S is the saturation current, n is the ideality constant (normally ranges from 1-2), V is the applied voltage, k is Boltzman's constant, and T is the temperature in Kelvins.

Saturation Current and Ideality Constant Measurements. Measuring the saturation current and ideality constant is straightforward and can be accomplished using only one measurement. To do this, first approximate Eq. 1 by assuming that the diode is forward biased and $V >> nkT$. With these assumptions the I-V relationship is simplified to

$$I = I_S \exp\left(\frac{V}{nkT}\right) \quad (2.)$$

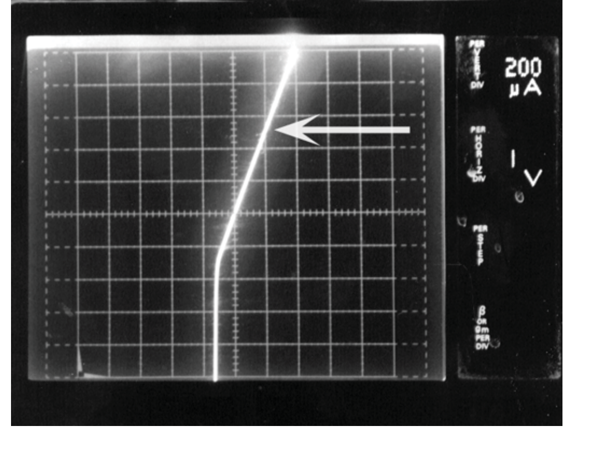
which can be rearranged into a more useful form

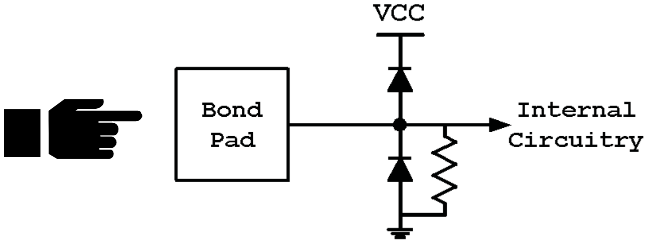
$$V = nkT \ln[I] - nkT \ln[I_S]. \quad (3.)$$

Eq. 3 is in the form of a line $(y = mx + b)$ from which the saturation current and the ideality factor can be extracted. The procedure to do this is as follows:

Set up a parameter analyzer or similar system to measure the I-V characteristics of the diode.

- Plot V vs. $\ln(I)$.
- The saturation current, I_S, is the x-intercept of the resulting curve.
- Measure the slope, *m*, of the curve.
- The ideality constant is *m/kT.*





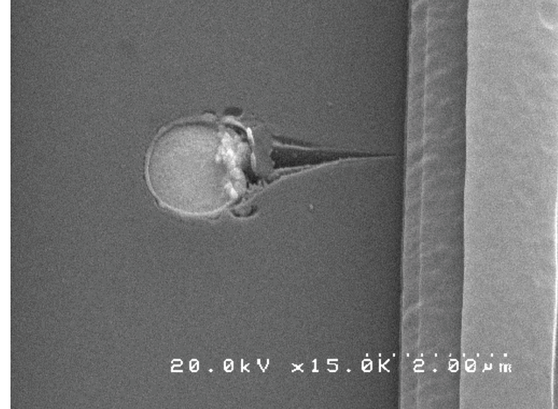


Figure 3. Standard curve trace of a damaged PN junction (this PN junction was part of the ESD protection circuitry.) The damage resulted in a parallel resistive path to ground. Physical analysis of the defect found it to be an ESD event resulting in contact punch through.

The ideality constant should range from 1-2. A measured ideality constant outside this range is indicative of an

abnormal junction [5]. Figure 4 shows an example measurement of a discrete diode. The saturation current is the x-intercept of the curve and can be read directly from the graph. For this diode, the saturation current measured 4 nA. The ideality constant is calculated from the slope of the linear portion of the curve. For this diode, the slope can be recorded directly from the graph and measures 113 mV. The ideality constant is then

$$n = \frac{m \log_{10}(e)}{kT} = 1.9 \tag{4.}$$

($\log_{10}(e)$ is a conversion factor to convert between base 10 logarithm and natural logarithm).

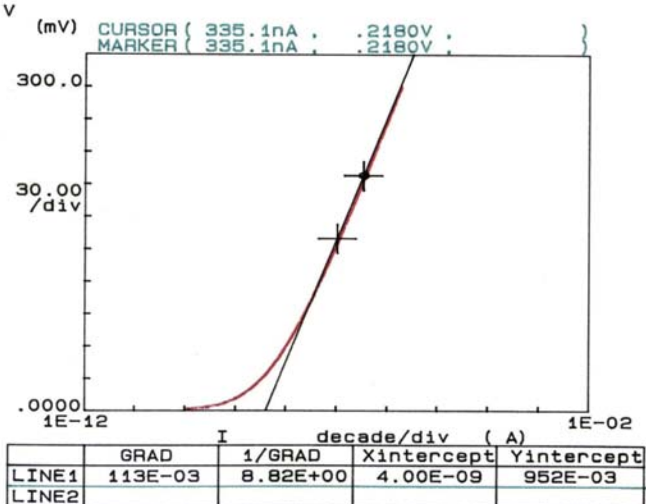


Figure 4. Measurement of the saturation current and ideality constant of a diode.

Bipolar Junction Transistors

The next semiconductor device examined is the bipolar junction transistor. BJTs are three terminal devices that are composed of two back to back PN junctions. Figure 5 shows the schematic diagrams for an NPN and PNP transistor showing the current and voltage conventions.

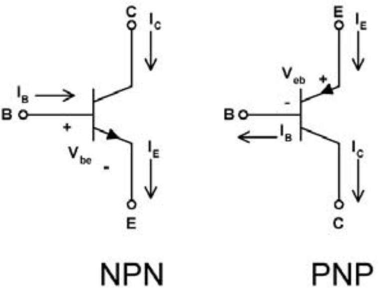


Figure 5. Schematic symbol for an NPN and PNP transistor showing the current conventions.

Theory of Operation. In DC operation, a bipolar transistor has three regions of operation – the cutoff region, the active region and the saturation region. The regions of operation are illustrated in Figure 6, which is a plot of the characteristic curves for an NPN transistor. The cutoff region occurs when the emitter-base junction and the base-collector junctions are both reverse biased. In this region the transistor is turned off and no collector current flows. In the active region, the emitter-base junction is forward biased and the base-collector junction is reverse biased. In this region of operation, the transistor has gain and acts like a current controlled current source. A small base current controls a large collector current. The terminal equations for the transistor operating in the active region are

$$I_C = I_S \exp\left(\frac{V_{BE}}{kT}\right) \tag{5.}$$

$$I_B = \frac{I_S}{\beta} \exp\left(\frac{V_{BE}}{kT}\right) \tag{6.}$$

$$I_E = \frac{\beta + 1}{\beta} \exp\left(\frac{V_{BE}}{kT}\right) \tag{7.}$$

where $\beta = I_C / I_B$ is the DC current gain of the transistor. Finally, in the saturation region both the emitter-base junction and the base-collector junction are forward biased. Here the collector current is dependent on the collector-emitter voltage.

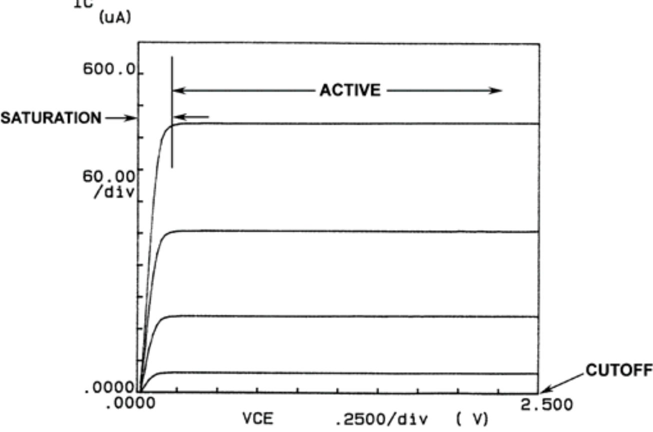


Figure 6. Characteristic curves of an NPN transistor showing the three regions of operation.

Bipolar Transistor Characterization. For failure analysis, there are three parameters that are useful to characterize. They are the characteristic I-V curves, the saturation current and ideality factor, and the DC current gain *(β)* [3, 4].

Transistor Characteristic Curves. Figure 6 shows typical characteristic curves for an NPN transistor. A curve tracer can be used to measure the characteristic curves of a transistor. Figure 6 was generated using a parametric analyzer that was set up to source a series of fixed base currents to the device and measure the resulting collector current as a function of the collector-emitter voltage.

Saturation Current and Ideality Constant. The measurement and calculation procedure of these parameters is the same as

for the diode; just replace I with I_C and V with V_{BE} in Eq. 1 and follow the diode procedure.

DC Current Gain (β). There are several ways to measure the DC current gain of a transistor:

- Use the characteristic curves. β is the ratio of the difference in collector currents to the difference in base currents for successive curves ($\beta = \Delta I_C / \Delta I_B$).
- Use a Gummel plot. Sweep the base voltage from 0 to 1V and plot I_C and I_B (log scale) vs. V_{BE} (linear scale). β is determined by the distance between the two curves. Figure 7 is an example of a Gummel plot. It shows the relationship between the base current and collector current for a discrete transistor. The measured β is 113.
- β can be measured by sweeping the base current of the transistor, measuring the resulting collector current and plotting I_C vs. I_B . β is the slope of resulting line.

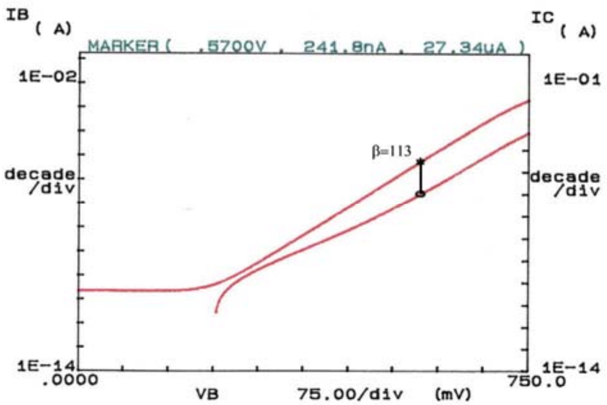


Figure 7. Gummel plot of an NPN transistor.

MOS Transistors

A MOS transistor is a four terminal device with a gate, source, drain, and back gate terminals. Figure 8 shows the schematic symbols for N –channel (NMOS) and P-channel (PMOS) MOS transistors.

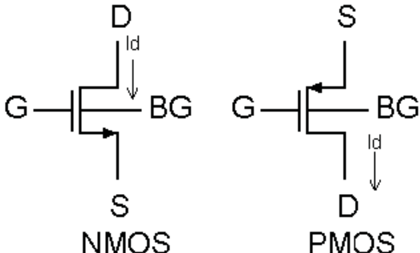


Figure 8. Schematic symbol for an NMOS and PMOS transistor showing the drain current flow convention.

Theory of Operation. In contrast to bipolar transistors where a large collector current is controlled by a small base current, in a MOS transistor the drain current is controlled by the gate voltage. Figure 9 shows a simple cross-section of an NMOS device. The gate of the transistor is polysilicon and is separated from the silicon by a thin dielectric layer (silicon dioxide). The source and drain are heavily doped n+ regions in a p–type substrate. The back gate is a contact to the p-type substrate. For a PMOS transistor the source and drain regions are heavily doped p+ regions of an n-type well. The back gate contact of the PMOS will be a contact to the n-well.

The MOS transistor works by creating a channel under the gate electrode so that current can flow between the drain and source terminals. The channel is created by inversion. For an NMOS transistor as the gate potential becomes more positive with respect to the source terminal positive charges accumulate in the gate and negative charges accumulate in the substrate beneath the gate. Thus the p-type region beneath the gate is "inverted" and a continuous n-type exists between the drain and source and current will flow. The gate voltage required to produce inversion is called the threshold voltage (V_t.)

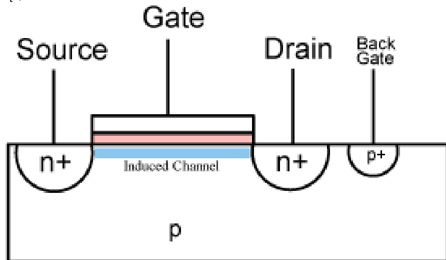


Figure 9. Cross-section of an NMOS transistor showing the gate, drain, source, and back gate contacts.

Like the bipolar transistor, there are three regions of operation for a MOS transistor (DC operation) – the subthreshold region, the triode region, and the saturation region. Figure 10 shows the characteristic curves of an NMOS transistor showing the triode and saturation regions of operation.

1. ***Triode Region:*** The triode region is characterized by

$$V_{GS} \geq V_t$$
$$V_{GD} \geq V_t .$$

V_{GS} is the gate to source voltage and V_{GD} is the gate to drain voltage. In this region the drain current is given by

$$I_D = \frac{\mu C_{OX}}{2}\frac{W}{L}\left[2(V_{GS} - V_t)^2 V_{DS} - V_{DS}^2\right] \qquad (8.)$$

where W and L are the width and length of the channel, V_{DS} is the drain to source voltage, C_{ox} is the capacitance per unit area of the gate, and μ is the electron mobility. This equation shows that in this region for small values of V_{DS}, a MOS transistor acts like a voltage controlled resistor with a resistance value proportional to the bias point, ($V_{gs} - V_t$).

2. ***Saturation Region:*** The saturation region is characterized by

$$V_{GS} \geq V_t$$
$$V_{GD} \leq V_t .$$

In this region the drain current is given by

$$I_D = \frac{\mu C_{OX}}{2}\frac{W}{L}(V_{GS} - V_t)^2 . \qquad (9.)$$

Here the MOS transistor acts like a voltage controlled current source with the drain current proportional to the bias point ($V_{gs} – V_t$).

3. ***Subthreshold Region:*** Although in the simplified model of the MOS transistor the drain current should be zero for a gate to source voltage less than the threshold voltage, there is a small non-zero drain current for values of V_{GS} that are less than V_t. The drain current in this region of operation is given by

$$I_D = \frac{\mu C_{OX}}{2}\frac{W}{L}e^{\frac{V_{GS}}{nV_t}}\left(1 - e^{\frac{-V_{DS}}{V_t}}\right). \qquad (10.)$$

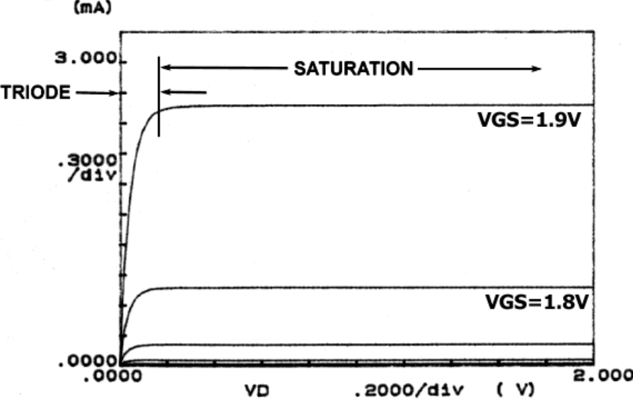


Figure 10. Characteristic curves of an NMOS transistor showing the triode and saturation regions of operation.

MOS Transistor Characterization. For failure analysis, it is often useful to be able to characterize a MOS transistor and determine its threshold voltage. There are several ways to measure the threshold voltage of a MOS transistor [3]. The easiest way is to measure the V_{GS} required to achieve a certain drain current, say 1uA. This V_{GS} is then said to be the threshold voltage.

Another method to measure the Vt of a MOS transistor is to plot the square root of the drain current as a function of V_{GS} with the transistor operating in saturation. Recall from Eq. 9 that

$$\sqrt{I_D} = \sqrt{\frac{\mu C_{OX}}{2}\frac{W}{L}}\left(V_{GS} - V_t\right) \qquad (11.)$$

which has the form of a line $(y = mx + b)$ with

$$m = \sqrt{\frac{\mu C_{OX}}{2}\frac{W}{L}}$$

$$b = -m \cdot V_t .$$

Therefore, to measure V_t with this method plot $\sqrt{I_D}$ vs. V_{GS} and measure the x-intercept of the linear portion of the curve. The x-intercept is V_t. Figure 11 is an example V_t measurement of an NMOS transistor.

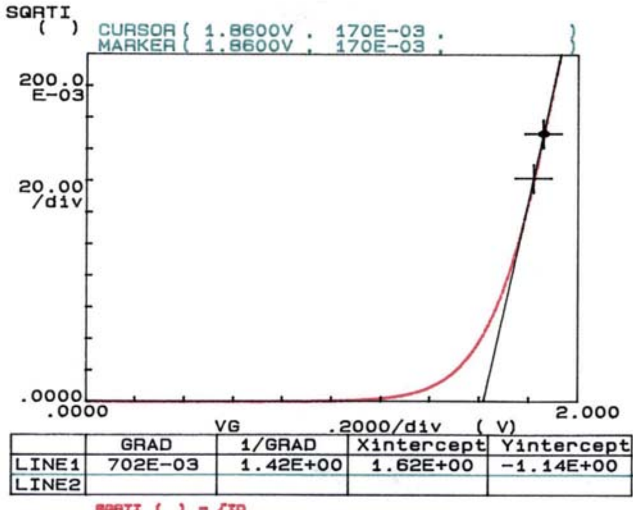


Figure 11. $\sqrt{I_D}$ vs. VGS for an NMOS transistor. The x-intercept shows that the threshold voltage of this transistor is 1.62 V.

ANALOG CIRCUIT CHARACTERIZATION

The remainder of this paper deals with characterization of analog circuits commonly encountered in failure analysis. The scope is limited to DC analysis of circuits.

Single Transistor Circuits.

The discussion of analog circuit characterization begins with the analysis of two simple single transistor circuits – the common collector and common emitter amplifiers with bipolar transistors and the common source and common drain amplifiers with NMOS transistors. Figures 12 and 14 show the schematics of the two circuits respectively. For each, *Vin* is the input voltage and *Out* is the output voltage.

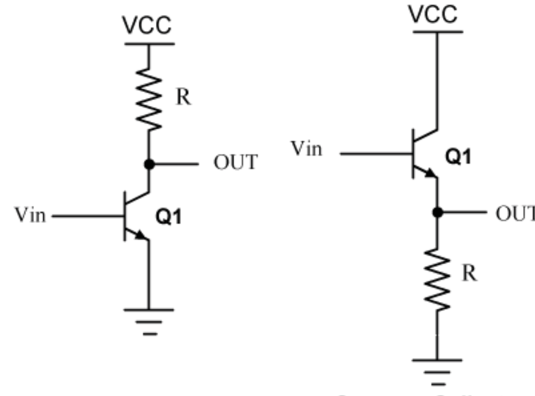


Figure 12. Schematic diagrams of a common collector and common emitter amplifier.

Common Emitter (CE) Amp. The common emitter amplifier acts like an inverter with three distinct regions of operation.

- *Region 1: Vin = 0V*
  - Q1 is cut-off.
  - IC1 = 0.
  - $Vout = V_{CC}$.
- *Region 2: Vin > 0V* and the base-collector junction is reverse biased.
  - Q1 is active.
  - $Vout = V_{CC} - I_S R\exp\left(\frac{Vin}{kT}\right)$.
- *Region 3: Vin > 0V* and the base-collector junction is forward biased.
  - Q1 is saturated
  - $Vout = V_{CEsat}$

Figure 13 is a measured DC transfer function for a common emitter amplifier clearly showing the three regions of operation. Operation in region 2 is required for amplification.

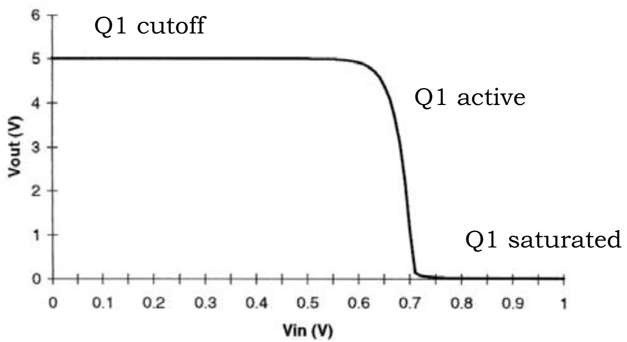


Figure 13. DC transfer function of a CE amplifier. The three regions of operation are marked.

Common Collector (CC) Amp. The common collector amp acts like a voltage follower with two distinct regions of operation.

- *Region 1: Vin – Vout < 0.7V*
  - Q1 is turned off.
  - $Vout = 0$.
- *Region 2: Vin –Vout > 0.7 V*
  - *Vout = Vin – 0.7V*

In region 2, the output of the CC amplifier is the input voltage minus a B-E voltage drop (typically 0.6 – 0.7 V).

Characterization for Failure Analysis. Characterization of analog circuits for failure analysis means "what type of characterization is needed to help isolate a failure". It has nothing to do with circuit characterization to determine if it complies with design specifications. With this in mind, there are two characterization strategies for the CC and CE amplifiers. First, a basic curve trace analysis of the transistor terminals with respect to ground and VCC is a good starting point. The I-V curves for each junction will detect leakage, shorts, and possibly opens in the circuit. Secondly, parametric testing of the circuit can be a method of characterizing CC and CE circuits. Parametric testing includes testing VOL, VOH for the outputs and IIH, IIL for the inputs. Table 1 details the response of a CE amp to several simulated defects. These defects are B-E leakage, C-E leakage, and the output clamped by a diode to ground. The table shows that not all tests are equally suited for all possible defect types. For detecting B-E leakage the input current tests are useful whereas for detecting C-E leakage, the output voltage, and IDDQ tests are useful (in this example IDDQ refers to the VCC current when Q1 is in region 1 of operation [cut-off].)

Table 1. Response of a CE amplifier to simulated defects. The (**) indicate detectable failure modes for each test.

Common Emitter

	GOOD	B-E leakage	C-E Leakage	Diode from output to ground
VOH (VI = 0.2)	5.0 V	5.0	450 mV (**)	700 mV (**)
VOL (VI = 1.0)	40 mV	40 mV	40 mV	40 mV
IIL (VIL = 0.2)	0.0	2.0 mA (**)	0.0	0.0
IIH (VIH = 1.0)	23 mA	35 mA (**)	23 mA	23 mA
ICCQ (vin = 0.2)	0.0	0.0	4.5 mA (**)	4.3 mA (**)

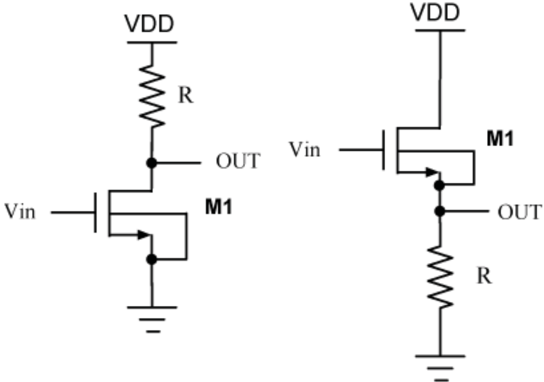


Figure 14. Schematic diagrams of a common source and common drain amplifier.

Common Source (CS) Amp. The common source amplifier acts like an inverter with three regions of operation (see Figure 15 for the DC transfer curve showing the regions of operation.)

- *Region 1: Vin < Vt*
  - M1 is cut-off.
  - $I_{D1} = 0$.
  - $Vout = V_{DD}$.
- *Region 2: Vin > Vt* and *Vin – Vout < Vt*.
  - M1 is in saturation.
  - $Vout = V_{DD} - \frac{\mu C_{OX}}{2} \frac{W}{L} R(V_{IN} - V_t)^2$.
- *Region 3: Vin > Vt* and *Vin – Vout > Vt*.
  - M1 is in triode region of operation
  - $Vout \approx 0$.

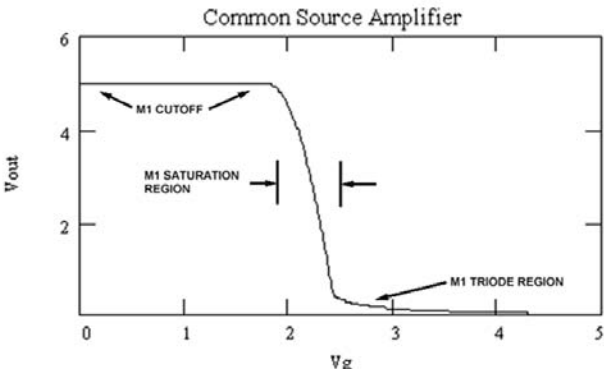


Figure 15. DC transfer curve of a Common Source Amplifier showing the three regions of operation.

Common Drain (CD) Amp. For DC inputs, a common drain amp acts like a voltage follower with two distinct regions of operation.

- *Region 1: Vin – Vout < Vt*
  - M1 is turned off.
  - $Vout = 0$.
- *Region 2: Vin –Vout > Vt*
  - *Vout = Vin – Vt*

Characterization for Failure Analysis. The characterization of common source and common drain amplifiers follows the analysis given for CE and CC amplifiers with one addition. The input current for these MOS amplifiers should be zero and any current measured is an indication of gate oxide leakage.

Current Sources.

Current sources are used for biasing and act as active loads. An active load would take the place of the resistor in a CE or CS amplifier. There are too many types of current source topologies to describe them all, so the following discussion will be limited to one simple current source. The basic functionality of various current sources are the same, they differ mainly in the implementation details.

Theory of Operation. Figure 16 shows a schematic of a simple bipolar current mirror. To simplify the analysis of this circuit, neglect the base current required for Q1 and Q2. The bases of the two transistors are tied together so that $V_{BE1} = V_{BE2}$. Since $I_{REF} = I_{C1}$ and $I_C = I_S \exp\left(\frac{V_{BE}}{kT}\right)$ then

$$I_C = \frac{I_{S2}}{I_{S1}} I_{REF} \text{ [4].} \quad (12.)$$

I_C is the output of the current source that is used to bias other circuits. It is generated from I_{REF} and is scaled by the factor $\frac{I_{S2}}{I_{S1}}$.

For a MOS current mirror, the result is similar with the scaling factor being the ratio of the transistor sizes

$$I_D = \frac{\left(\frac{W_2}{L_2}\right)}{\left(\frac{W_1}{L_1}\right)} I_{REF}. \quad (13.)$$

W and L are the channel width and length of each respective transistor.

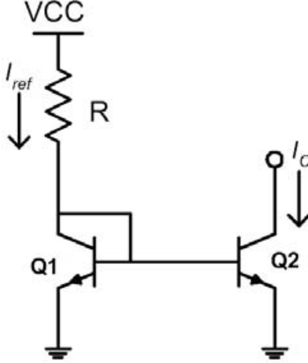


Figure 16. Simple bipolar current mirror.

Characterization for Failure Analysis. Characterization of current sources is straightforward. Usually it involves just measuring the bias current and comparing it to either a known good current source or to the designed value. Current sources that fail generally fall into two categories – they fail catastrophically or fail to have the correct bias current (parametric fails). Catastrophic failures may be due to EOS or gross misprocessing (open contacts, etc.) Parametric fails generally are caused by some type of nonideality (e.g. low β) or mismatch in the transistors that comprise the current source.

Voltage References.

A voltage reference is used to supply a stable voltage that is independent of temperature and power supply variations.

Bandgap Voltage Reference. Figure 17 is a block diagram of a bandgap voltage reference. The reference voltage is derived from the sum of two components – a base-emitter voltage and a voltage that is proportional to the thermal voltage. Since the base-emitter voltage of a bipolar transistor has a negative temperature coefficient (-2mV/°C) and the thermal voltage has a positive temperature coefficient (+ 0.0853 mV/°C), a constant is chosen which results in a zero temperature coefficient.

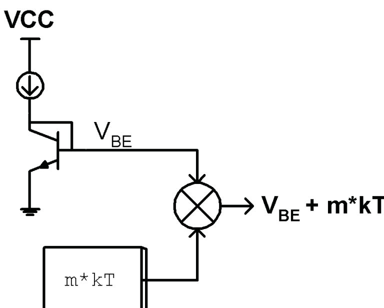


Figure 17. Block diagram of a bandgap reference voltage.

Theory of Operation. Figure 18 is a schematic of a Widlar bandgap voltage reference circuit. The node 'Vout' is the output of the voltage reference. Assuming the transistors have no base current the following current and voltage equations can be written.

$$Vout = V_{BE3} + I_{C2}R_2 \quad (14.)$$

$$I_{C2} = \frac{V_{BE1} - V_{BE2}}{R3} = \frac{\Delta V_{BE}}{R3} \quad (15.)$$

$$\Delta V_{BE} = kT \ln\left(\frac{I_{C1}}{I_{C2}}\frac{I_{S2}}{I_{S1}}\right) \quad (16.)$$

combining Eqs. 9-11 yield

$$Vout = V_{BE} + \frac{R2}{R3} kT \ln\left(\frac{I_{C1}}{I_{C2}}\frac{I_{S2}}{I_{S1}}\right) \quad (17.)$$

if R1 and R2 are chosen so that $I_{C1}R1 = I_{C2}R2$ then $V_{BE3} = V_{BE2}$ and

$$Vout = V_{BE3} + \frac{R2}{R3} kT \ln\left(\frac{R2}{R3}\frac{I_{S2}}{I_{S1}}\right) \text{ [4, 6].} \quad (18.)$$

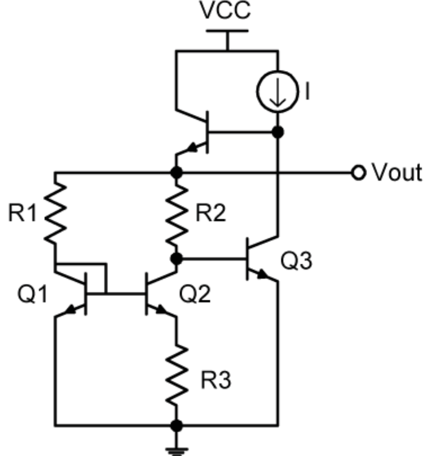


Figure 18. Schematic diagram for a bandgap reference voltage.

Characterization for Failure Analysis. From a failure analysis point of view, voltage references are not difficult circuits to characterize. You simply measure the output voltage. The only complication is that it may also be necessary to measure the output voltage over temperature. Failure modes associated with voltage references generally fall into two categories – no output voltage or an output voltage that is incorrect or doesn't track correctly over its specified temperature range. Bandgap voltage references operate in a feedback configuration and will have two stable states – one stable state is at the correct output voltage and the other when the output is zero. The existence of a stable output state with $Vref = 0$ requires a voltage reference to have a start-up circuit to 'nudge' the output into the correct stable state. If a reference voltage reads zero volts, then characterize the circuit as follows:

- Perform a standard curve trace from VCC to ground to detect any leakage current that may be due to catastrophic damage such as electrical overstress.
- If no leakage current is measured, characterize the start-up circuitry to see if it is working properly. Depending on the circuit architecture and layout, this may or may not be possible. If it is not possible to isolate the start-up circuitry, measure node voltages and compare them to a good unit.

Voltage references that have an incorrect output voltage or a reference voltage that doesn't track properly over temperature require different characterization strategies.

- Voltage references are typically trimmed to the correct voltage at wafer probe. The trimming process can take a variety of forms, but one common method is to trim resistor values via fuses

or zener diodes. An improperly blown fuse or zener diode may experience some unexpected behavior over time and temperature. The first step is to characterize the trims over temperature and see if they are stable.

- Frequently ΔV_{BE} is set up by a ratio of resistor values. If one of these resistor values is grossly off, say by an open contact, the reference voltage will be incorrect.
- A high input offset voltage in the summer circuitry will also affect the reference voltage. The characterization of input offset voltage is discussed in section on op-amps.

Voltage Regulators.

A voltage regulator outputs a known voltage that is stable over a wide range of output current [7].

Theory of Operation. Figure 19 is a block diagram of a series voltage regulator. Vo is the regulated output voltage. The sampling network, consisting of a voltage divider (R1 and R2) samples a portion of the output voltage and feeds it back to the error amplifier. The error amplifier is a differential amplifier (usually an op-amp) which compares this sampled voltage to a stable reference voltage and amplifies the difference of the two signals. If the gain of the error amp is high enough, the feedback loop will force the voltages at its inputs to be equal (V+ = V-). This results in the output of the voltage regulator being

$$Vo = \left(1 + \frac{R1}{R2}\right) Vref \ . \qquad (19.)$$

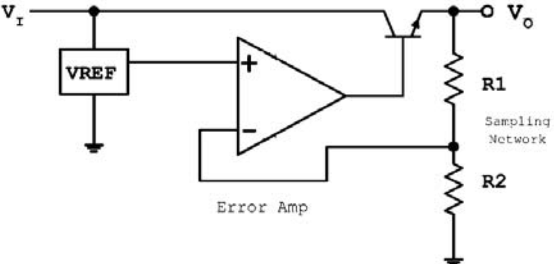


Figure 19. Block diagram of a series pass voltage regulator.

Characterization for Failure Analysis. When a voltage regulator requires failure analysis, a common complaint is that the output voltage is incorrect or the unit is not regulating. There are other possible failmodes for features not shown on the block diagram, such a overcurrent limiting, but these will not be discussed. The first step in characterizing a voltage regulator is to do an output vs. input plot. From this, useful information is gathered that helps determine the direction of future characterization.

- ***The regulator does not regulate.*** This means that the output voltage is a function of the input voltage for all values of input voltage. If the regulator doesn't regulate, then the next step is to curve trace the regulator between the input and output pins. Many times the pass transistor is damaged or shorted and the curve trace will quickly find this type of damage.
- ***The regulator regulates, but does so at an incorrect voltage.*** This type of failure generally indicates that the feedback loop is operating properly, but there is an error in one of the components that make up the regulator. The contribution of each of the possible error sources in a voltage regulator is discussed in the following section.

Error Sources in a Voltage Regulator. There are several possible error sources found in voltage regulators.

1. The reference voltage could be incorrect. The effect on the regulator's output due to this error is

$$\Delta V_o = \left(1 + \frac{R1}{R2}\right) \Delta Vref \ . \qquad (20.)$$

2. There could be errors associated with the sampling network. The effect on the regulator's output due to this error alone is

$$\Delta V_O = \frac{R2}{R1}\left(\frac{\Delta R2}{R2} - \frac{\Delta R1}{R1}\right) Vref \ . \qquad (21.)$$

3. Errors associated with the error amplifier will also have an effect on the regulator's output voltage.

$$V_O = \frac{Vref - V_{OS}}{\frac{1}{A_V} + \left(\frac{R2}{R1 + R2}\right)} \qquad (22.)$$

where A_V is gain of the error amplifier and V_{OS} is the input offset voltage of the error amplifier.

Op-amps.

The final analog building block is the op-amp. Op-amps are either stand-alone integrated circuits or may be found internally on large analog designs.

Theory of Operation. An op-amp is a differential amplifier whose output voltage is given by

$$Vout = A_v(V_+ - V_-) \qquad (23.)$$

where A_V is the gain of the amplifier and V_+ and V_- are the non-inverting and inverting inputs, respectively [4]. Although op-amps may have a variety of topologies, two and three stage op-amps are most common. In a three-stage op-amp the first stage is a differential amplifier that provides gain and often performs a differential to single ended conversion (some op-amps also have differential outputs.) The second stage is a gain stage and usually has a compensation network to ensure stability. The third stage, if present, is a buffer that supplies the drive current needed to drive external loads.

An ideal op-amp has these qualities:

- Infinite gain.
- No input bias current.
- Zero output resistance.

Unfortunately, real op-amps are not ideal and these aspects of op-amp behavior need to be understood by the failure analyst.

Characterization for Failure Analysis. Real world op-amps depart from the ideal and have

- Finite gain (Av).
- Input bias current (IIB).
- Input offset current (IOS)
- Input offset voltage (VOS).
- Non-zero output resistance (Ro).

Figure 20 illustrates these departures from ideality. Failures of these and other parameters (e.g. output voltage swing, common mode rejection ratio) are likely to be encountered by a failure analyst who works on op-amps. The following sections outline the characterization techniques for parameter failures most likely to be encountered.

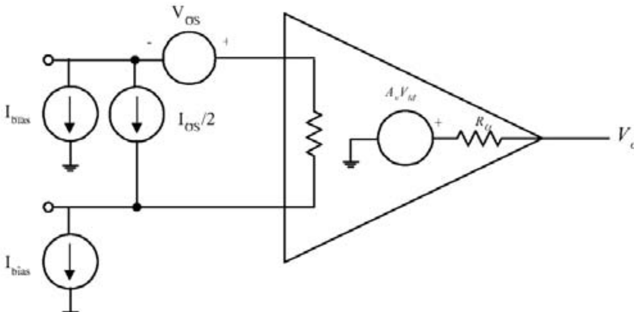


Figure 20. Block diagram of an op-amp showing the non-idealities.

Input bias current (IIB). The input bias current for bipolar op-amps is the base current required for the input differential pair. For example, Figure 24 shows a simplified schematic of an input differential pair. For the bipolar case, the input bias current is the base current required to bias Q1 and Q2 to $IEE/2$ (plus any possible leakage in the ESD protection circuitry). For the MOS differential amplifier, the input bias current should be essentially zero. To characterize IIB, use the following procedure:

- Do a standard curve trace on the input pins to see if there is any leakage current in the ESD structure. Finding leakage on a curve trace does not uniquely isolate the leakage to the ESD structure, but not finding leakage eliminates the ESD structure as being a potential failure site.

- If leakage current is measured during the standard curve trace, isolate the ESD structure from the op-amp circuitry. If the leakage goes away, then the failure is isolated to the ESD structure. If the leakage is still present, characterize the PN junctions of the input transistors to isolate the failure mechanism. For MOS input transistors, the leakage is due to gate oxide damage.

- If a standard curve trace does not measure any leakage current, perform an IIB test to verify the failure. The test set-up will usually be found in the op-amp specifications. If the IIB failure is verified, characterize the input transistors (possibly for low β for bipolar inputs.)

Input offset voltage (VOS). The input offset voltage is the input differential voltage required to bring the output voltage of the op-amp to zero (assuming positive and negative power supplies). The input offset voltage of an op-amp can range in magnitude from tens of microvolts to tens of millivolts. There are several different ways to measure the input offset voltage. The first is to configure the op-amp as a non-inverting amplifier whose inputs are grounded through resistors (see Figure 21). In this configuration VOS is given by

$$VOS = \frac{Vout}{\left(1 + \frac{R2}{R1}\right)} . \qquad (24.)$$

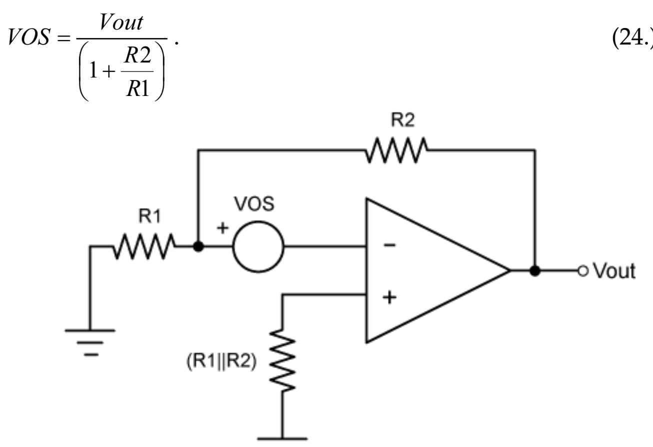


Figure 21. VOS measurement of an op-amp using a closed loop configuration.

The second method measures VOS directly in an open loop configuration. Figure 22 shows a setup diagram for this measurement. A low value resistor (~10Ω) is connected across the input terminals of the op-amp. The inverting input is grounded and the non-inverting input is driven with a current source.

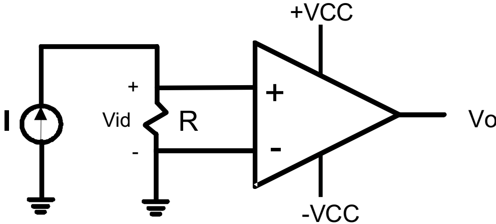


Figure 22. Measurement set-up for VOS using an open-loop configuration.

In this set-up the input voltage is generated by the voltage drop across the resistor and is given by

$V_{id} = IR$ (25.)

Vo is plotted against Vid and the resulting curve will look similar to Figure 23. For this op-amp, the input offset voltage is 800 μV. This method is also useful in determining the DC gain of the op-amp – it's just the slope of the linear portion of the curve as plotted in Figure 23.

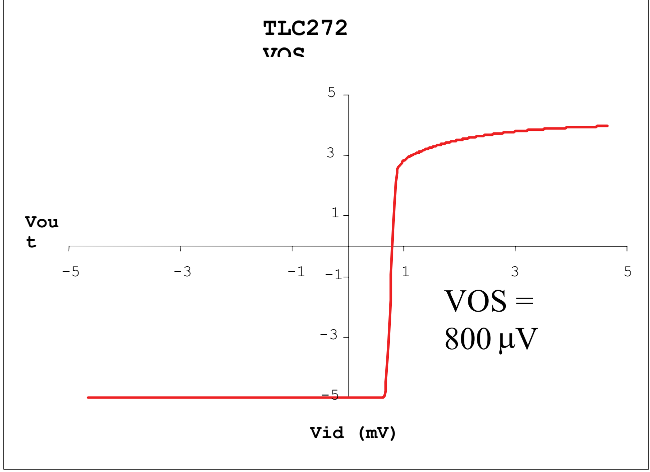


Figure 23. Measured VOS for a TLC272 op-amp using the open-loop method.

Factors that affect VOS. The factors that affect VOS are important to consider and will help an analyst isolate the cause of a VOS failure. Figure 24 is a schematic of a simple input differential pair. The input offset voltage is modeled as a voltage source in series with one of the inputs. For the bipolar differential pair and using a circuit loop consisting of VI1, VI2, Q1, and Q2 we have

$$VOS = V_{BE1} - V_{BE2}. \quad (26.)$$

Using Eq. 5 and solving for V_{BE} results in

$$VOS = kT \ln\left(\frac{I_{C1}}{I_{C2}} \frac{I_{S2}}{I_{S1}}\right). \quad (27.)$$

Recall that VOS is the amount of differential input voltage required to bring the output of the op-amp to zero. For the input differential stage in Figure 24, the output is VOD, which is $VO1 - VO2$. To make this output voltage to equal zero requires that $I_{C1}R1 = I_{C2}R2$. When this is satisfied, the offset voltage is

$$VOS = kT \ln\left(\frac{R2}{R1} \frac{I_{S2}}{I_{S1}}\right). \quad (28.)$$

By defining new variables, Eq. 23 can be derived in a way that is particularly helpful for failure analysis. With

$$\Delta R = R1 - R2 \quad (29.)$$

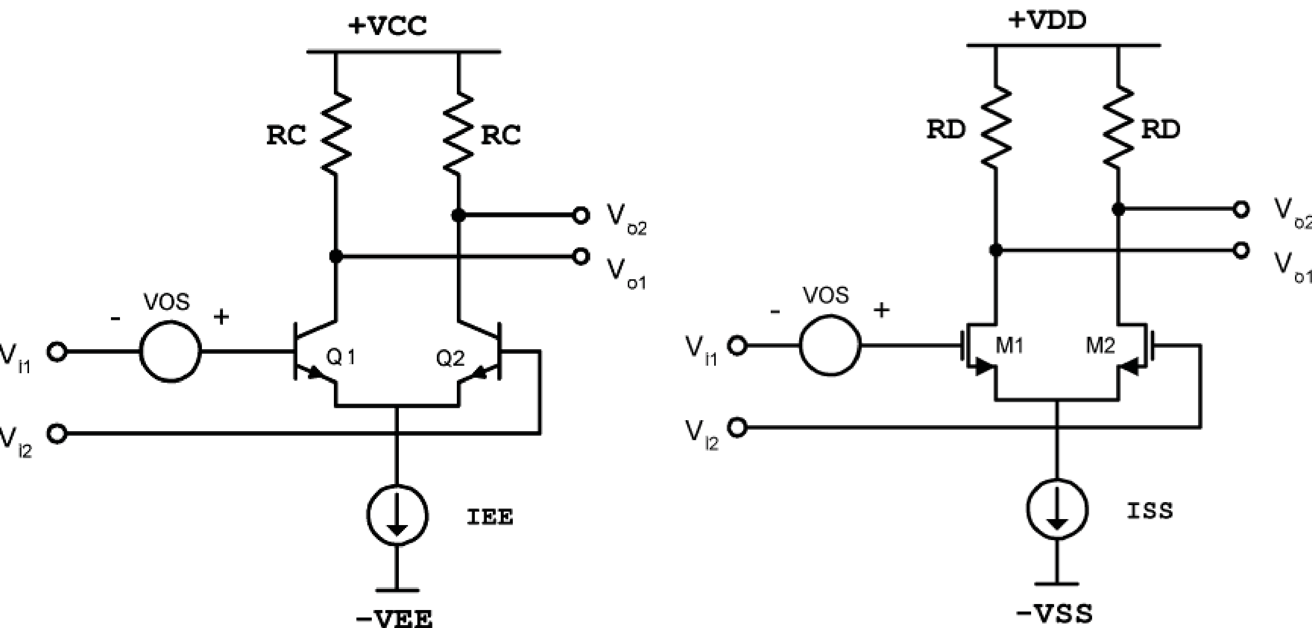


Figure 24. Schematics of a simple bipolar and MOS input differential amplifier.

$$R = \frac{R1 + R2}{2} \quad (30.)$$

$$\Delta I_S = I_{S1} + I_{S2} \quad (31.)$$

$$I_S = \frac{I_{S1} + I_{S2}}{2} \quad (32.)$$

the offset voltage is

$$VOS \approx kT\left(-\frac{\Delta R}{R} - \frac{\Delta I_S}{I_S}\right) \text{ [4].} \quad (33.)$$

A similar analysis of the MOS differential amplifier results in an offset voltage

$$VOS \approx \Delta V_t - \frac{(V_{GS} - V_t)}{2}\left(\frac{\Delta R_D}{R_D} + \frac{\Delta\left(\frac{W}{L}\right)}{\left(\frac{W}{L}\right)}\right) \text{ [4].} \quad (34.)$$

The input offset voltage is then proportional to the relative mismatch in the input transistors and the load resistors. And though this result was derived for this particular differential pair, in general the input offset voltage can be attributed to mismatches in the components that make up the input differential pair.

CONCLUSION

This paper covered the basics of analog device and circuit characterization. Though the breadth of the subject matter did not allow for an in-depth treatment of every subject, it is hoped that it is a good introduction for the failure analyst who is new to the subject.

REFERENCES

1. J. Beall and D. Wilson, "Curve Tracer Applications and Hints for Failure Analysis" in *Microelectronics Failure Analysis Desk Reference*, 3rd ed., ASM International, 1993.
2. D. Appleman and F. Wong, "Computerized Analysis and Comparison of IC Curve Trace Data and Other Device Characteristics" in *ISTFA 1990*, pp. 271-277.
3. D. Schroder, *Semiconductor Material and Device Characterization*, John Wiley & Sons, 1990, pp. 147 – 193
4. P. Gray and R. Meyer, Analysis *and Design of Analog Integrated Circuits,* 3rd ed., John Wiley & Sons, 1993
5. B. A. McDonald, "Avalanche Degradation of h_{FE}", *IEEE Transactions on Electron Devices,* VOL. ED-17, No. 10, October 1970, pp. 871-878.
6. P. Brokaw, "A Simple Three-Terminal IC Bandgap Reference", *IEEE Journal of Solid-State Circuits,* VOL. SC-9, No. 6, December 1974, pp. 388-393.
7. R. Widlar, "New Developments in IC Voltage Regulators", *IEEE Journal of Solid-State Circuits,* VOL. SC-6, No. 1, February 1971, pp. 2-7.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Screening for Counterfeit Electronic Parts

Bhanu Sood and Diganta Das
Center for Advanced Life Cycle Engineering (CALCE)
Department of Mechanical Engineering
University of Maryland, College Park, MD 20742
Phone: (301) 405 3498, email: bpsood@calce.umd.edu

Abstract

Counterfeit electronic parts have become a significant cause of worry in the electronics part supply chain. Most of the counterfeit parts detected in the electronics industry are either new or surplus parts or salvaged scrap parts. The packaging of these parts is altered to modify their identity or to disguise the effects of salvaging. The modification can be as simple as the removal of old marking and then adding new marking, or as complicated as recovery of a die and repackaging.

In this chapter, we discuss the type of parts used to create counterfeits and the defects/degradations inherent in these parts due to the nature of the sources they come from, proposed inspection standards, and limitations of these standards. The processes used to modify the packaging of these parts to create counterfeits are then discussed along with the traces left behind from each of the processes. We then present a systematic methodology for detecting signs of possible part modifications to determine the risk of a part or part lot being counterfeit.

Keywords: counterfeit electronics, inspection techniques

1 Introduction

A counterfeit electronic part is one whose identity (e.g., manufacturer, date code, lot code) has been deliberately misrepresented. Several factors contribute to the targeting of the electronic parts market by counterfeiters, including obsolescence; lead time (manufacturer or an authorized distributor unable to supply parts within the lead time requirement of customer); price issues (parts available at lower prices from independent distributors and brokers); absence of pedigree verification tools in the electronics part supply chain; availability of cheap tools and parts to create counterfeits; and costly inspection/testing procedures.

Easy availability of unauthorized parts is one of the prominent reasons for the growing problem of counterfeit electronic parts. There are relatively few incidents of illegal manufacturing in the electronics industry due to the high costs involved in manufacturing electronic parts such as integrated circuits. Counterfeit parts are generally relabeled part (e.g., marked as higher grade or with a recent date code or as RoHS[1] compliant), refurbished parts (i.e., used part reworked to appear as new), or repackaged part (e.g., recovery of die and repackaging). Counterfeiters have access to reclaimed, scrapped, and excess parts, which are easily available from unauthorized sources.

Excess inventories comprise of electronic parts that are no longer required by product manufacturers or contract manufacturers for normal production needs [1]. Excess inventories result due to a variety of reasons, such as differences between forecasts and actual production schedules, delay in discontinuations of slow moving product lines, and economic recessions [2][3]. Disposal options for excess inventories include alternate use within the company; returning the parts to original suppliers (manufacturers, distributors); disposing of the parts into the gray markets (unauthorized markets); and scrapping the parts. Out of the four disposal options, selling the parts in the gray market creates a source of parts for counterfeiters. Improper scrapping procedures used to scrap the excess parts (in the

[1] The RoHS Directive stands for the restriction of the use of certain hazardous substances in electrical and electronic equipment.

absence of other disposal options) can also result in counterfeiters salvaging the parts [4].

The pedigree of excess parts is often unknown due to the anonymous nature of transactions taking place in gray markets. The quality of excess parts depends on prior storage conditions, the duration of storage, and handling procedures. Depending on the construction, handling, and storage, excess parts can become unsolderable, contaminated, damaged, or otherwise degraded.

Part manufacturers and testing companies often scrap parts, which fail quality checks and other screening tests (e.g., functional tests, burn-in). Often companies do not destroy the parts in-house but rely on third parties. However, some parts escape destruction and are salvaged by counterfeiters. Examples of attributes of scrapped parts include manufacturing defects such as absence of die, lifted wire bonds, missing or no bond wires, and damaged terminations (e.g., broken leads, balls, or chip-out in the terminations of passive parts).

Reclaimed parts are parts that have been recovered from assembled printed circuit boards of discarded electronic assemblies and failed boards which are scrapped by contract manufacturers. The pedigree of these discarded assemblies is often unknown. Parts that are reclaimed from such products may have undetected defects or degradations. Reclaimed parts may also have defects induced during the reclamation procedures, such as damaged terminations, popcorn damage in the molding compound, and delamination of the molding compound from the die attach [5].

Apart from excess inventories, scrapped parts, and reclaimed parts, counterfeiters may also buy new parts and relabel or repackage them to make them appear to be a different part. Such parts may have handling or packaging related damages such as ESD[2] damage or poor workmanship issues.

Unlike material characterization (e.g., XRF[3]) and destructive tests (e.g., decapsulation) that require expensive tools and equipment, visual inspection can be carried out with a light optical microscope, Visual inspection can be a first step in the detection process, but should not be the only method. The visual inspection process also requires access to data sheets and support from manufacturers to obtain the actual attributes of parts, e.g., date code validity.

An electronic part that has been re-marked with good quality ink and without errors is hard to detect through the visual inspection method. Marking permanency tests will not work in the case of laser-marked parts. Even in case of ink-marked parts, marking permanency tests may erase the marking of an authentic part, thus giving the impression of the part being counterfeit. With the growing sophistication of technology, counterfeiters use better quality inks and laser equipment to create counterfeit parts. A salvaged scrap part, which has been scrapped because of internal quality problems, such as missing bond wires, may not be detected through the visual inspection method or marking permanency tests. A part that has been repackaged (from the die) may have discrepancies (e.g., different manufacturers) in the die and package markings. Such discrepancies can only be detected through destructive techniques such as delidding. Refurbishing techniques such as reballing and solder dipping may initiate failure mechanisms such as interfacial delamination or bond pad corrosion, which can only be detected through scanning acoustic microscopy. Visual inspection also cannot detect discrepancies in termination plating materials. Such discrepancies can only be detected through material characterization techniques such as XRF spectroscopy.

Table 1: Types of Parts Used to Create Counterfeits

Types of parts	Sources and attributes
Excess inventories	**Sources:** OEMs[4], Contract manufacturers **Attributes:** handling, packaging, and storage related damage; defects due to aging; no traceability; unknown pedigree
Scrapped parts	**Source:** part manufacturers, testing companies, contract manufacturers **Attributes:** internal quality problems such as missing die or bond wires; die contamination; part termination damage
Reclaimed parts	**Source:** recyclers **Attributes:** damaged terminations and body; inherent defects induced during reclamation; unknown pedigree

The Independent Distributors of Electronics Association (IDEA) has developed a document for acceptability of electronic parts that are distributed in the open market [6]. The document, IDEA-STD-1010A, provides visual inspection techniques (including marking permanency tests) and acceptance criteria for open market parts. Electrical and destructive or invasive inspection techniques (e.g., delidding) are out of the scope of this document and it only covers visual inspection of the markings, surface texture, mold pin, external packaging (tray or tube), and body of a part. This document or any other

[2] Electrostatic discharge.

[3] X-ray fluorescence spectroscopy.

[4] Original equipment manufacturers.

methods that use only external visual inspection are not sufficient for detecting counterfeit parts.

Some test laboratories depend on electrical tests for detecting sub-standard and counterfeit parts. Electrical tests include parametric testing, board-level testing, and hardware and software functionality testing. In most cases, electrical tests are used to detect non-functional or failed parts. Some counterfeit parts may function properly during the electrical tests, but they may have inherent defects (e.g., contamination) induced during refurbishing or re-marking. Inherent defects induced during the counterfeiting process can only be detected through a systematic packaging evaluation. In this paper, we present a counterfeit detection process that incorporates packaging evaluation using tools and methods to detect signs of possible part modifications.

Table 2: Limitations of Using Visual Inspection Alone for Detecting Counterfeits

Types of counterfeit parts	Examples of limitations of visual inspection
Repackaged	○ Cannot detect internal discrepancies such as bond wire misalignment or missing bond wires, missing or damaged die ○ Cannot detect die and package marking mismatches
Remarked	○ Fails if markings on counterfeit parts are good quality ○ Need access to datasheets or support from original manufacturer
Refurbished	○ Cannot verify RoHS compliance claims ○ Cannot detect termination plating discrepancies with original parts ○ Cannot detect internal failure mechanisms induced during the refurbishing processes such as interfacial delamination
Salvaged scrap parts	○ Markings may be original manufacturer's, thus difficult to detect any discrepancies ○ Internal problems such as missing die or bond wires cannot be detected

2 Creation of Counterfeit Parts

With easy availability of parts to create counterfeits, counterfeiters have developed inexpensive methods of counterfeiting that rely on modifying the packaging of the parts by processes such as relabeling or refurbishing. In this section, we discuss the three most commonly used methods used by counterfeiters to create counterfeits.

2.1 Relabeling

Relabeling is the process of altering the markings on a part to make it appear as a different part. A typical part marking includes part number, lot number, and the manufacturer's logo. In some cases, part marking also includes the country of origin mark. The relabeling process includes erasing the original marking by methods such as black topping, or sand blasting and applying a new marking to create a counterfeit part. Sandblasting is the process of smoothing, shaping, or cleaning a hard surface by forcing solid particles across that surface at high speeds. Blacktopping is a process in which a layer of material is applied to the top surface of a part to cover over old marking. Blacktopping may also be carried out after the part has been subjected to sandblasting.

Relabeling may be carried out according to the needs of the customer to have higher grade parts (e.g., changing processor speed), different parts with the same pin count and packaging type, different vintage parts (e.g., changing date code), or different military specifications. Some cases of relabeling also include dual part marking, i.e., the presence of part marking at two different places on the part.

GIDEP[5] issued an alert about operational amplifiers, LT1057AMJ8/883 with date code 0122 in 2006. Linear Technology Corporation (LTC) received the parts from a customer when the parts failed functional tests at the customer's facility. Destructive and physical analysis (DPA) of the parts revealed the die to be an original LTC die manufactured in October 1995 as a military lot. The parts were found to have been relabeled to make them appear to be new parts [7].

Relabeling leaves behind traces that can be detected through visual inspection or marking permanency tests. Some of the traces left behind are part marking irregularities such as spelling mistakes, different marking techniques used (e.g., laser marking instead of ink marking); dual part markings; part markings with invalid date codes or part numbers; parts (ink-marked) failing marking permanency tests; a filled-in or unclean pin-1cavity; die markings (date code, manufacturer) not matching with the package marking; and absence of country of origin marking.

2.2 Refurbishing

Refurbishing is a process in which parts are renovated in an effort to restore them to a like new

[5] GIDEP: Government industry data exchange program.

condition in appearance. The terminations of refurbished parts are realigned and re-finished (in the case of leads) or undergo reballing (in the case of ball grid array (BGA) type interconnects) to provide a new finish. Refurbishing is often carried out in conjunction with relabeling to sell used parts as new parts. Refurbishing is also carried out to hide defects that arise during reclamation of parts from circuit boards and improper handling. Refurbishing induces defects/degradations in parts such as bridged balls, missing balls, broken leads, popcorning, warpage, or localized delamination.

Realignment of leads (such as straightening) is often carried out on reclaimed or scrapped parts that have bent or non-aligned leads caused during reclamation of the parts from printed circuit boards or poor handling. Realignment of leads may cause damage to terminations such as broken leads or improperly aligned leads. The realignment process may also cause internal defects such as interfacial delamination and cracked passivation layer.

Solder dipping is frequently used to change the lead finish, e.g., from lead free (Pb-free) finish to a lead finish or vice-versa. Solder dipping is also used to improve or restore the solderability of the parts. But poor finish and thermal shock experienced during the solder dipping process can lead to defects in the terminations such as bridging across leads, internal delamination leading to package cracking, a cracked passivation layer, and deformation in die metallization [8].

Reballing is a process carried out on BGA parts to replace damaged balls or to change the termination finish from Pb-free to lead or vice-versa. Counterfeiters often use the reballing process to refurbish the part terminations (BGA) of reclaimed or used parts (with damaged balls) to make them appear to be new parts. Inconsistencies during reballing can cause defects such as missing solder balls, damaged pads, and bridged balls. Other defects caused by improper reballing are warpage, popcorning, and local delamination.

2.3 Repackaging

Repackaging is the process of altering the packaging of a part in order to disguise it as a different part with a different pin count and package type (e.g., dual-in-line (DIP) or plastic leaded chip carrier (PLCC)). The process involves recovery of die (by removing the original packaging) and molding the die into the desired package type. Counterfeiters generally do not use proper handling procedures, tools, and materials for repackaging the die, which may lead to defects or degradation in the repackaged parts such as die contamination, moisture-induced interfacial delamination, and cracks in the passivation layer. There may also be workmanship issues with the repackaged parts such as missing bond wires, missing die, bond wire misalignment, or poor die paddle construction. The marking on repackaged parts also may not match with the die markings. There may also be marking irregularities such as spelling errors, discrepancies in part number, or an incorrect logo. Counterfeiters may also use inferior quality materials to package the die, such as cheap filler materials and flame retardants.

Table 3: Processes Used to Create Counterfeits and Associated Defects

Process of counterfeiting	Associated defects
Relabeling	Marking irregularities, poor quality marking, filled-in or unclean mold cavities, discrepancies in package marking with the die marking, ESD damage
Repackaging	Discrepancies in package marking with the die marking; workmanship issues such as missing bond wires or poor die paddle construction; internal defects such as moisture induced interfacial delamination; poor materials used
Refurbishing	Bridged or improperly aligned terminations; internal defects such as interfacial delamination and cracked passivation layer induced during processes such as solder dipping, reballing, and realignment of terminations; differences in termination plating material with original part

3 Detection of Counterfeit Parts

Most of the counterfeit parts detected in the electronics industry are either new or surplus parts or salvaged scrap parts that are modified. The modification can be as simple as removal of marking and re-marking or as sophisticated as recovery of the die and repackaging. Most of these modifications leave behind clues that can be uncovered in order to establish the authenticity of the part. In this section we present a sequence of detection techniques that can be applied for detecting signs of possible part modifications. Detection is an important step to determine the risk of a part or part lot being counterfeit. The evaluation methodology begins with steps that can be implemented at the receiving department. The steps can include a thorough evaluation of shipping packages, inspection of humidity indicator cards, ESD bags, tube and tray materials and shipping labels. Inspection procedures of higher sophistication levels can be then be applied. These steps include external visual inspection, marking permanency tests for external compliance

and X-ray inspection for internal compliance. These inspection processes are followed by material evaluation in destructive and non-destructive manners such as XRF and material characterization of the mold compound using thermo-mechanical techniques. These processes are typically followed by evaluation of the packages to identify defects, degradations and failure mechanisms that are caused by the processes (e.g., cleaning, solder dipping of leads, reballing) used in creating counterfeit parts. This method of assessment is necessary since the electrical functionality and parametric requirements may be initially met by the counterfeit parts, but authenticity can only be evaluated after complete evaluation of the package. The latent damages caused by the counterfeiting process can only be detected by a thorough packaging evaluation.

Table 4: Inspection Methods, Severity and Tools or Equipment Needed

Inspection method	Severity and tools or equipment required
Incoming Inspection	**Severity:** non-destructive, may induce handling related damage such as ESD if precautions are not taken **Tools/Equipment**: Low power stereo macroscope, bare eyes, ruler, weighing balance. Information on original part material may be needed.
External visual inspection	**Severity:** non-destructive, may induce handling related damage such as ESD if precautions are not taken **Tools/Equipment:** low power optical macroscope, optical microscope, solvent for marking permanency tests, part datasheet information
X-ray inspection	**Severity:** non-destructive, may induce handling related damage such as ESD if precautions are not taken. Instances of part damage due to X-ray radiation exposure are also reported. **Tools/Equipment**: X-ray machine, X-ray images of an authentic part
Material evaluation and characterization	**Severity:** may be destructive or non-destructive depending on the type of equipment used **Tools/Equipment:** XRF, environmental scanning electron microscope (E-SEM), energy dispersive spectroscopy (EDS), differential scanning calorimetry (DSC), thermo-mechanical analyzer (TMA), thermomechanical analyzer (TMA), dynamic mechanical analyzer (DMA), hardness testers, Fourier transform infrared spectroscope (FTIR). Information on original part material may be needed
Packaging evaluation	**Severity:** non-destructive **Tools/Equipment**: scanning acoustic microscope (SAM), ion chromatography.
Die inspection	**Severity:** destructive **Tools/Equipment:** automatic chemical decapsulator, can also be carried out through manual etching; information on original die markings and attributes needed. wire pull, ball bond and solder ball shear testing, environmental testing and micro-sectioning.

3.1 Incoming Inspection

Incoming inspection is the process of verifying the conditions of materials used for shipping the suspect packages. Attributes to inspect for include the status of humidity indicator cards (HIC), moisture barrier bags or ESD bags. Not only should the as-received state of the above materials be checked, but their authenticity should also be verified. Instances of counterfeit or fake HIC cards are on the increase..

Incoming inspection should start with verification of the receiving documents and external labels on shipping boxes and matching the details in the purchase order with the shipping list enclosed with the shipment. Manufacturers' logs and shipping origin should also be checked and verified. Any certificate of conformance (CoC) should also be inspected for authenticity and cross-checked with existing CoCs from same distributor or part manufacturer. The next step is an inspection of the ESD and moisture barrier bags to check for any damage or sealing issues. The HIC should also be checked to verify that it is genuine and based on the color indicator, that the shipment has not been exposed to elevated levels of humidity that may prove detrimental to the functioning and reliability of the electronic part. Brand of tray, tube and reels used in the shipment should also be inspected. Single shipments of counterfeit parts have been known to be shipped in trays of different brands.

3.2 External Visual Inspection

External visual inspection is a process of verifying the attributes of parts such as package and part markings (part number, date code, country of origin marking), part termination quality, and surface quality. Visual inspection is performed on a sample of parts from a given lot. Resources required for carrying out visual inspection are standard tools for handling electrostatic sensitive parts [9], part datasheet information (part number format, dimensions, number of pins and package type), a microscope with at least 30X magnification (magnification of the microscope can be adjusted to inspect certain features of the part), a camera built

into the microscope (some of the processes of determining a counterfeit require sending copies of photos to different resources for their evaluation), and a solvent to check for part marking permanence.

The visual inspection starts with the inspection of the label on the packaging in which the parts are shipped. Features to inspect include spelling errors on the manufacturer labels, validity of manufacturer codes on the labels (such as codes that contain information on manufacturing location), verification whether the date codes on the external packaging match date codes on the parts, and validity of date codes. The packaging inspection also includes any part specific requirements such as the requirement of a dry pack and a humidity indicator card for moisture-sensitive parts.

The next step in the inspection process is the verification of whether the part markings, such as logo, part number, lot code, date code, and Pb-free marking (if any), conform to the shipping and purchase order information. This is followed by verification of the validity of the part number, date/lot codes, and Pb-free marking (if any) with the original part manufacturer requirements. In some cases counterfeiters may not place Pb-free marking on the parts (when they relabel the parts with newer date codes), though the original manufacturer may have shifted to Pb-free manufacturing. The part should also be inspected for any dual part marking, such as marking on the top as well as on the side of a part with different and often conflicting information. The markings should also be inspected for any irregularities such as spelling mistakes, font size differences compared with the original part, and the marking technique used on the part. For example, an authentic part may have ink marking, whereas the counterfeit part may have laser marking. Figure 1 provides examples of items to look for during visual inspection of a part.

Marking with inferior quality inks or laser equipment can be detected by conducting marking permanency tests on the parts or looking for any laser-induced defects on the parts, such as holes on the surface. Acetone is a common solvent used to determine if a part has been remarked, but a less harsh solvent is a combination of 3 parts mineral spirits and one part alcohol. This is the mixture that MIL-STD-883 (Method 2015.13) [10] requires part markings to withstand. Certain harsher solvents such as DynaSolv 711 are also frequently used for checking for marking permanency. If the result of the marking permanency test is a change in the surface texture or wiped-off marking, this is a possible sign of the part's being counterfeit.

The pin-1 cavity and other mold cavities (part of the plastic mold process) present on a part should be inspected for uncleanness or unevenness, because sandblasting or blacktopping leaves the mold cavities unclean or filled in. Verification of the pin-1 or other mold cavities on a part is a critical way to determine signs of relabeling on a part. In some cases, counterfeiters also etch a new pin-1 cavity in place of the filled-in cavity. Also, the presence of marking over the pin-1 cavity is a sign of the part's being counterfeit.

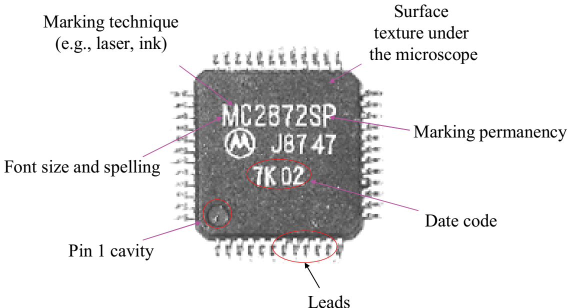


Figure 1: Examples of items to look for during visual inspection

The surface texture of a relabeled part is different from an authentic part. The surface of an authentic part when looked at with a microscope is usually sharp and rough (due to molding process residues and filler particles), whereas surface of a relabeled part is smooth because of relabeling methods such as sandblasting or blacktopping. Sandblasting also leaves marks that have a directional pattern on the surface of a part. Sandblasting also leads to rounded corners and edges.

Visual inspection also includes inspection of the part termination (leads or balls) quality to detect possible signs of counterfeiting. Part terminations should be inspected for any signs of refurbishing (solder dipping or reballing) or damage (broken or bent leads, bridged balls) due to reclamation. If the termination type is leads, things to look for are straightness, coplanarity, scratches, or other defects caused by reclamation or prior use. Termination refurbishing techniques such as solder dipping and reballing leaves behind traces that can be detected through visual inspection, such as bridged terminations and missing solder balls.

3.3 X-ray Inspection

X-ray inspection is carried out to conduct internal inspection on parts to verify the attributes of parts such as die size and bond wire alignment. X-ray inspection is also used to detect anomalies such as missing bond wires, missing die, or the presence of

contamination (Figure 2). Counterfeit parts are sometimes packaged without a die or with a different die. A die from a different manufacturer than the one listed on the package does not necessarily indicate a counterfeit since manufacturers sometimes institute a process change on a particular product (but production protocol requires a change in lot/date code). X-ray imaging is not the tool to resolve manufacturer logos and markings on the die surfaces to authenticate the device.

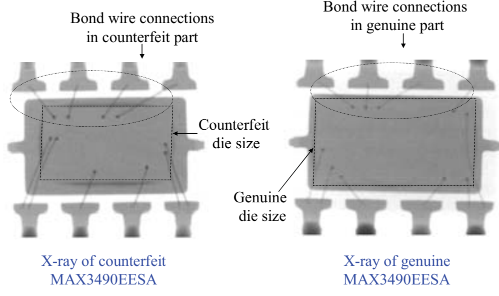


Figure 2: Example of X-ray inspection[6]

3.4 Material Characterization

Counterfeit parts often have discrepancies in termination material or molding compound material when compared with an authentic part. A part that has been relabeled with a newer date code may have tin-lead (SnPb) solder as the termination material, whereas the authentic newer version of the part has no lead in the termination. Similarly, the same counterfeit part may contain a halogenated flame retardant in the mold compound, whereas the authentic newer version of the part may be halogen-free to comply with RoHS directives. Similarly, a counterfeit part may also claim to comply with RoHS directives but may actually have Pb or halogens in the termination finish or mold compound.

X-ray fluorescence spectroscopy (XRF) can be carried out on the parts to evaluate the material composition of the terminations and the molding compound in order to detect the presence or absence of Pb and any other discrepancies with an authentic part. XRF can also be a useful tool to detect counterfeit passives. CALCE conducted authentication services on customer-returned multi layer ceramic (MLCC) capacitors using X-ray fluorescence spectroscopy. The capacitors were found to be similar to an authentic part except for low concentration of a critical rare-earth element, Yttrium. Figure 3 shows a plot of the variation in the amount of Yttrium among the various parts that were analyzed with XRF.

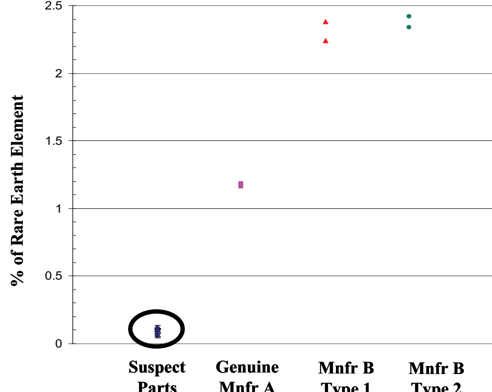


Figure 3: Plot showing the variation in the amount of Yttrium among the various parts

Another method of evaluating the material composition is through environmental scanning electron microscopy (E-SEM) and electron dispersive spectroscopy (EDS). E-SEM is conducted on parts after removing the encapsulants (decapsulation) or after delidding. For example, E-SEM microscopy can be used to verify the elemental composition of the metallization layers. E-SEM microscopy can also be used to verify the solder plating composition on the part termination. In certain cases E-SEM can also be used for inspecting the external part packaging for signs of sandblasting and for detecting topographical changes resulting from the black-topping process.

If a non-authentic raw material is used in a part, polymeric materials such as the component molding compounds, attach materials and coatings need to be evaluated in comparison with the authentic parts in order to detect counterfeit parts. Tools and equipment that aid in material characterization includes differential scanning calorimeter (DSC), thermo-mechanical analyzer (TMA), dynamic mechanical analyzer (DMA), hardness testers, and Fourier transform infrared spectroscope (FTIR). In the dynamic (temperature scanning) approach, a DSC can be used to study the cure reaction and glass transition temperature of the epoxy molding compounds which can be compared with the cure reaction of epoxy molding compound from a known authentic part. The TMA can be used to measure the coefficient of thermal expansion (CTE) of molding compounds of suspect parts which can then be compared with CTE of an authentic part. A DMA can be used to determine the visco-elastic material properties of an epoxy molding compound which can be compared to similar properties of expected molding compound. FTIR spectroscopy, by means of

an infrared spectrum of absorption and emission characteristics of the different organic functional groups within molding compound, can help in distinguishing between counterfeit and authentic parts.

It should be clarified that typical processing steps such as solder reflow, rework and burn-in testing can introduce changes to the thermo-mechanical and cure properties of epoxy molding compounds due to the significant high temperature exposure. While using tools such as DSC, TMA, DMA and FTIR, results can vary among genuine parts if they are sourced from assemblies that have been exposed to any of these processing steps and some variations in material properties is expected.

3.5 Packaging Evaluation to Identify Hidden Defects/Degradations

Processes used to create counterfeits such as relabeling, refurbishing, and repackaging often induce internal defects/degradations in parts due to a lack of proper equipment/tools used and improper handling procedures. In this section we provide techniques and procedures for packaging evaluation to identify hidden defects/degradations.

Delamination, voids, and cracks in plastic-encapsulated microcircuits lead to failure mechanisms such as stress-induced passivation damage over the die surface, wire bond degradation due to shear displacement, accelerated metal corrosion, reduction in die attach adhesion, intermittent outputs at high temperature, popcorn cracking, die cracking, and device latch up (hot spot formation). Defects such as delamination, voids, and cracks can be caused due to thermal and mechanical shocks during reballing, solder dipping, realignment of leads, and repackaging.

Moisture-induced interface delamination can occur during each of the processes of relabeling, refurbishing, and repackaging. Moisture-induced interface delamination begins with the package absorbing moisture from the environment, which condenses in micropores in polymer materials such as the substrate, die-attach, molding compound, and various adhesives along the interfaces. During the PCB assembly process, when the part is exposed to high temperatures associated with the soldering process, popcorning may occur.

Scanning acoustic microscopy (SAM) is a non-destructive method which can be used to detect delamination of the molding compound from the lead frame, die, or paddle (top side and bottom side separately); voids and cracks in molding compound; and unbonded regions and voids in the die-attach material. SAM can detect hidden defects such as delamination growing along the die, isolated voids (bubbles or from outgassing), lack of die attach material between die and substrate, and delamination growing along substrate. Procedures for acoustic microscopy for non-hermetic encapsulated electronic parts are provided in JEDEC Standard J-STD-035 [11] and NASA Standard PEM-INST-001 [12]. Examination of the package for voids, cracks, and delamination should be performed at multiple locations including the interface between the die surface and molding compound (top view), and interface between the lead frame and molding compound (top and back view).

3.6 Die Inspection

For die inspection, a preparatory method to expose the die is necessary. Once the die is exposed, the attributes of the die, such as die markings (e.g., manufacturer logo, date code), passivation layer quality, and interconnection quality, can be verified using a high power microscope (Figure 4). A part that is counterfeited using relabeling and repackaging will usually have discrepancies in the die and package marking.

Defects induced during the refurbishing process due to thermal and mechanical shock such as metallization layer damage (due to ESD, corrosion), contamination, bond wire defects, and cracks in the passivation layer can be detected by inspecting the die features. Repackaging-induced defects, such as chip-to-substrate attachment failure leading to voids and thermal stress problems, deformation of bond wires due to improper bonding, and cracks at the bond pad–bond wire junction, can also be detected by inspection of the die area.

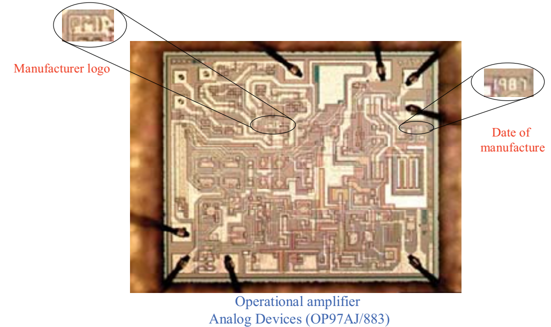


Figure 4: Example of die inspection[7]

4 Summary and Recommendations

Often there is damage inherent in parts that are used to create counterfeits. The damage may result from improper handling, storage, or packing procedures, as in the case of new or excess inventories or overruns. Damage may also occur when parts are reclaimed from assembled printed circuit boards. Parts may also have failed even before they were counterfeited, as in the case of parts that are scrapped by the part manufacturer during quality control (QC) checks. Parts may be counterfeited using processes such as relabeling, refurbishing, and repackaging, each of which leaves behind traces in some form or other.

A systematic methodology for detecting counterfeit parts has been presented in the paper. The methodology consists of external visual inspection, marking permanency tests, X-ray inspection, and material evaluation and characterization, followed by identification of defects or degradations that may have been induced during the counterfeiting process, and die-marking inspection. This methodology helps in detecting signs of possible part modifications to determine the risk of a part or part lot being counterfeit. Table 5 summarizes the methodology and tools.

Table 5: Inspection Methods and Traces or Defects to Inspect

Inspection method	Items of Review
External visual inspection	Spelling errors in part markings or labels; validity of logo, part number, lot code, date code, and/or Pb-free marking; marking technique; quality of marking; mold cavities; straightness, coplanarity, scratches, bridging or other defects in terminations; surface texture
X-ray inspection	Die size; bond wire alignment; anomalies such as missing bond wires, missing die, or presence of contamination
Material evaluation and characterization	Termination plating materials, molding compound, attach materials, coatings, laminate or substrate materials
Packaging evaluation	Delamination of the molding compound from the lead frame, die, or paddle; voids and cracks in molding compound; and unbonded regions and voids in the die-attach material
Die inspection	Die markings (e.g., manufacturer logo, date code), passivation layer quality, interconnection quality, metallization layer damage (due to ESD, corrosion), contamination, bond wire defects

We expect that organizations will evaluate the sources of parts prior to purchasing parts from them and thereby eliminating the biggest risk factor of obtaining counterfeit parts. To be effective, the inspection process needs to come to a conclusion within a relatively short period of time and hence a logistics plan of performing the evaluations needs to be in place since all the equipment and expertise may not reside in the same location.

The inspection methodology presented in this paper is a tool of last resort and it is no substitute for sound supply chain management methods. The cost of inspection can add up to be significant in relation to the cost of parts. The possibility of damage of parts from additional handling associated with inspection remains even when the parts are determined not to be counterfeit. All things considered, a strict and effective inspection method will help in finding suspect counterfeit parts but it will not necessarily save time and money for an organization.

5 References

[1] Parker, J. and Turgoose, C., “An Analysis of the Excess Electronic Components Market”, HP Laboratories, Bristol, February, 1999.

[2] Chatterjee, K. and Das, D., “Semiconductor Manufacturers' Efforts to Improve Trust in the Electronic Part Supply Chain”, Components and Packaging Technologies, IEEE Transactions on , vol.30, no.3, pp.547-549, Sept. 2007.

[3] Chatterjee K., Das D., Pecht M., Suorsa P., and Ricci C., “Solving the counterfeit electronics problem,” Proc. Pan Pacific Microelectron. Symp. (SMTA), Jan. 30–Feb. 1 2007, pp. 294–300.

[4] Pecht M. and Tiku S., “Bogus: Electronic Manufacturing and Consumers Confront a Rising Tide of Counterfeit Electronics,” IEEE Spectrum, vol. 43, no. 5, pp. 37–46, May 2006.

[5] Snider, K., “A Time for Change: The Not So Hidden Truth Behind the Chinese Open Market”, ERAI Special Report, April 2007.

[6] Independent Distributors of Electronics Association (IDEA), IDEA-STD-1010-A, “Acceptability of Electronic Components Distributed in the Open Market”, 2006.

[7] GIDEP (Government-Industry Data Exchange Program) Alert, Document no. CT5-A-07-01 dated November 27, 2006.

[8] Sengupta, S., Das, D., Ganesan, S., Pecht, M., Lin, T.Y., and Rollins, W., “Effects of Re-finishing of Terminations by Solder-dipping on Plastic Quad Flatpack Electronic Parts”, Proceedings of 7th IEEE Conference on

Electronic Packaging Technology, Vol. 2, Dec 2005.

[9] EIA/JEDEC, Publication JESD625-A, “Requirements for Handling Electrostatic-Discharge-Sensitive (ESDS) Devices”, Arlington, VA, December 1999.

[10] Military Standard MIL-STD-883, Rev, H, “Test Method Standard for Microcircuits”, Columbus, OH, February 2010.

[11] EIA/JEDEC, Publication J-STD-035, “Acoustic Microscopy for Nonhermetic Encapsulated Electronic Components”, Arlington, VA, May 1999.

[12] NASA, PEM-INST-001, “Instructions for Plastic Encapsulated Microcircuits (PEM) Selection, Screening, and Qualification”, Greenbelt, MD, June 2003.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor
Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

An Overview of Analog Design for Test and Diagnosis

Stephen Sunter
Mentor Graphics

Abstract – ***Test time and diagnosis effort for analog circuitry now consumes much more than its area proportion of most ICs, so analog design-for-test (DFT), especially built-in self-test (BIST) has become more important recently. This paper first describes the most commonly used general analog DFT techniques, which are loopback and analog buses, and then describes function-specific DFT techniques for the most common analog functions, which include PLL, SerDes, ADC, and RF. It briefly explains the allure of analog BIST and then presents seven essential principles for BIST of analog functions, including some aimed at diagnosing analog functions in more systematic and automated ways.***

I. Introduction

In the last 20 years, there have been tremendous advances in designing digital circuits to make them more testable. By "testable", we don't mean just the ease with which we can sort good (defect-free) from bad (defective) devices; we also mean how easily we can diagnose manufacturing defects and excessive variations, to correct deficiencies in the circuit design, masks, manufacturing process, or test.

The improvements in digital DFT led to engineering design automation (EDA) tools that automate the creation of scan paths in random logic circuits, connecting the flip-flops into long shift registers in test mode, so that the logic values of a circuit's internal nodes can be shifted off-chip to allow deduction of any defective logic gates. Inserting a scan path requires adding multiplexers in most logic paths – the added delay is usually insignificant (or zero) thanks to automated logic optimization. After a scan path is inserted, test patterns that achieve 100% coverage of all stuck-at faults (equivalent to a short between a signal node and a power rail) and stuck-open faults can be generated automatically. (Though, as analog effects, like capacitive coupling between signals, inductance in power rails, and mismatch in delay paths become more significant in ever-shrinking CMOS processes, the complexity of automatically generating these patterns is greatly increasingly.)

Today, analog test has become the bottleneck in getting systems on a chip (SoC) to market [7][8]. Typically, less than 20% of these chips is analog circuitry, but it incurs 70~80% of the test cost (test time and ATE capital expense) and up to half of the DFT and test engineering effort. Diagnosis can be a large part of this effort.

A key reason for this disparity between analog and digital testability is the lack of automatable, systematic, and general DFT methods for analog circuits. Now that analog has become the bottleneck in reducing test time and time-to-market, industry is beginning to focus on this area.

An "analog" circuit should be defined for this context. Obviously, a filter comprising resistors, capacitors, and an operational amplifier must be considered an analog circuit. Switched capacitor filters, and switched-mode voltage regulators are also considered purely analog, even though they have digital clocks and control signals. Radio frequency (RF) circuits are perhaps the only remaining pure analog circuits commonly implemented in ICs. All these pure analog circuits have analog inputs, outputs, and functions.

Another class of circuits has an analog core but a digital input *or* output. This class includes the analog-to-digital converter (ADC) and DAC. A third class has an analog core but digital inputs *and* digital outputs – this includes the phase-locked loop (PLL), delay-locked loop (DLL), serializer-deserializer (SerDes), and double data rate (DDR) interfaces. This class is easier to create DFT techniques for because the access circuitry can be entirely digital.

A last class of circuits that can be considered analog in at least some respects is digital circuits. When we consider delay, drive, leakage, switching threshold, and simultaneous switch noise (SSN), digital circuits can appear quite analog.

The development of systematic analog DFT methods has generally progressed from addressing analog aspects of digital gates, to analog functions with digital inputs and outputs, to ADCs and DACs, and lastly to pure analog.

The next section of this paper will describe the most common, general-purpose analog DFT methods, and show their advantages and limitations in diagnosability and automatability. Then DFT and BIST techniques for the most common analog functions will be described. Lastly, the challenges of creating BIST for analog functions are summarized, along with seven principles that must be exploited to meet these challenges.

II. Common Analog DFT Methods

Loopback is one of the oldest systematic DFT methods still widely used today. It was probably used first in modulator-demodulators (modems) that allow computers to communicate via telephone lines. In one loopback mode, the transmitted signal is looped back into the receiver, without being transmitted on the telephone wire. This allows the attached computer to detect whether the signal is, in fact, being transmitted. In another mode, loopback occurs at the far-end of the telephone wire, thus allowing the computer to diagnose whether the problem lies in the modem or in the telephone wire.

In an IC with an ADC and DAC, the DAC's analog output can be connected to the ADC's analog input so that a

test stimulus can be purely digital, as shown in Figure 1. This permits testing the IC on a purely digital tester. By selecting different internal paths for the loopback, the location of any defects can be narrowed to a single block. The main shortcoming of this approach for testing is that defects in one direction might be compensated by defects in the other direction. For example, if the DAC has excessive gain and the ADC has insufficient gain, the overall gain might be within test limits and incorrectly pass the test.

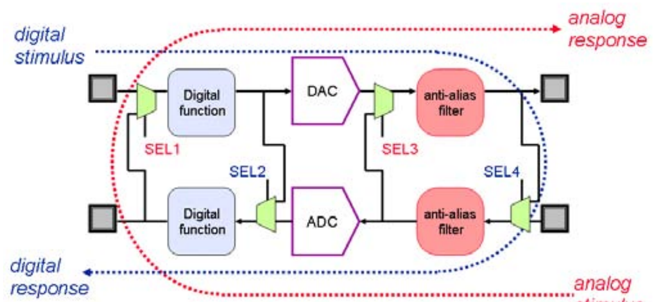


Fig. 1 Loopback paths for an IC with an ADC and DAC

The loopback technique is standard practice for testing high-speed serializer-deserializers (SerDes). A digital BIST circuit generates a pseudo-random bit sequence (PRBS) applied as parallel words to the serializer at 100~500 MHz clock rates, the serialized output is looped back to the deserializer, where the BIST latches the data as parallel words and compares them to expected words to measure the bit error rate (BER). Loopback avoids connecting the tester to the 2~20 Gbit/s transmitted serial data.

A key advantage of a PRBS is that it is not truly random, and is actually quite predictable. The receiver portion of the BIST can simply latch any digital word received (it will very likely be error-free) and then predict all the correct subsequent words in the sequence. This allows any amount of delay to be tolerated in the loopback path. Many companies simply test a SerDes by checking that a few hundred million bits are received error-free, without testing any AC parameters at all. Characteristics like jitter, frequency-dependent phase delay, slew rate, and rise/fall mismatch (which causes duty cycle distortion) are assumed to be acceptable if no bit errors are detected. The minimum acceptable BER for most applications is 10^{-12}, or one in a trillion bits, so testing a billion bits is testing a tiny fraction of the number needed to measure BER. But industry experience seems to show it is sufficient, at least for data rates below 5 Gb/s in nanometer CMOS technologies. Above that rate, the ability of SerDes to equalize the frequency-dependent delay through board wiring becomes paramount, and it is not tested by digital loopback BIST.

Control of test mode functions, like loopback or BIST, is widely implemented using the IEEE 1149.1 [1] standard (often called "JTAG", after the joint test action group that initiated its development). This standard specifies a 4-pin or 5-pin standard test access port (TAP) that is used widely for boundary scan but also to control test functions and in-system programming (ISP).

The analog test bus is another classic analog DFT technique that is widely used [8][9][10]. A single-wire monitor bus (or sometimes a differential pair of wires) is connected via CMOS transmission gates to various signal nodes of interest, as shown in Figure 2. A second signal wire (or differential pair) might also be used to simultaneously inject signals at nodes of interest, though this is less common than the monitor bus. To avoid connecting relatively large off-chip capacitances to the node-of-interest, an analog buffer is usually used, often one of the function's buffers connected to the bus in test mode. The transmission gates are controlled by digital signals, possibly derived from a shift register. The IEEE 1149.4 standard [2], often referred to as analog boundary scan, was developed based on this concept, primarily for accessing the analog pins of an IC for board-level testing. The standard has not been accepted widely for that purpose, though a few companies use its standardized control structure for access to internal signals.

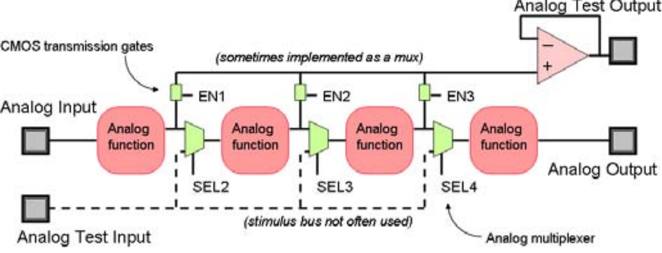


Fig. 2 Analog test bus

An analog test bus allows detailed diagnosis of circuit functionality, even if its bandwidth is much lower than most signals of a circuit. The bus can be used to observe (or adjust) DC reference voltages, low frequency signals, or the average value of high frequency signals. The main disadvantages of analog buses are their impact on sensitive signal nodes, crosstalk between circuits near or connected to the bus, excess leakage current when many nodes are connected to a single bus, distortion added to signals by mismatch in each transmission gate's transistors, and low bandwidth.

BIST is a DFT technique that is widely used for testing digital circuitry, especially embedded memories and random logic. Digital BIST is accomplished by shifting pseudo-random bit patterns into scan paths, applying the bits in parallel or combinational logic gates, capturing the resulting output bits in parallel, shifting out the bits, and then combining the bits to produce a digital signature. There is no comparable, widely-used analog BIST, though it has been an area of extensive research since about 1990 [11].

III. DFT FOR COMMON FUNCTIONS

This section briefly describes techniques used for DFT of the most common analog functions. Each technique is very different, unlike for digital circuitry where scan path access is used for almost all logic functions except memory, and BIST is used for almost all memories.

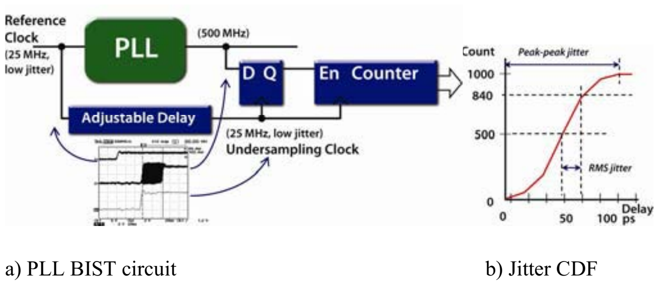
A PLL

The PLL is probably the most common “analog” function on ICs. PLLs are mostly used to generate high frequency clocks for digital core logic from a low frequency reference, to save power compared to using a high-frequency reference clock at the board level, since $P = fCV^2$, and on-chip capacitances and voltage swings are smaller than off-chip. The two most important parameters of a PLL are lock time and jitter. Aside from having a major impact on the whole IC’s performance, lock time is a good indicator of the loop gain in the PLL, and jitter is a good indicator of the PLL’s phase detector and thermal noise in its voltage-controlled oscillator. Although some companies monitor circuit nodes inside a PLL via an analog bus, this is impractical for many PLLs because these nodes are very sensitive to extra capacitance and noise.

The most common way to test PLL performance is to detect lock time by simply waiting an equivalent time after the PLL is initialized and then start testing the logic core that is clocked by the PLL – if the first few vectors fail then either there is a logic defect or the PLL has not locked yet, so in either case the IC is defective. Of course, this provides little diagnostic resolution. A more direct way to measure lock time is to count clock cycles from initialization (or preferably a forced loss-of-lock) until the PLL’s lock indicator changes state. This is often done on an *ad hoc* basis.

Measuring PLL output jitter is more complex. Some companies compare the output phase of the PLL to that of an adjustable delay line comprising a series of inverters tapped at various points by a multiplexer, as shown in Figure 3a. The delay is incremented after every 1000 samples, for example, until the delay has swept the entire range of the jitter. A plot of the resulting number of logic 1 samples *vs* delay is the cumulative distribution function (CDF) of the jitter and looks like that in the graph of Figure 3b.

As jitter levels decrease to picosecond levels, delay lines become complex. Their jitter may exceed that of the PLL, and they can induce additional jitter in the PLL if they are placed in close proximity (they usually are). Designing a “quiet” delay line is almost as complex as designing a PLL, and then it is difficult to diagnose which is at fault: the PLL or the delay line.



a) PLL BIST circuit

b) Jitter CDF

Fig. 3 **Sampling a PLL output using the PLL input delayed** (circuit, waveforms, and the resulting CDF)

B SERDES

As SoC gate counts reach hundreds of millions, higher pin counts are needed. Industry is choosing to use SerDes to reduce pin-count. Instead of using perhaps 32 single-ended data outputs, each switching at say 100 MHz with a 3.3 V swing, a differential pair of pins is used, switching at 3.2 GHz with a 500 mV swing. Assume a total capacitance of 10 pF per pin, power for the parallel case is $P = 32 \times fCV^2 = 348$ mW, but only 16 mW for the serial case, primarily due to the square-law impact of reduced voltage swing. Other advantages, like less switching noise (due to constant current differential drivers), more noise tolerance (due to subtraction of noise common to both wires of a differential pair), and less electro-magnetic radiation (due to each wire’s magnetic field cancelling the other’s), clearly make SerDes the better choice. However, at higher speeds, signals become more analog and non-deterministic, as can be seen in the typical SerDes received signal eye of Figure 4, so different DFT and test techniques are required.

In addition to the loopback technique described earlier, parametric testing is needed for SerDes. One common approach is to reuse the phase interpolator function already in many SerDes receivers. A phase interpolator is a circuit that derives equally spaced points in time (typically 32) between clock edges using analog techniques. Consider the waveforms in Figure 5. Two digital clocks, one delayed 90° relative to the other, are low pass filtered by capacitances, and then added. By adjusting the weights, A and B, of each waveform in the sum, any delay can be derived.

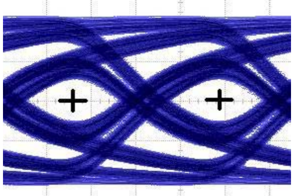


Fig. 4. SerDes signal eye (many waveforms overlaid)

Fig. 5. Phase interpolation

Phase interpolation works well in a SerDes receiver because the phase can be continually adjusted to center the sampling mid-way between data edges (ideally at the crosses in Figure 4). To use the interpolator in test mode, the transmitter output is looped-back to the receiver and they use the same reference frequency; pseudo random data is transmitted, phase adaptation is disabled in the receiver, a specific interpolated phase is selected, and the BER is measured. As the selected phase approaches the received data edges, the BER increases due to jitter. The disadvantages of this approach, aside from requiring specific analog circuitry in the receiver, are that the phase increments are determined by analog circuitry and are hence uncalibrated, the phase increments are coarse relative to typical random jitter, the measurement does not discriminate between different frequencies of jitter, and additional analysis in the tester is

needed to derive values for slew rate and duty cycle distortion, for example.

The only other reported DFT technique [3] that can measure jitter for SerDes is digital and does not require a specific SerDes design. It exploits a second mode that most SerDes designs have, usually called lock-to-reference or freeze mode, in which the receiver samples the incoming data using a clock derived from the receiver's own reference clock. In the DFT technique, the receiver is driven by a slightly different reference frequency than the transmitter (almost all SerDes must tolerate this in system use, but the capability is rarely tested in production because it requires a second clock). Thus, the receiver undersamples the waveform from the transmitter – it samples at a slightly different time for every serially-received bit, including the area around the data edges that contain jitter. As a result, the output parallel data can be analyzed digitally to measure parameters such as jitter, phase delay, and duty cycle.

C ADC AND DAC

The first attempt at BIST for analog/digital converters was published [11] in 1991. In that paper, it was simplistically proposed that random patterns could be applied to a DAC with its output looped-back to an ADC whose output was expected to contain the same bit sequence, especially if the least significant bits were ignored (since they were the most sensitive to noise). But these bits are the primary area of interest in measuring ADC and DAC performance.

Applying a constant current to a capacitor to generate a linear ramp has been reported multiple times [18][19], but is practical only for less than 10-bits accuracy because the technique offers no way to account for or cancel non-linearity in the ramp. A few purely-analog circuits have been proposed to generate sine waves on-chip, but not for testing purposes.

Using a digital signal processor (DSP) already on the IC is a practical way to generate a sigma-delta bit stream that produces a sine wave when low pass filtered by an analog filter [20], even though the filter might be relatively complex (e.g., fourth order). The output of the circuit under test (CUT) is analyzed using frequency domain analysis (Fourier transforms). Frequency domain analysis is used because measuring differential non-linearity (DNL) and integral non-linearity (INL) requires a linear ramp, and the low-pass filter required for a sigma-delta-derived ramp adds non-linearity to the ramp [6].

Another approach [6] applies a step-like exponential waveform derived from four different bit patterns filtered by a first-order low-pass filter in a way that does not add non-linearity. Variations up to about 20% in the resistance and capacitance have minimal effect on the result. The technique, as reported, used an off-chip op-amp resistor and capacitor, but an on-chip anti-alias filter of the ADC could be used instead. The approach measures INL, in effect, but does not measure DNL.

In another technique, called stimulus error identification and removal [16], a ramp with unknown linearity is applied to an ADC-under-test, the response is captured, and then the ramp is applied again but with an unknown, small, stable DC voltage offset. Mathematical analysis of the two responses can deduce the non-linearity in the ramp stimulus so that it can be subtracted to obtain the ADC's linearity.

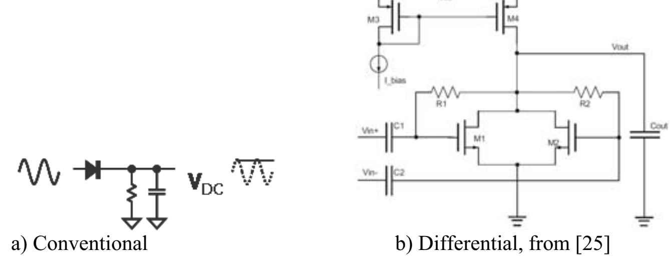
Very few approaches for DAC BIST have been proposed. Measuring the output voltage (or current) of a DAC, for thousands of different output voltages, is very time-consuming. Using a comparator is impractical because the offset voltage of the comparator is almost always voltage dependent and hence adds non-linearity. Ultimately, using a comparator is equivalent to converting the DAC into an ADC, in which case ADC BIST techniques can be used.

D RF

The most intensely studied new area for DFT in academia is for RF circuitry. It is the most difficult and time-consuming because all approaches must be verified in silicon – off-the-shelf circuits are unsuitable as are simulations. One emerging common DFT technique is power detection via a diode or diode-equivalent.

Accurately measuring properties of RF analog signals is much more difficult than for audio signals, for example, because RF waveforms have very small amplitude (typically 20~100 mVp-p) and very high frequency. DFT circuitry cannot accurately and economically convey RF signals off-chip, so first the RF signal properties must be converted to signals that can be conveyed more easily, and that means detecting only the peak amplitude of the signal (as done by an AM radio), or lowering the frequency by mixing (equivalent to multiplying) the RF signal with a similarly high frequency oscillator signal to produce a low frequency signal (as done by spectrum analyzers and also by radios).

Measuring the power of a waveform requires a squaring function, but if it is sinusoidal, a peak amplitude detector can be used. Figure 6a shows a conventional diode-based peak detector for single-ended signals, and Figure 6b shows a CMOS equivalent for differential signals. These circuits permit monitoring RF signals without adding a 50 ohm load, and they produce a DC output voltage proportional to the peak voltage of the RF signal.



a) Conventional b) Differential, from [25]

Fig. 6 RF power detectors

Many RF circuits contain both a transmitter and a receiver, so it is tempting to use loopback to facilitate test generation at the low frequency input to the transmitter and response analysis at the low frequency output of the receiver.

However, this approach prevents diagnosis of many potential problems because on-chip coupling between the transmitter and receiver cannot be distinguished from the off-chip loopback.

E Miscellaneous Analog

A long neglected area for DFT is miscellaneous analog, or, to use a digital analogy, 'random' analog. What makes this area most difficult is that the functions can occupy minimal silicon area (their advantage), so any dedicated BIST circuit is considered impractically large. No general technique has emerged, like scan has for digital circuits.

Oscillation BIST has been proposed for filters, op-amps, and many other functions [26]. The technique requires connecting the output of a CUT to its input, adding inversion or gain as needed to ensure oscillation occurs. The oscillation frequency is measured relative to test limits derived from analog fault simulation. The technique is attractive because it does not require a stimulus, and frequency measurement requires little more than a binary counter. Unfortunately, a single measurement can not provide diagnosis if it is incorrect, and it can not measure noise.

Almost all other proposals for DFT and BIST of general analog functions focus on filters and amplifiers – circuitry that conveys sinusoidal stimuli – which excludes functions like voltage regulators, mixers, comparators, etc.

IV. Principles of Analog BIST

Analog BIST is poorly understood by digital designers. Reusing digital BIST techniques, such as applying pseudo-random patterns to analog circuits has been tried by researchers [11][12] but analyzing the resulting output has proven too complex and undiagnostic. Diagnosis is a key requirement of any analog DFT or BIST.

Analog BIST is poorly understood by most analog designers too. Analog design usually requires designing the most accurate, highest speed, lowest power, circuit that can be designed into a given process technology. So it can be difficult to understand how BIST could measure performance even more accurately if it is necessarily implemented on the same IC. (A PLL designer asked me at a conference in 2009, "If your PLL BIST is so accurate, why don't you design PLLs?") This section will attempt to explain this paradox.

For the inaccuracy of a measurement to be insignificant, in an engineering sense, compared to a CUT, the measurement's inaccuracy must be at least an order of magnitude less than that of the CUT. In an ADC or DAC, this means more than three bits higher resolution.

But the challenge for designing analog BIST is even greater, especially compared to BIST for digital circuitry. Whereas digital BIST stimulus and response is always a logic value, 0 or 1, an analog stimulus could be a DC voltage, a linear ramp, or a sinusoid waveform. The response could be very different from the stimulus: for a voltage-controlled oscillator (VCO), the stimulus is a DC voltage and the response is a frequency; for a delay-locked loop (DLL), the stimulus is a clock frequency and the response is a phase delay; for a phase-locked loop (PLL), the stimulus could be a change in frequency and the response could be a lock time (the time until the output phase is continuously aligned to the input's new phase). And there are many datasheet specification-based analyses that BIST may be required to perform, ranging from computing harmonic distortion and signal-to-noise ratio, to computing the duty cycle, phase delay, gain, or impulse response.

Diagnosing a manufacturing defect, based on a failure of one these specification-based tests, may be impossible, so industry has been trying to develop BIST that measures performance more structurally, while maintaining correlation to specification-based tests.

The challenge of designing analog BIST is daunting but the benefits that industry enjoys for digital BIST are very enticing. Digital BIST usually permits faster test and easier diagnosis of speed-related defects, in a systematic, general, and automated way. BIST facilitates more test parallelism, especially multi-site testing, which dramatically decreases the average test time per IC. Another benefit may be higher yield, due to lower pin-count test access (less pad damage, and no contact resistance variance), timing accuracy that inherently scales with technology, and fewer signal loading effects.

Fortunately, BIST does not need to be simultaneously more accurate, and faster, and lower power. The following seven principles show how the design constraints for analog BIST circuitry can be more relaxed than for the CUT, while delivering the required test accuracy and speed.

A Test the BIST

Automatic test equipment (ATE) is calibrated before it leaves the factory, and is recalibrated periodically, typically weekly because the calibration may take an hour. For BIST test results to be trustworthy, the BIST circuitry itself must be tested, but in milliseconds because the time is incurred for every device tested.

BIST for digital circuitry is essentially a digital finite state machine (FSM) comprising flip-flops and combinational logic that are tested via scan path access in less than a millisecond. If BIST for analog is purely digital, then it too can be tested via scan. It is possible for parametric BIST for functions like SerDes and PLLs to be purely digital because those functions have digital inputs and outputs, even for BIST with picosecond accuracy [3][4], however, most PLL BIST circuits include delay lines whose delay must be calibrated [13][14] or are assumed to be accurate [15]. Delay lines can be calibrated by connecting their input to their output, while ensuring the output is inverted relative to the input, to create a ring oscillator. The oscillation period is proportional to the delay line's delay, and it can be measured using a binary counter whose output latches and resets, once every N oscillation cycles, a counter clocked by a known frequency.

Testing or calibrating other analog parameters of the BIST is more difficult. For example, if a linear ramp is generated on-chip to measure linearity of an ADC, calibrating

the linearity of the ramp is complex. The stimulus error identification and removal technique [16] mentioned earlier is the only technique reported that can deduce the non-linearity in the ramp stimulus by analyzing only the ADC output.

One way to avoid calibration of the analog circuitry is to ensure that significant variations in its performance have no significant effect on the measurement. This is the principle behind sigma-delta converters which transfer signal noise to a higher frequency to permit more performance variation in their analog low-pass filters. Another way to avoid calibration of the analog circuitry is to use as little analog circuitry as possible.

B UNDERSAMPLE

In analog, there is always a trade-off between speed and accuracy: a 24-bit accuracy, 20 kHz ADC IC can be purchased for about the same price ($10) as a 12-bit, 20 MHz ADC. Consequently, for BIST to achieve higher accuracy than a CUT, the BIST must operate more slowly. This can only be achieved with undersampling. According to the Nyquist criterion, a signal must be sampled at a rate that is greater than twice the highest frequency of interest to capture all the information in the signal. Sampling at a lower frequency is called undersampling, and can still capture the same signal waveform but some frequency information is lost. As illustrated in Figure 7, when a sine wave is sampled slightly later in each of its cycles (therefore, fewer than two samples per cycle), and only the samples are viewed, the samples create the original wave shape but at a much lower frequency.

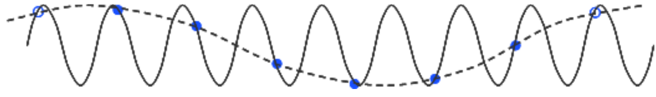


Fig. 7 Undersampling a sine wave
(sampled signal, sample points, and reconstructed output waveform)

There are limitations: the input bandwidth of the sampler must be greater than the signal's frequency, the sampler's edge placement accuracy must be finer than the reciprocal of twice the frequency of interest, noise at higher frequencies will fold into the frequency band of interest, and the signal must be periodic.

In oscilloscopes, undersampling is used to achieve equivalent-time sampling (ETS) wherein a periodic multi-gigahertz signal might be sampled less than a million times per second, yet an accurate waveform is displayed. After each sample, the oscilloscope's sampling delay is changed with picosecond accuracy, and the sampler has time to settle to the required accuracy.

Similarly, in the PLL BIST discussed earlier for Figure 3, jitter in the output of a PLL is sampled by the PLL's own input reference frequency. The BIST circuitry operates at the reference clock rate, and requires both the reference clock and the delay line to have comparable or lower jitter than the PLL output. As technologies shrink, jitter in a delay line does not because it is proportional to the delay line's delay [17].

In another PLL BIST technique, the PLL's output is sampled by a flip-flop clocked at a frequency, f_S, that is slightly asynchronous relative to the PLL's reference frequency. The asynchronous sampling clock is derived using an off-chip clock conditioner PLL (often needed anyway to reduce jitter in the clock from the tester) or another PLL in the DUT. The output will be a square wave at a much lower frequency, equal to the difference between the sampled frequency and the nearest multiple of the sampling frequency. Jitter in either clock, if the jitter has a high enough frequency, will result in "jitter bits", as shown in the waveform of Figure 8, which can be analyzed by relatively slow circuitry that operates at the sampling rate [3], and requires the sampling clock to have comparable or lower jitter than the PLL output. This approach allows a single latch to be placed close to the PLL-under-test and all other logic to be more distant, where its operation will have no effect on the PLL. As technologies shrink, the sampling clock path delay can decrease, therefore, so can its jitter. The technique is capable [3] of measuring picosecond jitter, as well as duty cycle, frequency, lock time, and phase delay.

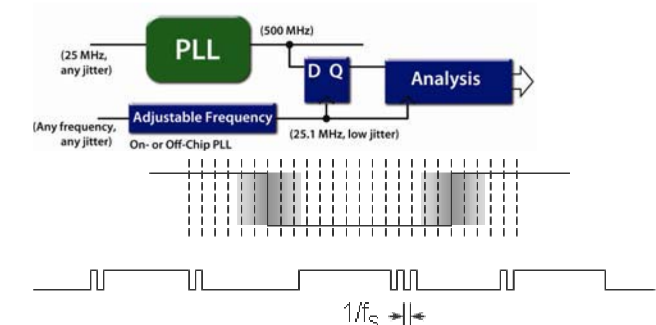


Fig. 8 Undersampling a clock that has jitter
(circuit, equivalent sampling resolution, and Q output)

C SUBTRACT TO IMPROVE ACCURACY

Accuracy is the closeness of a measurement to the true value. In ATE, calibration is intended to measure and then remove systematic errors, such as leakage current, voltage offset, gain errors, and wire delays; this must be done by a BIST too.

A few subtraction techniques are used widely in analog functions. In auto-zeroing, a node voltage is sampled while it is grounded and then the result is subtracted from the result sampled while the signal-of-interest is present. In correlated double sampling, a node's voltage is sampled while the signal of interest is inverted and then the result is subtracted from the result sampled while the signal is not inverted. In offset cancellation, a feedback loop is closed to force the input of a gain circuit to the offset voltage that would produce a non-offset output, and this offset is saved and subtracted from subsequent input samples. In differential signaling, inverted and non-inverted versions of the signal are sampled and subtracted simultaneously.

Analog subtraction techniques are best when the signal properties vary continuously but digital circuits can implement

subtraction more accurately if the signal properties of interest are stable during the measurement. The circuit shown earlier in Figure 3 can be used to subtract the delay line's delay when the counter's output is 840 from its delay when the counter output is 500 – this will correspond to the root mean square (RMS) value of jitter if the jitter's histogram has a stable Gaussian shape, which is typically true. However, the peak-to-peak value of jitter typically is not a stable value and is rarely used to screen electronic components for this reason.

D ADD TO IMPROVE PRECISION

Precision is the variation observed when a parameter with a constant true value is measured multiple times, and it affects measurement repeatability. The variation is typically caused by random noise, periodic interference, and quantization.

Consistent with the Central Limit Theorem, measurement variation caused by random noise can always be reduced by adding multiple samples to obtain an average result – the variance improvement is linearly proportional to the number of samples, and the improvement in sigma (or RMS) is hence proportional to the square root of the number of samples.

Averaging can also reduce the impact of periodic noise, but not in all circumstances. As reported in [3], if the sampling is asynchronous to the periodic interference, then measurement repeatability is unaffected. But when sampling is made synchronous to the interference, the impact of the interference can be completely eliminated.

If measurement quantization is much coarser than the true signal variation and random noise, then no amount of averaging can improve the precision. But if the random noise is larger, or the quantization resolution is improved, then the effects of quantization can be improved just as they are for random noise.

Analog techniques that accomplish averaging include low pass filtering, using resistance and capacitance, and integration, using constant currents and capacitance. Similarly to subtraction, analog techniques are best for continuous averaging and short term averages but digital techniques are best for averaging samples over longer time intervals.

E BIST CIRCUIT MUST HAVE HIGHER YIELD THAN CUT

To be economically viable, any DFT and BIST circuitry must have no significant impact on yield. The yield of digital circuitry is dominated by random defectivity so is inversely proportional to the circuit's area or gate count. Logic BIST that occupies 1% of the chip area (a typical case) reduces yield by a similar amount.

Analog functions that have less design margin than digital circuits can suffer from parametric yield loss disproportionate to the function's area. For example, if an analog test bus uses an analog buffer to separate off-chip capacitance from an on-chip CUT and that buffer has a parametric yield of 99%, then regardless of the area of the whole IC, the buffer will have the same yield impact as logic circuitry that occupies 1% of the whole IC. Therefore, it is essential that DFT and BIST blocks use minimal analog circuitry, and that their parametric margin is better than or comparable to that of digital circuitry. The best way to achieve this is to use predominantly digital techniques for analog BIST.

Circuits [18][19] that use on-chip linear ramp generators for ADC BIST can be designed to deliver very linear ramps, but as their designed accuracy is increased, the more important it is to test that accuracy on each device, and the more likely that there will be parametric yield loss for the BIST circuitry alone. Ramp generation techniques that use sigma-delta bit patterns [20] or pulse-width modulation [6] comprise digital circuitry followed by a low pass filter whose characteristics are unimportant, so they can achieve a higher yield if their digital gate-count is small. However, low gate-counts are difficult to accomplish because analog circuits-under-test may be comparatively very small and have many parameters that must be tested.

One way to minimize BIST circuit area is to use a single BIST block to test many analog functions, but the variety of analog functions and the difficulty of conveying analog signals accurately by analog bus (as discussed earlier) makes this approach challenging. Serial data busses that carry sigma-delta bit-streams or undersampled data are an alternative [5]. Figure 9 shows how the responses from multiple analog circuits could be captured by a digital bus, serially or in parallel, and shifted to a shared analysis circuit.

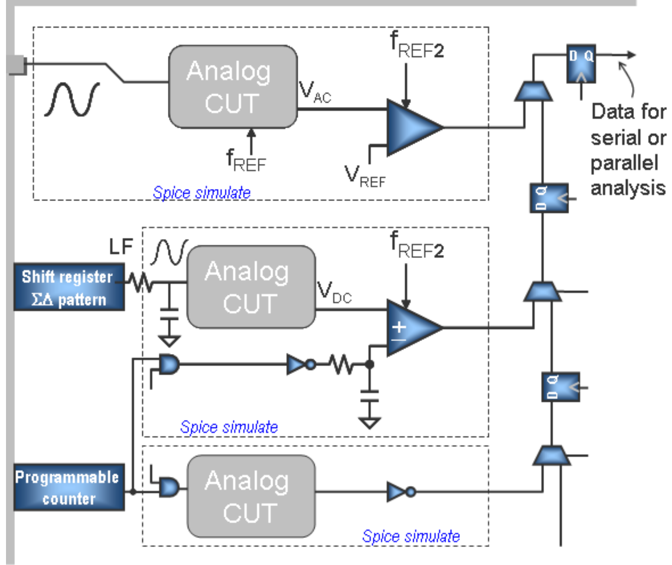


Fig. 9 Serial digital bus DFT technique for random analog

F SPECIFICATION-BASED STRUCTURAL TESTS

Digital structural tests are aimed at detecting and diagnosing defects in the topology of logic gates, independent of their eventual function. Automatically-generated, scan-loaded patterns verify that logic gates between flip-flops deliver the required combinational function, within a clock period. To verify that AC and DC coupling between logic

gates and interconnect are not excessive, test patterns are typically generated to detect each structural fault multiple times (denoted as N-detect patterns), each time with different logic values and transitions in adjacent nodes.

A key characteristic of structural tests is that the stimuli are quite different from stimuli received during functional operation – scan-based test of logic is a prime example. Two examples of analog structural tests are oscillation BIST [22] in which an op-amp or filter's output is connected to its own input to cause an oscillation whose frequency is measured, and applying a random digital pattern to an analog filter [23] while convolving the input and output signals to derive the filter's impulse response.

Purely structural tests are most efficient when there is lots of parametric margin (as is typical for digital circuits). But yield and fault coverage for purely structural tests are usually unacceptable for analog because the correlation between structural variations and functional performance varies greatly from design to design, and from structure to structure within a design. When design margin is slim, the yield difference between specification-based tests and structural tests is usually unacceptable [21].

However, function specification-based tests too are impractical for analog circuits in an SOC because too many tests are required and analog access prevents applying many of the tests. The solution is to use structural tests that approximate specification-based tests.

There are many specification-based structural tests in common usage. An ADC might never receive a linear ramp signal in its application, but a ramp test is very diagnostic and similar to function signals. A SerDes might never transmit and receive a long 1010 pattern, but loopback testing with this pattern facilitates much easier measurement of random jitter [4] because that pattern contains no data dependent jitter, it is similar to function-mode signals, and it easily generated.

G Deliver Digital Measurements and Pass/Fail

BIST that only delivers a pass/fail result prevents diagnosis, characterization, and optimal choice of test limits. BIST that only delivers measurements prevents use of simple digital test patterns, and requires the tester to perform calculations which greatly increases test time. BIST that produces an analog result, such as a DC voltage proportional to a parameter of interest, must convert that analog result to a digital value on-chip to ensure that the result is accurately conveyed to ATE and can be monitored remotely for in-system test. One approach is to compare on-chip an analog voltage result to a voltage provided by the ATE, and then shift out the comparator's one bit output. This approach can be cumbersome if comparisons to upper and lower limits are needed: delivering a DC voltage typically adds 3~15 ms to test time, requires an extra pin for each simultaneous comparison, increases sensitivity to noise and voltage offset, and prevents reusing the test in-system.

If an analog measurement is conveyed as a binary equivalent value to ATE, the test pattern must examine all bits of the binary word, then compare the word to upper and lower test limits to produce a pass/fail result. This cannot be achieved in ATE with a simple test pattern – a sub-program is needed to do the comparison which slows the test and limits the choice of ATE. The fastest and simplest solution, from an ATE perspective, is for all comparisons to be performed on-chip and then output as a serial binary value along with pass/fail bits. The binary value is used for diagnosis and analysis; only the pass/fail bits are used for go/no-go testing.

Standard, system-level accessible test access is increasingly important for in-system diagnosis of IC problems. Many IC problems that are caused by noisy power rails or inter-IC timing become "no trouble found" (NTF) after the IC has been removed from its board and retested on ATE. The 1149.1 digital TAP is widely used for conveying digital results because its timing is simple and robust, but a few other robust ports have been used, such as I^2C and proprietary buses. IEEE P1687 ("IJTAG") is an emerging, proposed standard [24] that supports diagnosis of inter-IC problems by enabling 1149.1 access to on-chip "instruments" during IC and in-system test.

V. Conclusions

This paper presented the two most common general analog DFT methods, analog test buses and loopback, and DFT approaches used for the most common analog functions, which are PLLs, SerDes, ADCs, and RF. A few methodologies were described for 'random' analog functions, but most companies use *ad hoc* techniques to provide more diagnosis – this is an area still in need of an approach as systematic as the scan path DFT used for random digital logic.

DFT for RF circuitry is presently the area receiving most academic research attention, but attention may eventually shift back to the common analog functions because they are now the bottleneck in test development and device test time.

The paper finished by describing seven principles essential to implementation of practical analog BIST, showing how analog BIST can be more accurate than the circuit it is testing on the same IC.

References

[1] Std. 1149.1a-1993, *IEEE Standard Test Access Port and Boundary-Scan Architecture*, IEEE, New York, 1993

[2] Std. 1149.4-1999, *IEEE Standard for a Mixed Signal Test Bus*, IEEE, New York, 2000

[3] S. Sunter, A. Roy, "Noise-Insensitive Digital BIST for any PLL or DLL", *J. of Electronic Testing: Theory and Applications*, vol. 24, pp. 461-72, Oct. 2008

[4] S. Sunter, A. Roy, "Structural Tests for Jitter Tolerance in SerDes Receivers", *Proc. Int'l Test Conf.*, Oct. 2005

[5] S. Sunter, "A General Strategy for BIST of High-Speed Analog Functions", *Informal Proc. of Workshop on Test and Verification of High-Speed Analog Circuits* (TVHSAC), Nov. 2009

[6] A. Roy, S. Sunter, “High-Accuracy Stimulus Generation for A/D Converter BIST”, *Proc. Int’l Test Conf*, pp. 1031-39, Oct. 2002
[7] S. Sunter summary of statement by K. Arabi, “ITC 2009 Panels – Part 1, ‘Can EDA Help Solve Analog Test and DFT Challenges?’” *IEEE Design & Test*, Jan/Feb 2010
[8] F. Poehl, F. Demmerle, J. Alt, H. Obermeir, “Production Test Challenges for Highly Integrated Mobile Phone SOCs – A Case Staudy”, *Proc. of European Test Symp*., pp. 17-22, May 2010
[9] V. Zivkovic, F. van der Heyden, G. Gronthoud, F. de Jong, “Analog Test Bus Infrastructure for RF/AMS Modules in Core-Based Design” , *Proc. of European Test Symp*., pp. 27-32, May 2009
[10] R. Andlauer, P. Vu, “Analog Test Bus Grows in Importance”, *Electronic News*, May 2002
[11] M. Ohletz, “Hybrid Built In Self Test (HBIST) for Mixed Analog/Digital Integrated Circuits”, *Proc. European Test Conference*, May 1991
[12] C-Y Pan, K-T Cheng, “Implicit Functional Testing for Analog Circuits” *Proc. of VLSI Test Symp.,* pp. 489-494, 1996
[13] K. Jenkins, A. Jose, D. Heidel, “An On-chip Jitter Measurement Circuit with Sub-picosecond Resolution”, *Proc. of ESSCIRC*, pp. 157-160, 2005
[14] K. Nose, M. Kajita, M. Mizuno, “A 1-ps Resolution Jitter-Measurement Macro Using Interpolated Jitter Oversampling”, *J. of Solid State Circuits*, pp. 2911-20, Dec. 2006
[15] B. Casper, A. Martin, J. Jaussi, J. Kennedy, R. Mooney, “An 8-Gb/s Simultaneous Bidirectional Link with On-Die Waveform Capture,” *IEEE J. Solid-State Circuits*, vol. 38, no. 12, pp. 2111-2120, Dec. 2003
[16] L. Jin, K. Parthasarathy, T. Kuyel, D. Chen, R. Geiger, “Accurate Testing of Analog-to-Digital Converters Using Low Linearity Signals With Stimulus Error Identification and Removal,” *IEEE Trans. Instrum Meas.*, vol. 54, pp. 1188 – 1199, June 2005
[17] A. Abidi, “Phase Noise and Jitter in CMOS Ring Oscillators”, *J. Solid-State Circuits*, vol.41, pp. 1803-1816, Aug. 2006
[18] B. Provost, E. Sanchez-Sinencio, “Auto-Calibrating Analog Timer for On-Chip Testing”, *Proc. of Int’l Test Conf.*, pp. 541-548, Sept. 1999
[19] F. Azaïs, S. Bernard, Y. Bertrand, X. Michel, M. Renovell, “A Low-Cost Adaptive Ramp Generator for Analog BIST Applications”, *Proc. of VLSI Test Symp.*, pp.266-71, Apr. 2001
[20] M. Hafed, G. Roberts, “Test and Evaluation of Multiple Embedded Mixed-Signal Test Cores”, *Proc. of Int’l Test Conf.*, pp. 1022-30, Oct. 2002
[21] M. Sachdev, “A Realistic Defect Oriented Testability Methodology for Analog Circuits”, *Journal of Electronic Testing: Theory and Applications*, vol. 6, no. 3, pp. 265-276, June 1995
[22] K. Arabi, B. Kaminska, "Testing Analog and Mixed-Signal Integrated Circuits Using Oscillation Test Method", *IEEE Trans. on Computer Aided Design*, vol. 16, no. 7, pp. 745-753, July 1997
[23] C. Pan, K. Cheng, “Test generation for linear time-invariant analog circuits,” *IEEE Trans. on Circuits and Systems-II*, vol. 46, no. 5, pp. 554-564, May 1999
[24] A. Crouch, “IJTAG: The path to organized instrument connectivity”, *Proc. of Int’l Test Conf.*, pp. 1-10, Oct. 2007
[25] C. Zhang, R. Gharpurey, J. Abraham, “Low Cost RF Receiver Parameter Measurement with On-chip Amplitude Detectors”, *Proc. of VLSI Test Symp*., May 2008
[26] G. Sánchez, D. de la Vega, A. Rueda, J. Diaz, *Oscillation-based test in mixed-signal circuits*, The Netherlands: Springer, 2006

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

An Overview of Integrated Circuit Testing Methods

Anne Gattiker
IBM Austin Research Lab
gattiker@us.ibm.com

Phil Nigh
IBM Systems & Technology Group
nigh@us.ibm.com

Rob Aitken
ARM
Rob.aitken@arm.com

INTRODUCTION

Increasing complexity, speed and pincount and decreasing time-to-market pose many challenges for integrated circuit (IC) manufacturers. IC testing is one of those challenges growing in importance and in overall contribution to the cost of a manufactured die. IC speeds that exceed automated test equipment (ATE) capabilities and sheer number of transistors to be tested are examples of new realities that are making test more difficult and more challenging. However, as IC failure mechanisms and ICs themselves become more complex, testing is becoming not only more difficult, but also more important. Specifically, the increasing difficulty of physical failure analysis (PFA) coupled with the time-to-market-driven need for rapid yield-learning is creating an increasingly important role for test not only in separating good dies from bad, but also in giving feedback to the manufacturing process about imperfections occurring during fabrication.

This chapter presents an overview of microprocessor and application specific integrated circuit (ASIC) testing. The chapter begins with a description of key industry trends impacting testing. Next, a basic overview of logic and memory test is given, where technical issues that are causing methodology changes are emphasized. The overview is followed by a section higlighting strategies for overcoming several emerging problems in test created by today's technologies.

TEST-RELATED TRENDS

The majority of this chapter is aimed at providing an overview of the test methods commonly used in the industry today. It is important, however, not only to understand where the industry is today, but also where it is heading. In this section, we highlight some key trends that will impact how ICs will be tested in the future.

There are some major challenges facing the IC industry related to the manufacturing test of future devices [1]. These key challenges include:

- **On-chip frequencies are in the multi-GHz range exceeding the speed capabilities of testers and test fixtures.** The cost of directly testing these ICs at their full speed is becoming prohibitive.
- **Chip densities are starting to exceed 100M logic circuits.** Without test methodology changes, test times could be in the "greater than one minute" range and the size of test vectors could exceed 20 Gbytes.
- Manufacturers are moving toward **integrating many different types of circuits on a single die and combining multiple die into an integrated component (such as a 3D IC)**--including putting analog and DRAM circuits on previously digital ICs. Test methods that have conventionally been used for various circuit types will change when these same circuits are embedded on a device with many other circuits.
- **Some testing methods**--particularly I_{DDQ} testing and high-voltage screening--**are becoming less effective** (or are projected to become less effective) due to device and voltage scaling. In addition, defects causing more subtle electrical fail types (e.g., timing-related failures) are becoming more common in future technologies.
- **Rising test costs** may force new testing methods to become adopted. These rising costs are caused by the trends listed above, their impact on the capital cost of automatic test equipment (ATE) and the increased test times.

Despite these trends toward greater complexity, semiconductor market forces are trying to drive down testing costs and to shrink the time between product introductions. Meanwhile, the manufacturing cost per transistor is decreasing, but the cost to test a transistor remains at best constant over time. Hence, market forces are directly in conflict with technology trends--either improved methods must be adopted or market expectations must change.

TESTING BASICS

In this section, we provide a brief description of the most common tests applied in the industry today. At least two levels of testing are normally applied to products before they are shipped: a test at the wafer level, often called "wafer test" or "probe test" and another test after the die is packaged, often referred to as "package test" or "final test." Some products may see additional test steps due to additional requirements related to temperature coverage, memory test redundancy verification or reliability screening. Figure 1 shows a typical test flow in which the IC undergoes a wafer level test, pre- burn-in package test, burn-in and post-burn-in final test. Often extra "data collection" or "characterization" tests are applied for a sample of parts. The tests described in this section may be applied at either wafer or package test or both.

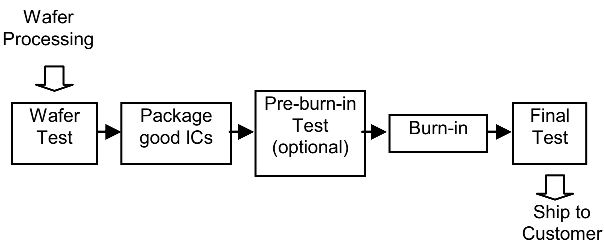


Figure 1. Typical IC test flow.

Types of tests

Contact / power-up tests. Normally the first test applied is aimed at ensuring that there is reliable contact between the tester and the device. Each device I/O should be tested for a low resistance connection to the ATE.

I/O parametric tests. These tests ensure that the parametric characteristics of the device's I/Os are within specification. Some of these specifications include guaranteed noise margins, drive capabilities and input leakage currents. For example, each I/O is functionally tested at its minimum voltage level to ensure it meets specification.

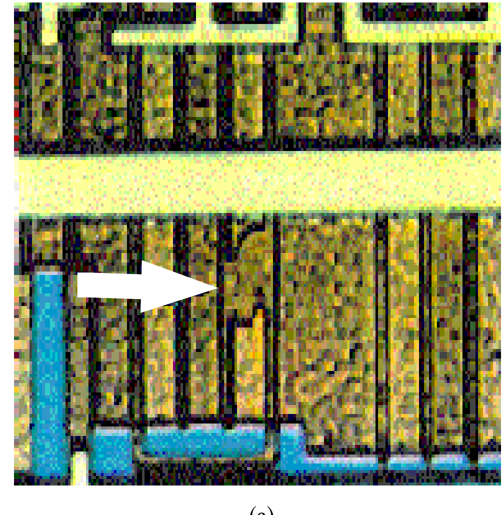
I_{DDQ} testing. Detecting defects by measuring the quiescent current of the power supply (typically V_{DD}) is called I_{DDQ} testing. This test method is most effective for fully complementary static CMOS circuits where there are no static paths of current, outside of unavoidable leakage paths.

There are many papers describing the effectiveness of I_{DDQ} testing [1] -[25] . Most defects cause abnormally high I_{DDQ}. Most defects are activated by some, but not all, circuit states so multiple I_{DDQ} measurements must be performed. Performing 4 to 30 I_{DDQ} measurements is most common in the industry (although many more measurements are performed for some products).

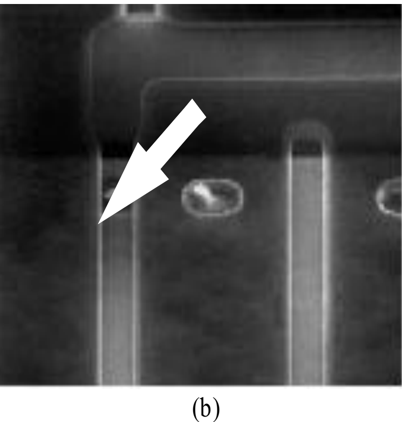
A benefit of I_{DDQ} testing is that it can detect reliability defects, i.e., defects that do not initially cause a circuit failure, but which will do so in the future. Some reliability failures pass all other tests, but have abnormally high I_{DDQ} during initial testing. Detecting these defective ICs at wafer or package test is a tremendous advantage.

There is some controversy over whether or not chips should be rejected if they fail only IDDQ testing. An argument for rejecting them is the defects detected may turn into reliability-related failures as described above. An argument against rejecting IDDQ-only fails is that the defected flaws may be "nuisances" that never cause functional failure. Performing physical failure analysis on IDDQ-only fails provides insight into their nature. Figure 2 shows physical failure analysis results for two IDDQ-only failing chips. Results such as these demonstrate that indeed the IDDQ test is detecting real defects in the manufactured circuit.

As technology continues to advance, however, the effectiveness of conventional I_{DDQ} testing is getting worse since the normal background leakage currents are going up [25]. For technologies beyond 90nm, I_{DDQ} testing is losing effectiveness in its ability to detect single spot defects. Low power products (with lower background leakages) may continue to use I_{DDQ} testing even for more advanced technologies. Figure 3 shows typical I_{DDQ} levels for high performance devices across a range of technologies. Note that I_{DDQ} levels have roughly been increasing by a factor of 10 with each technology generation.



(a)



(b)

Figure 2. Photographs of two defects that behaved as "IDDQ-only" failures. (a) Extra polysilicon that shorted two input to a NOR gate. (b) Gate oxide short on the input of an XOR gate.

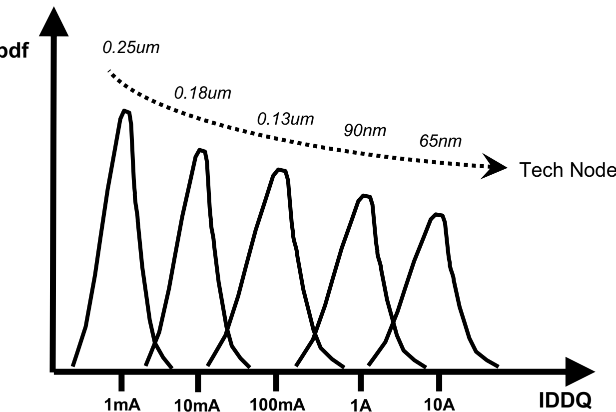


Figure 3. I_{DDQ} levels for defect-free chips across technology generations.

I/O performance testing. These tests are aimed at ensuring that the IC's inputs and outputs operate at the required speeds. Typical I/O timing specifications include the setup and hold times between signal inputs and clock lines. Historically, I/O performance testing has been a functional test driven by the tester, but companies are now adopting "on-chip I/O Wrap" methods for testing I/Os at speed [26][27].

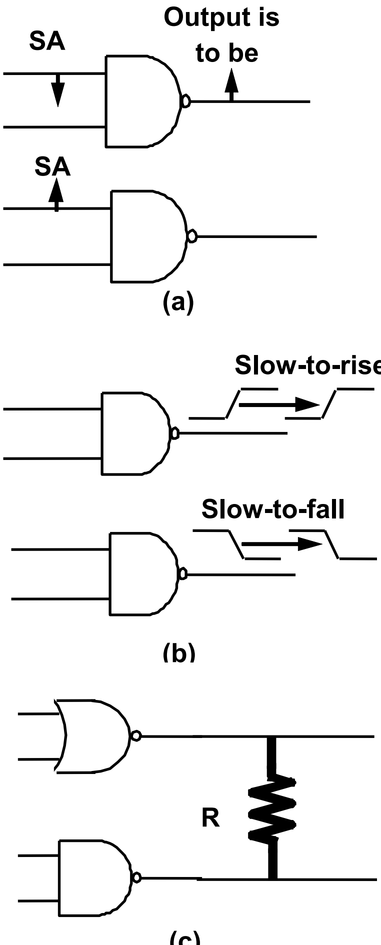


Figure 4. Fault models: stuck-at (a), transition fault (b), and bridging fault (c).

Logical tests. The most widely known test method is where logical data is applied to the chip and the response is compared with expected values. The purpose of this testing is to verify that the boolean operation of the device is correct. Typically the logical input and output values are stored in the ATE. The speed of the device may or may not be measured during logical testing. There are two basic forms of logical testing: functional and structural testing. They are described in more detail later in this chapter.

Speed and power binning tests. In addition to determining if a chip has a defect or not, it is often required to determine the speed at which the device fully operates. For example, the same microprocessor design may operate at 1.6GHz, 2Gz or 2.4GHz depending on the precise processing conditions seen by the IC. Manufacturers are eager to support speed binning since faster parts can be sold at a much higher price. In some applications, such as battery powered devices, frequency is fixed, but binning may still be possible on power consumption (leakage based) or operating voltage required to achieve a given frequency. In each of these cases, it is desirable to use test to separate parts into categories (bins) based on additional product value (performance and/or battery life).

Characterization tests. In addition to determining the passing and failing devices and the device speed, manufacturers often apply additional tests to learn as much as possible about the semiconductor processing conditions, for both control and optimization purposes. Such tests may include additional timing measurements, leakage current measurements, saturation current measurements and maximum/minimum operational voltage measurements. Typically manufacturing test stops with the first test failure to minimize test times. During characterization testing, however, all tests may be applied even for failing parts to gather information about all tests. Examples of such information are failing bits during logical tests and memory bitmaps, both of which can be used for fault localization.

Functional and Structural Testing Definitions

The basic idea of functional testing is to test the device in the same way that it will be used in the final application. For example, an adder circuit is tested by adding a sample of numbers. Typically, functional test patterns are manually created by designers whose goal it is to validate the correctness of the design, although vector generation may also be assisted by automated verification tools. A subset of these functional verification patterns are then reused for manufacturing test of the ICs. The test vectors for functional testing are usually applied at the full speed of the device.

The goal of structural testing, on the other hand, is to ensure that the logical circuit structure is defect-free. Test patterns are generated automatically. Automatic test pattern generation (ATPG) uses a gate-level representation of the circuit and a representation for defective circuits called a fault model. Figure 4 illustrates several fault models. The most commonly used fault model (called the stuck-at fault model) represents defects as causing logical gate inputs or outputs to be stuck at either logical '0' or logical '1' levels. The transition fault model is being used extensively in the industry today. (Fig. 4b.) Other fault models (e.g., I_{DDQ}, transition, path delay, bridging) have been proposed, but are used much less frequently [28].

While test generation based on the stuck-fault and the transition fault model has so far resulted in tests with adequate quality, new technologies may bring new fail mechanisms that require extensions, as discuss later in this chapter.

Scan Design

A key enabler of structural testing is scan design [29]. As illustrated in Figure 5, this is a design methodology in which the internal storage elements (flip-flops and/or latches) are connected into serial chains. Scan-based testing is performed by serially shifting input bits into the scan chains, applying a functional clock to capture the response of the combinational logic between the scan chains, and then serially shifting the responses out of the chains. Since the state of the circuits (i.e., the register contents) can directly be controlled and observed, the test generation problem is effectively reduced from a sequential one to a combinational one. In many chip designs, the scan inputs and outputs are shared with functional I/O through multiplexing, either in hardware or managed by the ATE.

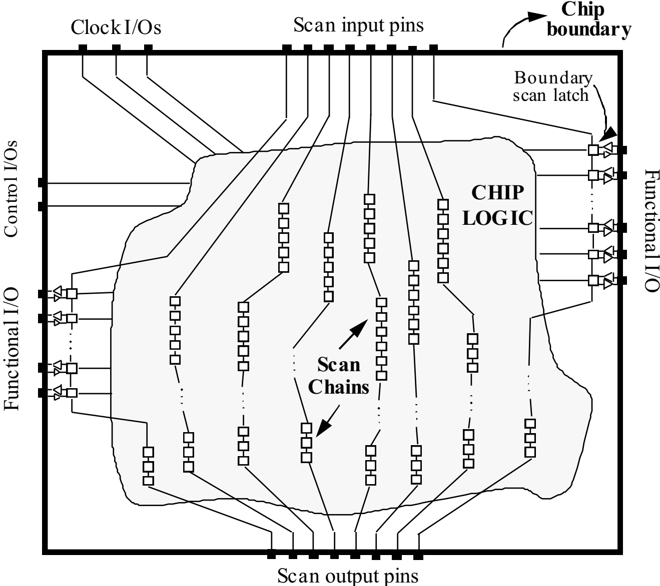


Figure 5. Scan Design

Designs for which sequential tests can be generated are limited to a size in the range of 50,000-100,000 logic gates. The design size limit is higher for combinational logic test generation where ATPG can be applied even for the largest ICs being designed today. Thus, scan design effectively enables automated test pattern generation for all logic ICs. Without scan design, ATPG would be limited to small macros only.

Scan design, however, does come with a price of area overhead and additional circuit delay. The improved test coverage, lower cost of test generation, better diagnostic capabilities, possibility of adding built in self-test (described below), and reduced tester requirements are benefits that help offset this cost and result in scan-based test's wide adoption in the industry today.

Scan-based Delay Testing

Note that delay testing is also possible using scan design. To implement scan-based delay testing, inputs are shifted into the chain in such a way that they are seen only at an intermediate node in each scan latch--their values are hidden from the combinational logic driven by the latches. Then a fast sequence of clocks launches the data from the scan latches to the combinational logic between the chains and then quickly captures the results into the scan chain registers [30]-[32]. This method is known as "launch-capture" delay testing and takes advantage of the excellent controllability and observability of scan design and therefore has the potential to provide very thorough delay test coverage.

A number of companies have observed and published that there is an increasing number of delay defects with advanced technologies [33]. Due to this trend, delay testing is becoming a requirement for most logic ICs built in advanced technologies.

Functional vs. Structural Testing Trade-offs

Choosing functional versus structural testing requires considering a variety of trade-offs. Even long-time advocates of functional test such as Intel are revisiting their test approaches [34]. Some of the trade-offs are described below.

First, an advantage of functional testing is the possibility of "at speed" testing, i.e., operating the chip during test at system speeds. Functional testing can be used by the designer to verify that the most used or the most critical operations meet design specifications. A goal of functional testing is to mimic the ultimate system environment as closely as possible. For example, functional test attempts to run the device at the final frequency of the system (or a little faster to provide a guardband). This "at-speed" testing also allows the already-mentioned speed binning--the process of classifying the chips according to the range of frequencies at which they will operate.

A disadvantage of "through-the-pins" functional testing is that at-speed functional testers are needed. Due to the required speeds, timing accuracy and power supply requirements, as well as the number of I/O pins, these testers are increasingly expensive, yet quickly outdated: commitment to at-speed functional testing entails huge capital investments.

The other reason functional testing is becoming too expensive is that it relies heavily on manual test generation. Often the designer is needed for the intimate understanding of his or her part of the design for the test generation. But the designer will also have to have a thorough understanding of the whole architecture of the chip in order to know how to get the necessary information to and from the part of the design. ATPG is critical to satisfy short product development cycles and to reduce the engineering effort in developing test patterns. Only structural, not functional, testing enables ATPG.

Since functional testing requires no changes to the design for testing, there is no area overhead involved. On the other hand, for structural testing additional design specifications are introduced to facilitate testing and making the test generation process automated. The area overhead incurred has to be compared with the lower test generation costs and reduced time to market.

A major advantage of structural test is reduced ATE requirements. For example, scan design can be implemented at the design's I/Os (called boundary scan) to enable the entire chip to be tested with only a few I/O pins [35]. This type of "reduced pin-count testing" can decrease ATE cost both by requiring fewer tester pins and enabling easier ATE reuse for many different designs.

Similarly, verifying the speed of the circuit by performing launch-capture delay tests rather than applying the test vectors at the full-speed of the IC [30][31] can greatly reduce ATE cost and enable ATE reuse.

Built-in Self-Test (BIST)

An increasing number of logic designs are implementing logic built-in self-test (LBIST). LBIST dramatically reduces the ATE requirements and has the opportunity to reduce product development times, provide higher speed test capability and enable the devices to be retested at the system level. As design reuse becomes more common, LBIST is an attractive opportunity to simplify the testing problem. The most common LBIST implementation is called STUMPS [36] where the random pattern generators and signature analyzes are added to full scan designs. In addition to LBIST, many designs incorporate some form of hardware-assisted test compression, where BIST-like vector generation and compaction is used to augment the bit stream provided by an ATE, enabling reduced test time, and/or more patterns to be applied in a given amount of test time. Test compression is discussed in more detail later in this chapter. Special test modes to assist diagnosis are often available in BIST and BIST-like methods.

Memory Testing

Historically, stand-alone memory ICs have been tested by applying a sequence of regular patterns to ensure that each of the cells can store and provide data. Various sequences of patterns are applied to ensure that the memory does not have any address or area sensitivities. Common test pattern types include marching 1s and 0s, gallup and neighborhood patterns [37]-[39]. Figure 6 shows a couple of simple examples of memroy test patterns. Memory test patterns are developed to verify address decoders (unique address generation), memory cell data retention, and absence of cross coupling between adjacent memory cells. Most memory architectures require an analog read-out of the memory cells which is achieved by a sense amplifier that may have unique test requirements. Normally, memory test patterns are applied at the full rated speed of the device.

DRAMs and Flash devices typically have more subtle sensitivities compared to SRAMs, so a wider variety of patterns and conditions may be applied to them. In addition, DRAMs typically require a larger effort to develop tests that cover all possible defective circuit behaviors.

For manufacturing test, memory ICs are tested in parallel (6 to 32 devices at a time) to reduce test costs. Normally SRAMs and DRAMs require an additional processing step of being exercised at high voltage and high temperature (burn-in). DRAMs (and some SRAM, particularly those containing large numbers of individual bits) normally contain "redundancy," or spare cells that can be used to replace defective cells. Bitmapping is widely applied to facilitate yield, test and reliability defect learning. Bitmapping is collecting all information about working/failing cells for the entire memory. Each of these steps is more difficult when the memory is implemented not as a stand-alone chip, but as a macro embedded within a logic chip.

Embedded Memory Testing

Embedded memory as registers or register arrays has long been part of designs. However, the size of these register arrays now along with use of SRAM memory (e. g., on-chip caches for CPUs) has increased. Today's designs may also include embedded DRAM or Flash.

It is fairly straight-forward and cost effective to use an ATE to test stand-alone memory ICs. For memories embedded on logic chips, built-in self-test (BIST) is the preferred alternative. Because of the regular nature of memories, on-chip BIST capability is feasible. The BIST engine generates the addresses and data-in signals and verifies the data-out signals. Test patterns similar to the ones applied for stand-alone memories can be applied. Programmability is often included to enable newly developed algorithms to be applied after release to manufacturing. Another capability that the BIST engine must support is wait periods that are needed between writing and reading data to verify data retention times. The BIST engine itself must be fully tested before memory testing can start.

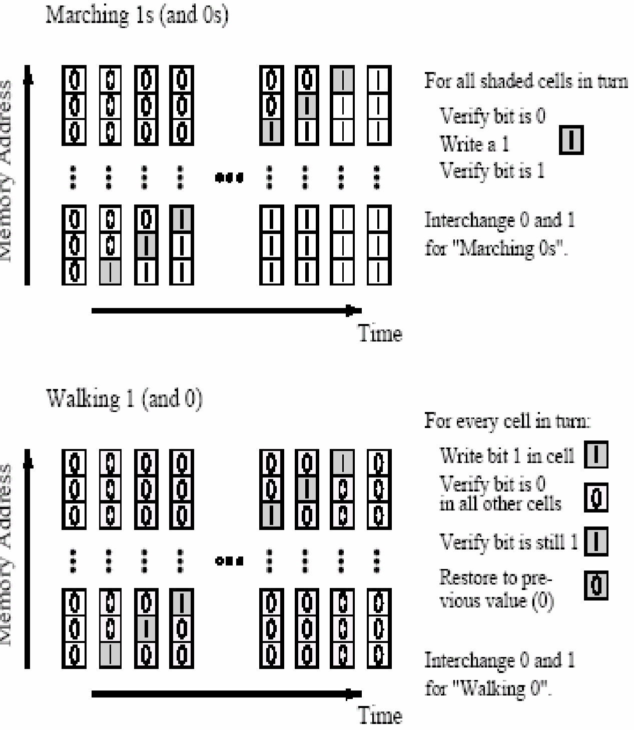


Figure 6. Memory Test Patterns.

Thorough testing and characterization of the embedded DRAM test is particularly important since DRAM requires unique process steps (e.g., capacitors in the memory cells) compared with digital logic circuits. During technology bring-up and yield ramping, the embedded DRAM may be the only vehicle to verify these additional steps. Yield-ramping requires an efficient approach for bitmapping using the BIST engine.

Analog circuit testing

Analog macros (e.g., phased-locked loops, digital-to-analog converters, analog-to-digital converters, sense amplifiers) are typically tested in a significantly different way from digital circuits. Instead of checking boolean responses, analog test methods are normally functional tests of parametric behavior such as voltage measurements, linearity, frequency measurements, and jitter.

Because of the difference in test methodologies compared with logic circuits, ATE for analog circuits and the test development process is significantly different. ATE must have highly precise and accurate parametric and frequency instruments that are available to a number of pins. These instruments are much more expensive compared with simpler, digital ATE. Issues like power supply noise and precise parametric measurements are much more significant for analog circuits than for digital circuits. Because of the unique functional test requirements, normally test development for mixed-signal products is not automated. Thus, test development effort is much larger and development times are longer for mixed-signal products. Also, capabilities such as automated fault diagnostics are normally not available.

System-on-a-chip (SOC) Testing

There has been much attention paid to the industrial trend toward systems-on-a-chip (SOC). The systems-on-a-chip concept suggests that functions that had previously been on independent dies are placed on a single die. SOC devices may include many circuit types including digital, memory and analog. With the recent development of system-in-package or 3D-IC technologies, SOC devices may also contain multiple process technologies--e.g., CMOS and Flash die in the same IC package.

SOC devices pose unique testing challenges. As unique circuit types (called embedded macros) are brought together on a single chip, it may not be possible to apply the same test methods that had been used to test the stand-alone chip. For example, the input and outputs (I/Os) of the macro may not be accessible from the chip I/Os. Inaccessibility of macro I/Os is a particular problem for testing analog circuitry. Standalone macro ("IP") providers need to provide tests for their designs as a key deliverable.

The test economics for SOC devices are challenging. As multiple macros are placed on a single IC, the cost to build the single die remains approximately constant (assuming the same die size and technology scaling where defect densities improve so that yield remains constant for a given die size). The test cost, however, may go up significantly if each embedded macro must be tested serially using conventional techniques. The testing-related costs may also go up if multiple, expensive testers are required for each individual macro. There are other technical issues associated with SOC devices such as embedded DRAM bitmapping, redundancy reconfiguration and mixed signal noise, which remain challenging.

Automatic test equipment

Automatic test equipment (ATE) can be a large part of the cost of testing ICs. Typical testing costs are $3K to $8K per pin for ATE which results in total capital costs of $700K to $5M per tester. The cost drivers of ATE are the number of pins, the maximum frequency, power delivery, pattern buffer storage and timing flexibility options.

Minimizing the cost of ATE is a high priority of semiconductor suppliers. This is done by optimizing manufacturing efficiency (e.g., parallel test) and reducing the requirements of the ATE cost drivers previously listed. For example, some test methodologies enable ICs to be tested using a reduced pincount interface. In general, structural test methods provide a better opportunity to exploit reduced ATE requirements than functional testing. Wider use of methodologies that significantly reduce test costs is expected to be a key trend in the next few years.

Reliability Testing

Some defects such as gate oxide shorts and high-resistance metal shorts may not disturb circuit functionality initially, but may later cause the IC to fail. Tests that look for indicators of defects that do not cause logical misbehavior are one way to find suchs defects. IDDQ testing, discussed earlier, is an example of such a test. Another way to find such defects is to apply stresses that accelerate the defects to the point of failure. Acceleration is accomplished through elevated voltage and/or temperature.

Elevated voltage stresses can be applied at wafer or package test. Operational tests (functional or structural) can be run and monitored at high voltages, or the chips can be put into a high-voltage state and then subsequently tested under nominal voltage conditions. Chang and McCluskey [40] discuss a method for applying high-voltage stressing for a short duration at wafer test.

Burn-in is a step aimed at inducing early life failures to become hard failures by applying high temperature and high voltage to ICs. Normally burn-in is an extra processing step where packaged ICs are placed in burn-in ovens for a significant length of time (e.g., 48 hours). Products may be "baked" without power applied to the chip and then re-tested. More rigorous burn-in tests involve applying power and operating the chip while it is in the burn-in chamber. Outputs may be monitored during burn-in for correct circuit operation or, more commonly, the circuit undergoes the stress voltage and temperature, and then is re-tested afterward. Burn-in is expensive both because it is an extra test step and because there are significant costs associated with supplying the power necessary to operate the chip at elevated voltage and temperature. Some circuits, e.g. SRAMs, require significant design modification to handle burn-in environments in modern processes. For some products, burn-in may be replaced by high-voltage stressing and I_{DDQ} testing.

SOLVING EMERGING TEST PROBLEMS

Trends discussed earlier in this chapter have driven the development of new focus areas in test. This section discusses several of those areas, including test data volume containment, test power containment and new methods of defect-based test.

Test Data Volume and Embedded Compression

The sheer number of transistors to be tested in today's circuits leads to a huge increase in the number of patterns that need to be applied. To mitigate the increase in test time and test data volume, "test data compression" methods are used. Figure 7 shows a the general architecture of test data compression for a full scan design. These methods generate input data using on-chip decompression and/or on-chip pattern generation. The decompression can be as simple as fanning out the input from a scan-in pin to multiple scan chains internally, or use more complex schemes that decode ATPG-based test patterns into appropriate values for each of the bits in the scan chains on the chip. On-chip test pattern generation typically uses pseudo-random pattern generators, which can be implemented on chip with a small area overhead. Output responses are compressed using multiple-input shift registers or XOR networks [41] - [45] . Typical test data compression rate today are 20X-80X and are quickly increasing to over 100X.

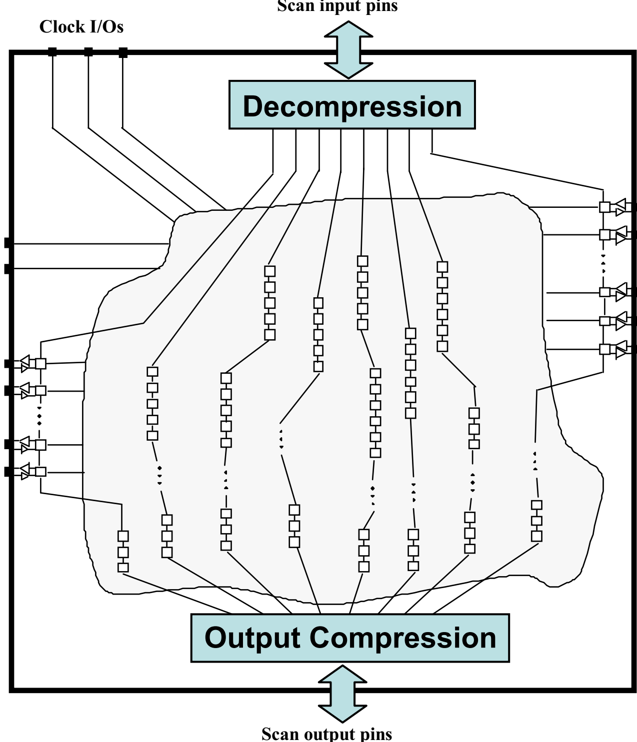


Figure 7. Test Data Compression Architecture.

Test Power

Today's large chips can easily draw 10's of Watts of power during functional operation. At test, the power draw can be exaggerated by the high switching of ATPG test patterns factors (compared to switching factors for functional operation, which are typically in the 10%-20% range). Power draw can be especially high during scan-based testing, as bits are shifted through the scan chain, causing transitions within the logic driven by the chain latch/flip-flow outputs. An active area of research in test is how to minimize scan shift power. Several promising approaches have been suggested that carefully choose the values of "don't-care" bits in the test pattens to minimize the number of 0-1 or 0-1 transistions that in turn consume active power. The choice of don't-care bits may, for example, try to put as many 0's or 1's in a row as possible. Keeping test power low is important because it is difficult to supply high current to the chip, especially during wafer test, and because of the cost of power supplies, meachanisms for delivering power the the chip and hardware for thermal control. In addition, power droop and, potentially, temperature elevation during that that exceeds what would exist during normal chip operation may lead to false fails, and in turn to yield loss.

Defect-based Test & Statistical Testing

More complex fault models and test generation methods are becoming more common in the industry. For example, some companies are targeting bridging faults that are extracted from the physical layout [46] . Others have proposed generating N-detect patterns wherein each stuck-fault is tested multiple times (in different ways) to improve the coverage of unmodeled defects [47] -[48] . Improved coverage can also be achieved by combining N-detect pattern generation with physically adjacent areas [49] .

Another ermerging trend is the use of statistical methods to optimized the testing process [21] -[24] , [50] -[52] . For example, parametric data can be collected during wafer test without pass/fail limits – and then a post-processing step is used to perform data analysis and to identify outliers [21] -[24] . These outliers can either be rejected or they could be subject to additional testing [50] . Automated real-time statistical analysis can enable the testing process to tolerate and compensate for increased process variability.

Statistical testing methods can be extended to perform "Adaptive Testing" where the testing, test content or test limits can be automatically changed in real-time. Adaptive testing also includes data feed-forward from one step to another and dynamic test routings. (See more background of Adaptive testing and examples at [1] .)

CONCLUSIONS

Test is a critical step in IC production. At each level of assembly of a saleable product (wafer, package, board, system) it becomes increasingly costly to find problems. The highest cost is incurred in customer-satisfaction if a bad chip is allowed to escape. This chapter has given a basic overview of microprocessor and SOC integrated circuit testing. IC testing is

being strongly impacted by semiconductor industry trends toward greater complexity, speed and integration. These trends are forcing changes in test methodologies that promise to continue to allow IC manufacturers to supply correct, robus and reliable products to the marketplace.

ACKNOWLEDGEMENTS

The authors wish to thank Wojciech Maly for providing the motivation for exploring many of these technical issues. We would also like to acknowledge Brian Kessler for providing information related to embedded memory testing.

REFERENCES

[1] The International Technology Roadmap for Semiconductors, 2010 Edition. Semiconductor Industry Association. (www.itrs.net)

[2] C. Hawkins and J. Soden, "Electrical Characteristics and Testing Considerations for Gate Oxide Shorts in CMOS ICs," *Int. Test Conf.,* Nov. 1985, pp. 544-555.

[3] C. Hawkins and J. Soden, "Reliability and Electrical Properties of Gate Oxide Shorts in CMOS ICs," *Int. Test Conf.,* Sept. 1986, pp. 443-451.

[4] L. Horning, J. Soden, R. Fritzemeier and C. Hawkins, "Measurements of Quiescent Power Supply current for CMOS ICs in Production Testing,", *Int. Test Conf.,* 1987, pp. 118-127.

[5] W. Maly, P. Nag and P. Nigh, "Testing Oriented Analysis of CMOS ICs with Opens," *Int. Conf. on Computer-Aided Design*, 1998, pp. 344-347.

[6] J. Soden, R. Treece, M. Taylor and C. Hawkins, "CMOS IC Stuck-open Fault Electrical Effects and Design Considerations*," Int. Test Conf.,* Aug. 1989, pp. 423-430.

[7] F. Ferguson, M. Taylor and T. Larrabee, "Testing for Parametric Faults in Static CMOS Circuits," *Int. Test Conf.,* 1990, pp. 436- 443.

[8] T. Storey and W. Maly, "CMOS Bridging Fault Detection," *Int. Test Conf.,* 1990, pp. 842-851.

[9] H. Hao and E. McCluskey, "'Resistive Shorts Within CMOS Gates," *Int. Test Conf.,* 1991, pp. 292-301.

[10] R. Rodriquez-Montanes, J. Segura, V. Champac, J. Figueras, J. Rubio, "Current vs. Logic Testing of Gate Oxide Short, Floating Gate and Bridging Failures in CMOS*," Int. Test Conf.,* 1991, pp. 510-519.

[11] C. Henderson, J. Soden and C. Hawkins, "The behavior and Testing Implications of CMOS IC Logic Gate Open Circuits*,"* Proc ITC., 1991, pp. 302-310.

[12] P. Maxwell, R. Aitken, V. Johansen and I. Chiang, "The Effectiveness of I_{DDQ}, Functional and Scan Tests: How Many Fault Coverages Do We Need?" *Int. Test Conf.,* 1992, pp. 168-177.

[13] J. Segura, V. Champac, R. Rodrigues-Montanes, J. Figueras and J. Rubio, "Quiescent Current Analysis and Experimentation of Defective CMOS Circuits*," Journal of Electronic Testing: Theory and Applications (JETTA),* Vol. 2, No. 4, Nov. 1992.

[14] H. Vierhaus, W. Meyer and U. Glaser, "CMOS Bridges and Resistive Transistor Faults: Iddq versus Delay Effects*," Int. Test Conf.,* 1993, pp. 83-91.

[15] H. Hao and E. McCluskey, "Analysis of Gate Oxide Shorts in CMOS Circuits," *IEEE Trans. of Computers,* Vol. 42, No. 12, December 1993, pp. 1510-1516.

[16] C. Hawkins, J. Soden, A. Righter and F. Ferguson, "Defect Classes -- an Overdue Paradigm for CMOS IC Testing," *Int. Test Conf.,* 1994, pp. 413-425.

[17] K. Wallquist, "On the Effectiveness of ISSq Testing in Reducing Early Failure Rate," *Int. Test Conf.,* Oct. 1995, pp. 910-915.

[18] S. Ma, P. Franco and E. McCluskey, "An Experimental Chip to Evaluate Test Techniques: Experimental Results," *Int. Test Conf.,* Oct. 1995, pp. 653-662.

[19] A. Gattiker and W. Maly, "Current Signatures: Application," International Test Conference, 1997, p.156-165

[20] P. Maxwell, P. O'Neill, R. Aitken, R. Dudley, N. Jaarsma, M. Quach, D. Wiseman, "Current Ratios: A Self-Scaling Technique for Production IDDQ Testing," IEEE International Test Conference, 1999, p.738.

[21] W. Daasch, J. McNames, D. Bockelman, and K. Cota, "Variance Reduction Using Wafer Patterns in IDDQ Data", International Test Conference, 2000, p.189.

[22] S. Sabade, D. Walker, "Improved Wafer-level Spatial Analysis for IDDQ Limit Setting, International Test Conference. 2001, p.82-91.

[23] W. Daasch, K. Cota, J. McNames and R. Madge, "Neighbor Selection for Variance Reduction in IDDQ and Other Parametric Data," International Test Conference, 2001, p.92-100.

[24] R. Madge, M. Rehani, K. Cota, W. Daasch, "Statistical Post-Processing at Wafersort - An Alternative to Burn-in and a Manufacturable Solution to Test Limit Setting for Sub-micron Technologies," IEEE VLSI Test Symposium, 2002, p.69

[25] T. Williams, R. Dennard, R. Kapur, M. Mercer and W. Maly, "IDDq Test: Sensitivity Analysis of Scaling," *Int. Test Conf.,* Nov. 1996, pp. 786-792.

[26] P. Gillis, et al., "Delay Test of Chip I/Os using LSSD and Boundary Scan", Proc ITC., 1998, pp. 83-90.

[27] M. Tripp, T. Mak and A. Meixner, "Elimination of Traditional Testing of Interface Timings at Intel*," Int. Test Conf.,* 2003, pp. 1014-1022.

[28] M. Abramovici, M. Breuer and A. Friedman, *Digital Systems Testing and Testable Design*, IEEE, 1998.

[29] E. Eichelberger and T. Williams, "A Logic Design Structure for LSI Testability," *Design Automation Conf,* 1978, pp. 165-178.

[30] B. Koenemann et al., "Delay Test: The Next Frontier for LSSD Test Systems," *Int. Test Conf.,* Sept. 1992, pp. 578-587.

[31] O. Bula et al., "Gross Delay Defect Evaluation for a CMOS Logic Design System Product," *IBM J. Res. Develop.*, Vol. 34 (No. 2/3) 1990, pp. 325-338.

[32] P. Maxwell, R. Aitken, K. Kollitz and A. Brown, IDDq and AC Scan: The War Against Unmodeled Defects, *Int. Test Conf.,* pp. 250-258, Oct. 1996.

[33] Special Issue on Delay Testing, *IEEE Design & Test of Computers*, Vol. 20, Issue 5, Sept.-Oct., 2003.

[34] S. Sengupta et al., "Defect-Based Test: A Key Enabler for Successful Migration to Structural Test," *Intel Technology Journal*, Q1, 1999, pp. 1-14.

[35] R. Bassett et al., "Boundary-Scan Design Principles for Efficient LSSD ASIC Testing," *IBM J. Res. Develop.*, Vol.

34 (No. 2/3) 1990, pp. 339-354.
[36] P. Bardell and W. McAnney, "Self-Testing of Multichip Logic Modules," *Int. Test Conf.,* Nov. 1982, pp. 283-288.
[37] R. Feugate and S. McIntyre, *Introduction to VLSI Testing*, Prentice Hall, 1988.
[38] A. Van de Goor, *Testing Semiconductor Memories*, John Wiley & Sons, 1991.
[39] A. Sharma, *Semiconductor Memories: Technology, Testing and Reliability*, IEEE, 1997.
[40] J. Chang and E. McCluskey, "SHOrt Voltage Elevation (SHOVE) Test for Weak CMOS ICs," *VLSI Test Symp.,* Oct. 1997, pp. 446-451.
[41] H. Hamzaoglu and J. Patel, "Reducing Test Application Time for Full-Scan Embedded Cores," Int. Symp. On Fault-Tolerant Computing, June 1999, pp. 260-267.
[42] A. Pandey and J. Patel, "Reconfiguration Technique for Reducing Test Time and Test Volume in Illinois Scan Architecture Based Designs," VLSI Test Symp., 2002, pp. 9–15.
[43] B. Keller et. al., "OPMISR: The Foundation for compressed ATPG Patterns", Proc. ITC, pp. 748-757, 2001.
[44] J. Rajski et. al., "Embedded Deterministic Test for Low Cost Manufacturing Test", Proc. ITC, pp. 301-310, 2002.
[45] S. Mitra and Kee Sup Kim, "X-Compact: An Efficient Response Compaction Technique for Test Cost Reduction", Proc. ITC, pp. 311-320, 2002.
[46] E. Tran et. al., "Silicon Evaluation of Logic Proximity Bridge Patterns", Proc. of VLSI Test Symposium, pp. 78-85, 2006.
[47] S. Ma, P. Franco and E. McCluskey, "An Experimental Test Chip to Evaluate Test Techniques Experimental Results," Proc. ITC, 1995, pp. 663-672.
[48] I. Pomeranz and S. Reddy, "On n-detection Test Sequences for Synchronous Sequential Circuits," VLSI Test Symposium, 1997, pp. 336-342.
[49] J. Nelson, J. Brown, R. Desineni, R. Blanton, "Multiple-Detect ATPG Based on Physical Neighborhoods," Conference on DesignAautomation, 2006.
[50] R. Richmond, "Sucessful Implementation of Structured Testing", Proc. ITC, pp. 344-348, 2000.
[51] R. Miller, "Good Die in Bad Neighborhoods", Proc. ITC, p. 1128, 2000.
[52] R. Edmondson et. al., "Optimizing the Cost of Test at Intel Using per Device Data", Proc. ITC, pp. 1-8, 2006.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Diagnosis of Scan Logic and Diagnosis Driven Failure Analysis

Srikanth Venkataraman
Intel Corporation
Hillsboro, Oregon, USA

Martin Keim and Geir Eide
Mentor Graphics Corporation
Wilsonville, Oregon, USA

Abstract

In this article we explore the world of diagnosis of digital semiconductors devices. After shortly outlining the technology behind diagnosis, the main part of this article describes key improvements to the basic diagnosis tools we knew from previous years. These improvements enable the failure analysis engineer as well as the yield analysis engineer to use diagnosis as a powerful software instrument supplementing his portfolio of tools. Throughout this article, we summarize several successful industrial applications of diagnosis. For example we recap applications where diagnoses helped reducing the search area for defect down to 3% of the original search area size; and in another case, diagnosis was used to improve at-speed tests, increasing the speed by 300MHz, plus several other cases.

1. Introduction

Scan diagnosis is an established technique identifying and localizing defects in digital semiconductor devices that fail manufacturing test. Traditionally, scan diagnosis tools relied on the stuck-at fault model and could diagnose down to a logic net in the design. Although this approach is useful for localization, it has some limitations. Stuck-at patterns typically detect a vast variety of defect types, including bridges and opens, but this fault model is not always sufficient for effective diagnosis, since an actual defect may not behave as a stuck-at fault. This typically results in a large number of candidates, typically referred to as suspects. Even with a single suspect net, that net could cover a large area of the die. Early literature references are [1-3].

Over the last few years, these limitations have been addressed through technologies such as

- layout-aware diagnosis [4, 5],
- cell-internal diagnosis [6-9],
- scan chain diagnosis [10-20],
- at-speed diagnosis [21], and
- iterative diagnosis [22].

An advanced tool is capable of diagnosing defects down to a logic net segment and physical polygon, and it can differentiate between defects internal to the cells vs. defects in the interconnect. Newer tools can also accurately classify a wide range of defect types. In addition, iterative diagnosis can be used to increase the initial diagnosis resolution.

Diagnosis can now be used in conjunction with

- on-chip compression [23-25] and
- logic built-in-self-test (BIST) [26].

These advances have significantly improved the value of diagnosis for failure analysis. Current diagnosis tools produce results directly applicable to the failure analysis task at hand. They represent the results for example in terms of physical location, including the x/y/layer coordinates of a possible defect location. They also consume production test patterns even for designs using embedded pattern compression techniques, no longer requiring special 'diagnosis only' test patterns bypassing the compression circuitry. In the remainder of this article, we will review many of advances in more detail and discuss their merits for the failure analysis engineer.

Beyond the immediate needs of the failure analysis engineer, diagnosis advances have also paved the way for using this technology in other areas like yield

loss analysis and subsequent yield improvements [4, 27-29] or test quality improvements [21, 30]. ([28] can be found in this edition of the desk reference manual.) The key here is the low cost of diagnosis and the significantly faster time to root cause compared to traditional yield improvement methods.

In the following, we use terms and methods from scan testing. Please see test overview article "IC Testing: Background and Directions" [31] of this edition of the desk reference manual for an explanation of the terms and methods.

2. How Diagnosis Works

Scan diagnosis leverages some of the same technology that is used in automatic test pattern generation (ATPG). Most scan diagnosis tools can only diagnose failures in ATPG or BIST patterns that leverage internal scan chains or scan-based DFT such as logic BIST. A typical diagnosis system requires a gate level representation of the design, ATPG test patterns, and fail data from the tester. In addition to scan ATPG tests, functional tests may be used for screening manufacturing defects. A snap-shot of observable scan cells or scan-dump can be captured at a give cycle of a functional test, and analysis of faults around the window of failing cycles can be used effectively to diagnosis silicon defects [32].

Each failing pin/cycle on the tester corresponds to a failure captured in a scan cell for a particular scan pattern. The first step in the diagnosis process is to trace the design backwards from the failing scan cell then to identify which suspects that could cause these failures. A suspect is typically a particular fault type (such as stuck-at, bridge, or open) in a particular location (net). An example suspect would be "stuck-at-0 fault on the A input of the OR gate /design/module/inst023." The suspect lists are then compared across multiple failures (multiple scan patterns, multiple scan cells), and this process helps narrow down the suspect list. Simulation is also performed to see if any of the suspects would cause any additional failures other than those observed on the tester.

Suspects are then ranked and scored based on how well the suspect explains the failures. For instance, suspect /design/module/inst023/A stuck-at-0 might explain 8 out of the 10 failures on the testers, and would cause 3 additional patterns to fail. This suspect would receive a lower score than a suspect that explains all the 10 failures and would not cause any additional failures. Since the suspects are based on fault models and not the actual defects, it is common to observe less-than-perfect diagnosis results.

The two most common metrics in measuring diagnosis results are accuracy and resolution. Accuracy is a measurement of how well the diagnosis tool has identified the true defective net(s) and the true defect type. For a given list of candidates, accuracy is a binary decision: Is the true candidate on the list, yes or no. Resolution measures the area of the defective location captured by a suspect, that is, the bounding box of a defect. For logic-based diagnosis tools, the resolution equals the combined area of all reported nets. Logic-based diagnosis tools do not know where two nets potentially bridge or where an open on a net might be located; thus, the whole net must be searched to find the defect. Layout-aware diagnosis with its access to the polygon level easily finds and reports, for example, the bridge location in x/y/layer terms, or the few polygons on a net, where an open defect might be.

3. A Typical Diagnosis Flow

A typical diagnosis flow requires three sets of inputs: A gate level description (i.e. model) of the design, the test stimuli applied on the tester and the failures recorded by the tester (a.k.a. datalogs) when these stimuli are applied.

Diagnosis flow starts by processing the datalogs. The datalogs record failures in terms of pin names and test cycle numbers which are dependent on the specific tester that is being used. By analyzing the DFT architecture of the design, these failures are mapped to scan operation ID numbers that have failed, and particular scan cells or primary output

signals in the design where the faulty behavior was observed. In the end, a list of failures in terms of test stimuli and design signals, which are independent of the tester architecture, is obtained.

Diagnosis execution starts by consuming the list of failures produced by datalog processing step. The tool identifies the cone of logic that supports the observed failures and reasons that the defect location must be contained in that cone. It enumerates candidate locations, simulates each candidate using the test stimuli and ranks them by comparing the simulated behavior to actual faulty behavior. In the end, a list of candidate locations is reported by decreasing likelihood of containing the actual defect.

4. Making Diagnosis Work in your Workflow

In this section we described the various requirements and other considerations that typically need to be taken into account to set up a full working scan diagnosis system.

Scan test content is applied on the tester along with the other manufacturing test content which includes cache tests, functional tests, IO tests, and parametric tests. Failures can be captured in-line during production tests by setting up a fail flow which is exercised on failing parts (dice at wafer sort or packaged parts). However, in order to minimize overall test time budgets, the increase in test-time due to fail data collection needs to be limited relative to the total test-time. A test flow to collect fail data is illustrated in Figure 1. The scan test content includes a scan-system test to determine if scan-system is functional. After the scan-system test, scan chain shift tests and scan ATPG capture tests are applied.

Usually just the production tests can be used for fail data collection. In addition some other tests are usually added with negligible impact to tester memory. In addition to the traditional (1100) repeat scan chain pattern which detects stuck at and all transition fault types, some additional chain tests are added. If the transition defects cause delays longer than one bit, then shift out value of this traditional chain test pattern will be a constant 0 or constant 1, which will be incorrectly interpreted as stuck-at-0 or stuck-at-1. To avoid this pitfall, (0) repeat and (1) repeat chain patterns are also included. These two patterns will not cause any failures for transition type defects. Also a slow changing pattern, 16 zeros followed by 16 ones repeat pattern are added. This pattern can detect large delay defects. Table 1 shows these patterns and the type of defects they detect including stuck-at (SA), slow-to-rise/fall (STR/F) and fast-to-rise/fall (FTR/F).

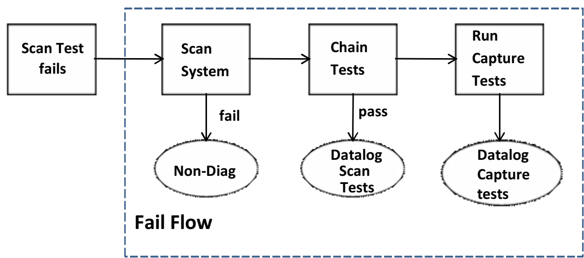


Figure 1: Failure Bucketing of Scan Tests

Pattern	Detects
(1100)R	SA0, SA1, Single bit delay (STR, STF, STRF, FTR, FTF, FTRF)
(0)R	SA1
(1)R	SA0
(16x1+16x0)R	SA0, SA1, Multiple bit delay (STR, STF, STRF, FTR, FTF, FTRF)

Table 1: Scan Chain Tests

Since failing data collection requires additional test time, a trade-off needs to be done to ensure adequate data collection while minimizing costs. To maintain diagnosis quality and resolution, we need to collect adequate failure information from failing scan test. However, too many failures collected can result in wasted test time with minimal additional benefit.

The data logging the first 100 failing scan operations are usually adequate for diagnosis of logic failures. Diagnosis of scan chain defects require capture tests

as well as shift tests. The number of failures from capture tests in this case is usually much higher than in the case of logic defects since scan load and unload operations are corrupted by the defect.

If we denote the length of a scan chain by L, then on average we expect to see about L/2 failures per scan-unload due to a defect on that chain. About 10-20 scan unload operations are usually needed to obtain good diagnosis resolution for scan chain defects.

An additional strategy to reduce fail data volume and the test time impact was to use sampling. In the initial phase of the yield learning when volume of material being processed was low, all failing dies on all wafers in a lot were analyzed. Later in the middle phase with volume picking up, fewer wafers are sampled and data logged.

5. Improving FA Success Rate

Early diagnosis tools were driven from the test point of view. There, the stuck-at model is very successful in detecting many different defects. For example, it is well known that low-resistive bridges can be detected with high confidence by stuck-at faults, in particular if each fault is targeted multiple times. The inverse is not true however. A bridge defect cannot be explained by a stuck-at fault model, since most of the bridges do not behave like a stuck-at fault. As a consequence, the reported diagnosis suspects might miss the defect entirely or report suspects with low scoring. Both of which is not acceptable for failure analysis, in particular for cases of field returns, where there's only one chance to find the defect.

Tool, know your defects better

Current diagnosis tools do understand the concepts of bridges and opens, speed related defects, as well as the difference between defects internal to the cells and in the interconnect. They no longer try to match everything to slow-speed stuck-at faults. In the following, we visit each of these concepts.

The case of opens

On logic level, diagnosis algorithms can figure out that an observed defective behavior of the die does not match a stuck-at behavior. The question is then, what kind of (static) defect could explain the observations? The only other two explanations are an unintended connection of a net to a neighboring net or an unintended disconnection of a net, i.e. a bridge or an open. The actual, physical reason for a short or the bridge is not relevant for the diagnosis tool. For example, as long as two nets have a (low-resistive) electrical connections of whatever physical cause, a diagnosis tool will classify this defect as a bridge.

For the case of opens, diagnosis usually determines the defective net very precisely. However, it usually cannot tell the open from a bridge based only on the observed failing bits, since both have similar observable behavior at the logic level.

For opens, the tool will report back to the user nothing better than the entire net – and somewhere on it is the open. In some cases, this result might already be sufficient for the failure analysis engineer to proceed for example with probing. In many cases however, a better resolution is desired. In the case of opens, this means reporting to the user, which segments of the net could have the open, and which segments can be assumed defect free.

Intel presented a method [37] to isolate interconnects opens by extending the stuck-at fault model to create a net fault model. It was shown that interconnect opens to could be isolated to single nets with information on portions of the interconnect that contain the defect.

AMD reports in [5] case studies on bridges an opens, where so-called *layout-ware diagnosis* is used to improve the accuracy and resolution. In layout-aware diagnosis the common, logical analysis engine is paired with an integrated layout based engine, which has access to exact routing of each net, each placement of cells, and all other layout components. Through this tight integration, diagnosis can correlate logically determined results against physical (i.e. layout) possibilities. Through this, the tool can quickly eliminate impossible candidates. Further, the layout-aware diagnosis can find

additional suspects derived from layout, which are beyond the reach of traditional, logic-only diagnosis.

Finally, layout-aware diagnosis also gains access to the exact topology of a net, i.e. it knows through which layers it routs, where single and double vias are located, and where the net branches out. Using the net topology, an open segment of a net can be determined, instead of declaring an open somewhere on the net. For nets with branches, these segments are usually significantly smaller than the total net.

In one of AMD's case studies a suspective net can be divided into 14 segments based on its topology. This net runs through three metal layers and spans a maximum of 36µm with a total search area for the suspect of 1192$µm^2$. Layout-aware diagnosis of the device leaves only a single net segment that could contain the open defect. Also, all bridge candidates from an earlier non layout-aware diagnosis run were eliminated. Overall, layout-aware diagnosis reduced the suspect area down to 11% (130$µm^2$), with a maximum span of 12µm.

The case of bridges

In the same study AMD [5] reports layout-aware results for a bridge-type of defect. In this case, the non layout-aware diagnosis reports a possible open or dominant bridge as the defect behavior. Dominant bridge means that the reported net is the victim net of a bridge defect. Victim nets are the ones over which the faulty values are propagated to an observation point, which is a primary output of the die or a scan flip-flop, whereas the aggressor net never carries any faulty values to any observation point. Hence, the aggressor net is not visible for a non layout-aware diagnosis tool. The reported victim net could potentially bridge to any of the neighboring nets at any location. The non layout-aware diagnosis tool cannot provide more detail. Furthermore, it is possible that an open-type defect could be the cause for the observed defective behavior of the device.

For the failure analysis process, this means that not only does the engineer need to look out for two very different types of defects, an open or a bridge, but also knows only one side of the potential bridge. The engineer still has to look for the other net and the defect somewhere along the victim net, potentially forming bridges to any of the neighboring net polygons. In other words, although the diagnosis report contains only one dominant bridge suspect, the number of potential defect locations the failure analysis engineer has to inspect is bounded only by the number of neighboring polygons.

While the logic report is ambiguous between a dominant bridge defect and an open defect, the layout-aware diagnosis report in AMD's case study shows only a bridge type defect. All open candidates were dismissed. Further, layout-aware diagnosis identifies all of the *relevant* aggressor nets. In this particular case, two neighboring nets are found and all other impossible aggressor-victim net combinations are eliminated.

Overall, instead of searching a suspect area of 1012$µm^2$ (the whole net), failure analysis can focus the activities to only the 26$µm^2$, due to the tight bounding boxes computed by layout-aware diagnosis. This is a reduction down to only 2.6% of the original area. This area is divided only between two metal layers, instead of four, in which failure analysis is looking for a bridge-type defect. The total span of the suspect area is reduced from 39µm to 16µm.

Diagnosing at-speed failures

In order to gain the high level of outgoing product quality, it became common practice to execute additional type of tests. Building on top of the stuck-at fault, other fault models are used. In devices manufactured at 130nm and below, it is common to use a structural at-apeed test, typically leveraging the transition fault model (some use path-delay). The title of this section reflects that not all at-speed failures identified during testing, and subsequently diagnosed, are actual, physical defects. The Freescale test case summarized below is a good example of companies using the convenience and performance of at-speed diagnosis for design

optimization, timing analysis, timing debug, and other related tasks.

In [21] Freescale reports the successful application of diagnosis of at-speed test patterns. Interestingly, this paper reports that the root cause for the failure of the dice on the ATE was not a physical defect, but an incomplete description to the ATPG tool of false and multi-cycle paths. Freescale claims that eliminating these paths based on the diagnosis result accumulated in a 300MHz increase in speed. For some observed failures, the root cause was a design problem which needed design changes.

A slightly different aspect of at-speed diagnosis is reported by Infineon in [33]. In this paper, special 'diagnosis friendly' test patterns are used to improve the efficiency of physical fault isolation of timing failures using a time-resolved emissions system. This paper presents a case study, where the identified defect was most likely a resistive via. Analysis of the topology of the net in question further reduced the number of possible defective vias by 40%. Failure analysis eventually identified and verified a resistive via within the identified area.

Cell internal defect or interconnect defect?

At 90 nm and beyond, a significant number of manufacturing defects are inside the library cells (standard cells) themselves. This is in part caused by the increasing use of custom cells designed to deal with higher process variations. The ability to differentiate between defects in the interconnect ("back-end defects") and internal to the cells ("front-end defects") is therefore crucial to ensure that failure analysis can be performed in a timely fashion [6].

Intel [7] presented techniques to distinguish between cell-internal and interconnect failures, and methods to isolate to within a cell using transistor-level information and switch-level simulation.

According to TSMC ([6]), the process improvement comes from two sources. (a) Knowing the defect is in the cell, any de-layering can go straight to the metal layer where the cell interconnects are routed, without the need to investigate any higher metal layer, and (b) having failure analysis investigate fewer layers reduces the overall cost of the examination. Further, being able to distinguish cell-internal defects from interconnect defects also provides useful cell statistics for yield analysis and subsequent improvements of the cell library. Additional references on this topic are [8, 9].

In [6], TSMC outlines at first the methodology how defect locations can be determined to be inside or outside of cells, then presents results based on emulation of defective cell behavior. In the end of the report however, the methodology is put to the test on a 90nm technology, AMD graphics chip. From a set of diagnosis reports, 7 were selected that had a cell as the one of the highest suspects reported. For these dice failure analysis was performed. In all 7 cases, a defect was found in the indicated instance of the cell. TSMC concluded additionally that using the cell-internal diagnosis classification, the overall failure analysis process was faster and less costly, because diagnosis eliminated the need to investigate any layer higher than metal 1.

Defects in scan chains

So far the discussion was about finding the defect in the functional logic. In this section, we focus on defects in the scan chains themselves. It is well known, that the cause of a significant portion of dice failing logic test relates to 'issues' in the scan chains themselves. Literature reports numbers in the 10% to 30% range [20]. Not all of these issues are actual defects in the physical sense. Some issues are simply setup or hold-time violations, i.e. the scan chains are operated incorrectly. Nonetheless, at the moment of testing on the ATE, this is not obvious. Hence, a diagnosis tool for scan chain failures must consider and report all possibilities.

The method of scan chain diagnosis were already discuss in the 1990s [10-13], and improved since, see for example [14-20], with an alternative approach to scan chain diagnosis reported by Intel in [36]. Additionally, it is possible to diagnose errors from scan enables and multiple failing scan chains [15, 16]. Scan chain diagnosis is a very active field of

academic and industrial research. The reader is encouraged to use these literature references as a starting point.

In principle, scan chain integrity testing comprises of shifting in a repetitive sequence of 0s and 1s, typically '0011' and observing the shift-out values at the output of the scan chain. From these observed chain test values, one can conclude defects in the scan chain. To name a few possibilities, if the scan out values are '0000', most likely a stuck-at 0 fault is somewhere on the scan chain. If the scan out values are '0001', a possibility is a 'slow to rise', whereas a '0111' could indicate a 'fast to rise' problem.

There are many different conclusions scan chain diagnosis tools compute nowadays. Reporting the cause is only half of the answer of a scan chain diagnosis tool. The other half of the answer is the location of the problem. Scan chain diagnosis usually reports the location by the ordinance number of the scan cells, counted from the scan output of the chain (the cell closest to the scan out usually has the ordinance number zero). In this sense, a scan chain diagnosis report might say 'slow to rise problem between cells 25 and 21'.

It is the goal of these diagnosis tools to compute the range of possible locations as tight as possible. But they do not always succeed. In the papers listed above, many algorithmic methods are proposed how to tighten the range, some papers present case studies underlining the success of their method.

Overall, diagnosis tools are in good shape when it comes to scan chain failures. One recent key improvement is the inclusion of scan logic tests in addition to the scan chain integrity test. At first it seems odd to execute scan test patterns, knowing that some scan chains are not operational. However, it has been demonstrated that the additional information gained from this is sufficient to considerably improve the quality of scan chain diagnosis. Some methods, like [18, 19], are actually able to diagnose and distinguish a scan logic and a scan chain failure at the same time.

Shortening the list of suspects

Up to now, we assumed that diagnosis uses the production test pattern set. There is good reason to start with this pattern set, since it detected the defect in the first place. Yet, the purpose of a production test pattern set is to detect as many faults as possible with the smallest number of tests. In other words, each test pattern potentially detects many different faults simultaneously. As long as there are several failing patterns, a diagnosis tool can determine the intersection of these sets of possible faults. Usually this reduces the number of reported suspects to a meaningful quantity. In some cases however, there are just too many suspects reported to be meaningful for any failure analysis work.

LSI describes in [22] the concept of *iterative-diagnosis*. In this method, additional test patterns are generated, with the goal of reducing the number of suspects reported by diagnosis, before any (destructive) failure analysis work should be commenced. These new patterns together with the previous ones are brought to the ATE and applied to the device-under-test. The new test response is diagnosed again, and if the number of suspects is still too large, the process repeats. It is an iterative process, which stabilizes after a few cycles.

The analysis performed in [22] concludes that in 41% of the cases this method is successful in reducing the number of suspects. Actual failure analysis case studies reported in this paper underline, according to LSI, the usefulness of iterative diagnosis for both, scan and chain defects.

It should be noted that the principles LSI uses in [22] are applicable to all kinds of diagnosis including at-speed and layout aware. Especially for layout-aware diagnosis, which by itself usually reduces the suspect count, the benefit of iterative diagnosis is in addition to the benefit the user gets from using layout in diagnosis.

Bringing in other data

Correlating the diagnosis result with other data, like inline inspection or layout image overlays has proven

to increase the failure analysis success. Early papers describing successful case studies are [34, 35].

The principle behind this is straight forward. Diagnosis is a very fast tool, generating a list of candidates. Using standard FA equipment with a link to a design representation, the engineer can quickly scan from position to position, looking for indications of a defect. Besides large scale manufacturing issues, like scratches, this method applies more to larger technology nodes, with more-or-less visible defects.

In technology nodes of 90nm and smaller and with the increase of non-visible defects and design-manufacturing interactions causing dice to fail, this simple correlation between an observable indicator like a crack or inline inspection results and the actual defect is no longer valid for the majority of the defects. Still, it is a quick check and should therefore not be dismissed. Sometimes the root cause is as simple as having a mechanically defective package.

6. Designs with Embedded Compression

Many devices use one or another type of "embedded compression" technology. The article [31] in this desk reference manual explains in detail this technology. Here, we summarize only the principles, see Figure 2.

Test patterns are stored on the ATE (Automated Test Equipment) and then applied to the device-under-test. The test response of the device is read from the device and then compared on the ATE to the expected response as usual. On the device however, the test patterns provided by the ATE are fed through some logic, the so-called decompressor, which computes the bits shifted into the scan chains of the device. On the output side, the bits shifted out of the scan chains are first 'compacted' before sent to the ATE for comparison as usual. The benefit of embedded compression are reduced ATE run time and reduced pattern volume to be stored on the ATE. This enables industrial designs to achieve several 100x compression. In slightly simplifying

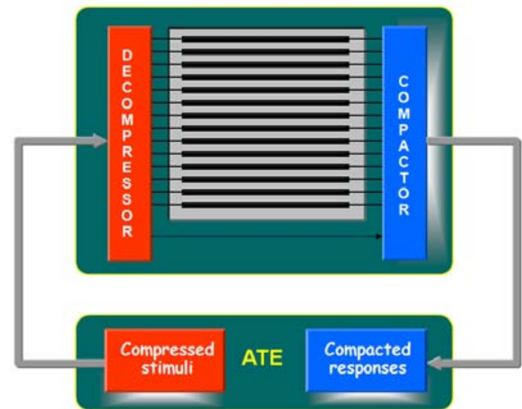


Figure 2: Principle of embedded compression

terms, for every 1 observable bit at the ATE, there are several 100 internal bits (i.e. scan cells) it is derived from, see Figure 3.

Embedded compression changes significantly one key ad-hoc diagnosis methodology, which is the chain intensity plot, or sometimes also called the flop-plot. In such a plot, the simulated values of all scan flops of a scan chain of interest are reported and compared to the actual shift-out values of that particular scan chain. In designs with embedded compression, the shift-out values no longer correspond to any one scan chain or scan cell for that matter. Nevertheless, such a comparison with the compressed outgoing chain bits can still be of value, e.g. to determine the offset of the failing bits.

All commercial embedded compression technologies include one mode of operation in which the on-chip compression logic is bypassed, and the internal scan chains become accessible from the outside again. The flip-side is that all compression is lost, i.e. several magnitudes of additional ATE memory are needed to apply the complete bypass test pattern set. This explosion of pattern volume can be prevented if focusing only on the few failing patterns but, inadvertently trough this, potentially reducing the success of a high quality diagnosis result. Even if all uncompressed patterns are used, still, the operation of the device has changed. Although the patterns are still the same, it is not guaranteed that the defect is still observed.

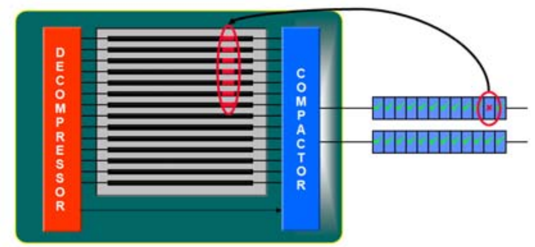


Figure 3: In embedded compression, one observed value relates to many internal scan chain positions.

Intel [24] proposed a compression architecture which can enable reasoning backwards through the compactor to determine failing internal scan cells starting from external compressed channel failures. This enables the usage of diagnosis solutions with no modifications. An alternative is described in [23] and [25], and also used in [6]. In [25] Infineon outlines the technical details how diagnosis of compressed test patterns works. This paper also provides in-depth analysis of this technology, comparing the quality of diagnosis results between the compressed and uncompressed cases. Overall, Infineon's analysis shows that executing diagnosis with compressed or uncompressed patterns has no significant impact on the quality, neither to the resolution nor to the accuracy of diagnosis result.

7. Summary

In this overview article of diagnosis of scan logic and diagnosis driven failure analysis we presented several industrial case studies and successful applications. Table 2 summarizes these cases.

The authors hope that our selection of studies and explanations help the reader to see diagnosis as a powerful tool worth adding to the portfolio of tools any failure analysis engineer should have.

References

1. R.G. "Ben" Bennetts, D.W. Lewin, "Fault Diagnosis of Digital systems – a review," Computer July/August 1971.
2. J. Waicukauski, E. Lindbloom, "Failure Daignosis of Structured VLSI," IEEE Design & Test of Computers, August 1989.
3. Butler, K.M. ; Johnson, K. ; Platt, J. ; Kinra, A. ; Saxena, J., "Automated diagnosis in testing and failure analysis," Design & Test of Computers, IEEE, Jul-Sep 1997.
4. J. Mekkoth et.al, "Yield Learning with Layout-aware Advanced Scan Diagnosis," International Symposium of Testing and Failure Analysis (ISTFA), 2006.
5. Y.-J. Chang, et.al., "Experiences with Layout-Aware Diagnosis," Electronic Device Failure Analysis, May 2010..
6. M. Sharma, W.-T. Cheng, T.-P. Tai, Y.S. Cheng, W. Hsu, L. Chen, S.M. Reddy, A. Mann, "Faster defect localization in nanometer technology based on defective cell diagnosis", IEEE International Test Conference, Oct.21-26 2007.
7. M. E. Amyeen, D. Nayak, S. Venkataraman, "Improving Precision Using Mixed-level Fault Diagnosis, Proc. IEEE Int'l Test Conf., 2006, Paper: 22.3
8. X. Fan, W. Moore, C. Hora, M. Konijnenburg and G. Gronthoud, "A Gate-Level Method for Transistor-Level Bridging Fault Diagnosis," in Proc. VLSI Test Symp., 2006.
9. X. Fan, W. Moore, C. Hora and G. Gronthoud, "A Novel Stuck-at Based Method for Transistor Stuck-Open Fault Diagnosis," in Proc. Intl. Test Conf., 2005, pp. 253-262.

Case Study or Application	Result
Layout-Aware Diagnosis	Search area reduction for open defect to 11% Search area reduction for bridge defect to 3% In both cases, diagnosed type and location confirmed by failure analysis
At-Speed diagnosis	Improve test speed by 300MHz by finding ATPG setup error Identified design marginalities
At-Speed diagnosis	Identified defect combining TRE an diagnosis, reduced number of suspect vias by 40%
Cell-Internal diagnosis	Identified correctly 7 out of 7 cases, failure analysis confirmed defects Significant analysis time improvement, cost saving
Iterative diagnosis	Reduces number of suspects in 41% of cases, improving failure analysis success rate
Embedded compression	Confirmed no impact on resolution or accuracy

Table 2: Summary of presented case studies or applications and their result

10. J. Schafer, F. Policastri and R. Mcnulty, “Partner SRLs for Improved Shift Register Diagnostics", Proc. VSLI Test Symposium, 1992, pp. 198-201.
11. S. Edirisooriya, G. Edirisooriya, “Diagnosis of Scan Path Failures," Proc. VLSI Test Symposium 1995, pp. 250-255.
12. S. Narayanan, A. Das, “An Efficient Scheme to Diagnose Scan Chains," Proc. Int'l Test Conference, 1997, pp. 704-713.
13. Y. Wu, "Diagnosis of Scan Chain Failures," Proc. Int'l Symp. On Defect and Fault Tolerance in VLSI Systems, 1998, pp. 217-222.
14. Chris Eddleman, Nagesh Tamarapalli, Wu-Tung Cheng, “Advanced Scan Diagnosis Based Fault Isolation and Defect Identification for Yield Learning,” International Symposium of Testing and Failure Analysis (ISTFA) 2005, Nov. 6-11, 2005, Page(s): 501-509.
15. K. L. Lee, N.Z. Basturkmen, S. Venkataraman “Diagnosis of Scan Clock Failures” Proceedings of VLSI Test Symposium, pp. 67-72, 2008
16. N.Z. Basturkmen, R. Guo, S. Venkataraman “Diagnosis of Multiple Scan Chain Failures“, Proceedings of European Test Symposium, 2008
17. J. Hwang, D. Kim, N. Seo, E. Lee, W. Choi, Y. Jeong, J. Orbon, S. Cannon, “Deterministic Localization and Analysis of Scan Hold-Time Faults”, International Symp. On Test and Failure Analysis, Nov. 2008, paper 13.3.
18. F. Wang, Y. Hu, H. Li, X. Li, Y. Jing, Y. Huang, “Diagnostic Pattern Generation for Compound Defects”, ITC 2008, paper 14.1.
19. X. Tang, R. Guo, W.-T. Cheng, S.M. Reddy, Y. Huang, “On Improving Diagnostic Test Generation for Scan Chain Failures,” Asian Test Symposium, 2009.
20. R. Guo, S. Venkataraman, “A Technique For Fault Diagnosis of Defects in Scan Chains”, ITC, 2001, pp. 268-277.
21. Nandu Tendolkar; et.al, “Improving Transition Fault Test Pattern Quality through At-Speed Diagnosis,” ITC 2006.
22. K. Gearhardt, C. Schuermyer, R. Guo, “Improving Fault Isolation using Iterative Diagnosis”, 34th International Symposium for Failure Analysis, Nov. 2-6, 2008, Pages 390-391.
23. W.-T. Cheng, K.-H. Tsai, Y. Huang, N. Tamarapalli, and J. Rajski, “Compactor independent direct diagnosis”, Proc. of Asian Test Symp., pp. 15-17, 2004.
24. Z. Stanojevic, R. Guo, S. Mitra, and S. Venkataraman, “Enabling yield analysis with X-compact,” Proc. IEEE Int'l Test Conf., 2005, pp. 734-742.
25. A. Leininger, P. Muhmenthaler, W.-T. Cheng, N. Tamarapalli, W. Yang, K.-H. Tsai, “Compression Mode Diagnosis Enables High Volume Monitoring Diagnosis Flow,” IEEE International Test Conference, Nov. 6-11, 2005, Paper: 7.3.
26. W.-T. Cheng, M. Sharma, T. Rinderknecht, l. Liyang, C. Hill, “Signature based diagnosis for logic BIST”, IEEE International Test Conference, Oct. 21-26, 2007.
27. D. Appello, A. Fudoli, V. Tancorre et al., “Understanding yield losses in logic circuits”, IEEE Design & Test Magazine, May/June issue, 2004, pp 208-215.
28. D. Carder, S. Palosh, R. Raina, “High-Volume Scan Analysis: Methods to avoid Failure Analysis,” Microelectronics Failure Analysis Desk Reference Manual, 6th edition, ASM International
29. H.P. Erb, C. Burmer, A. Leininger, “Yield enhancement through fast statistical scan test analysis for digital logic,” IEEE Advanced Semiconductor Manufacturing Conference and Workshop, April 2005
30. Eichenberger, S.; Geuzebroek, J.; Hora, C.; Kruseman, B.; Majhi, A., “Towards a World Without Test Escapes: The Use of Volume Diagnosis to Improve Test Quality” Test
31. A. Gattiker, P. Nigh, R, Aitken, “An Overview of Integrated Circuit Testing Methods,” Microelectronics Failure Analysis Desk Reference Manual, 6th edition, ASM International
32. M.E. Amyeen, S. Venkataraman, and M. W. Mak, “Microprocessor System Failures Debug and Fault Isolation Methodology,” Proc. IEEE Int'l Test Conf., 2009.
33. C. Burmer, R. Guo, W.-T. Cheng, X. Lin, B. Benware, “Timing Failure Debug using Debug-Friendly Scan Patterns and TRE,” 34th International Symposium for Failure Analysis, Nov. 2-6, 2008, pp. 383-389
34. A. Kinra, A. Mehta, N. Smith, J. Mitchell, F. Valente, “Diagnostic Techniques for the UltraSPARCm Microprocessors,” ITC 1998
35. A. Kinra, H. Balachandran, R. Thomas, J. Carulli “Logic mapping on a microprocessor,” ITC 2000.
36. C.L. Kong, M.R. Islam, “Diagnosis of Multiple Scan Chain Faults,” International Symposium for Failure Analysis, Nov. 6-10, 2005
37. S. Venkatarman and S. B. Drummonds, “A Technique for Logic Fault Diagnosis of Interconnect Open Defects”, in Proc. VTS 2000, pp. 313-318

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Interpretation of Power DMOS Transistor Characteristics Measured With Curve Tracer

Hubert Beermann
Infineon Technologies AG, Munich, Germany

Abstract

For the investigation of discrete DMOS power transistors (Double-Diffused Metal–Oxide–Semiconductor) Curve Tracer (CTR) measurement can give a quick overview of their electrical behavior. At a glance, it can be seen if the chip type inside matches the printing on the package by comparing the device under test (DUT) with a reference. Different failure characteristics deliver first hints on the nature of the defect e. g. which components of the device are involved and how they are affected (short, leakage or threshold voltage shift). Typical characteristics between gate, source, and drain are discussed and how they are related to principle defect locations or to the nature of defects.

1. Introduction

The bipolar transistor is focused in the paper by D. Wilson, Curve Tracer Data Interpretation for Failure Analysis [1]. The behavior of MOSFETs is different to that. Target of this paper is an introduction for measuring power MOSFETs and the interpretation of the curves.

In failure analysis of discrete power MOSFETs it is essential to obtain a quick overview of the behavior of the investigated device. The interpretation of Lissajous figures, characteristics of gate defects, drain-source ON resistance, or the inverse diode can provide useful information for the understanding of the device to be analyzed and its possible malfunction.

1.1 Power DMOS Characterization

As the gate oxide is the most sensitive component of the MOS transistor, every measurement should start with gate-source and gate-drain measurement. This excludes artifacts when performing drain-source tests where higher voltages or energies are used. For a pass or fail measurement an I_{GSS} test can be performed. This works quickly because gate-source and gate-drain can be measured simultaneously. But for failure analysis it is important to see the difference between each circuit and its characteristic behavior. If the gate circuits are not shorted, an I versus V (V-I) curve trace can be measured and displayed as a Lissajous figure. These figures can be used to verify if the chip corresponds to the indicated type on the marking. Furthermore they can tell if the threshold voltage is shifted or if a drain-source leakage or a short can be expected. Hints of gate-source leakages in the low µA range can be seen; however, these leakages have to be measured in DC mode or e. g. by a parameter analyzer. In case of high leakages it is important to distinguish between the direct leakage path (through the damaged components) and the collateral path (between damaged and not damaged components). Examples are shown below.

For a better comparability of the following characteristics the same device type was used for all curves. Exceptions are indicated.

2. Interpretation of Lissajous Figures

The Lissajous Figures which are obtained by the curve tracer are capacitive V-I curves with a 50Hz sinus voltage applied to a gate circuit of the device. The capacitances between the gate-source (G/S)

and gate-drain (G/D) circuits are different and some vary depending on the applied voltage (variation of space charge regions). A change of capacitance also occurs when drain and source are temporarily connected by switching ON the channel. This happens while the gate voltage reaches the threshold voltage level. This capacitance change results in a current peak visible in the Lissajous figure. This effect can be used for comparing threshold levels of different or altered devices. It gives an approximate but not exact value of the $V_{GS\,(th)}$ given in the data book specification because the measuring conditions are different. A typical test condition according to the data book for $V_{GS\,(th)}$ is $V_{GS}=V_{DS}$, $I_D=1mA$. A peak in Lissajous figure is already visible if the channel current (I_D) is in the low µA range.

Figure 1 and Figure 2 illustrate in principle the components and the capacities which are in the focus of the following discussion of the Lissajous figures. The circuit diagram is an approximation since a large number of individual transistor cells are connected in parallel on a chip. Distributed capacitances are therefore involved, and these vary, for the most part, as a function of the drain-source voltage [2].

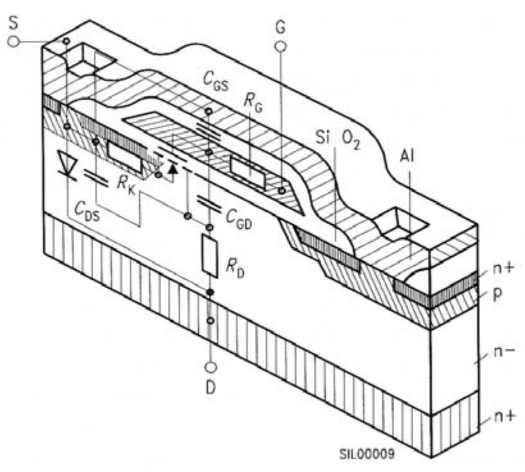


Figure 1: Sectional drawing of an N channel MOS transistor cell showing the conductances of the equivalent circuit diagram (source: Infineon data book).

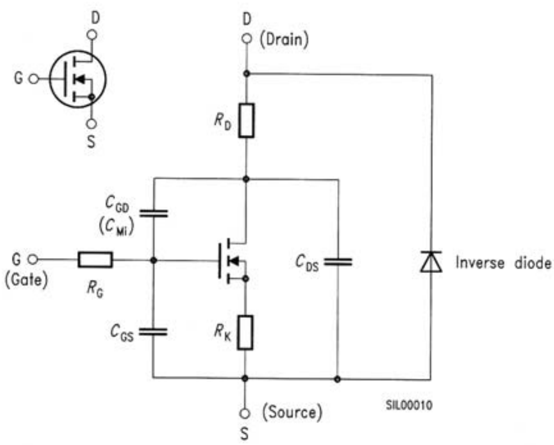


Figure 2: Simplified equivalent circuit diagram of an N channel MOS transistor and the graphical symbol (source: Infineon data book).

In the following AC V-I curve traces of an error free device are measured between G/S and G/D and displayed as Lissajous figures.

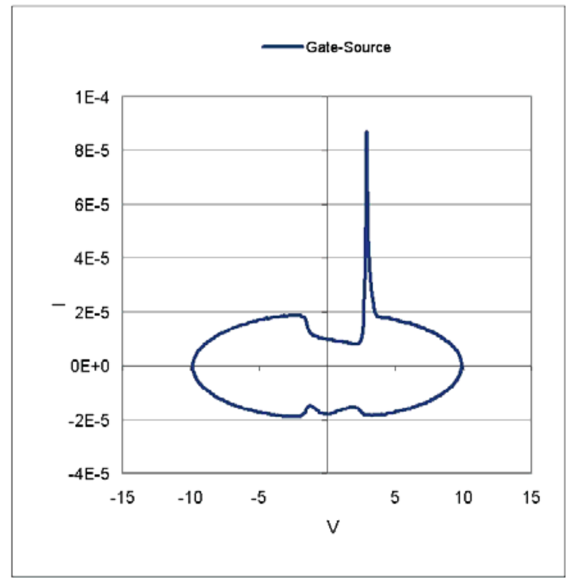


Figure 3: Typical AC behavior of the gate source circuit with floating drain pin (error free device).

The transient current shown in Figure 3 is analogue to the capacitances and the threshold voltage ($V_{GS\,(th)}$) of the device. The peak is due to the change of capacitance during the onset of the channel.
The curves of Figures 3 to 9 are measured using the following CTR parameters: AC mode, f=50Hz, max. amplitude voltage =±10V.

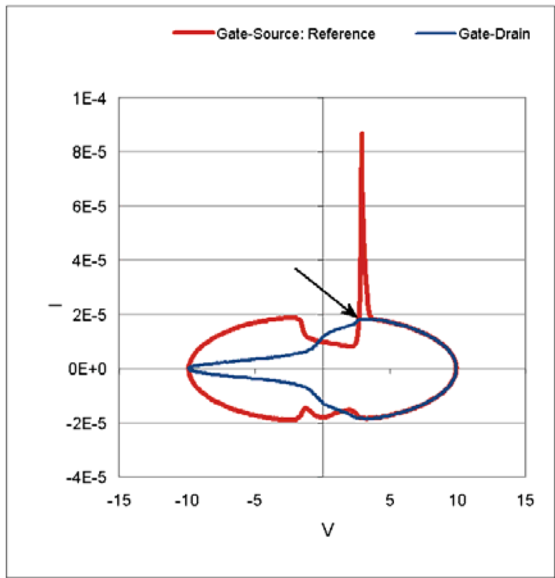


Figure 4: Gate-drain Lissajous figure with floating source pin (error free device).

In the G/D curve trace in Figure 4 a small mark is visible (arrow) due to the onset of the channels. As long as the channel is in on-state there is no difference to the gate-source curve. When the gate is negative the body-Epi diodes are reverse biased. The space charge regions are increased and the capacitance is reduced. Since the channel is OFF C_{GS} and C_{GD} are only connected via C_{DS}.

2.2 Lissajous Figures of Different Devices

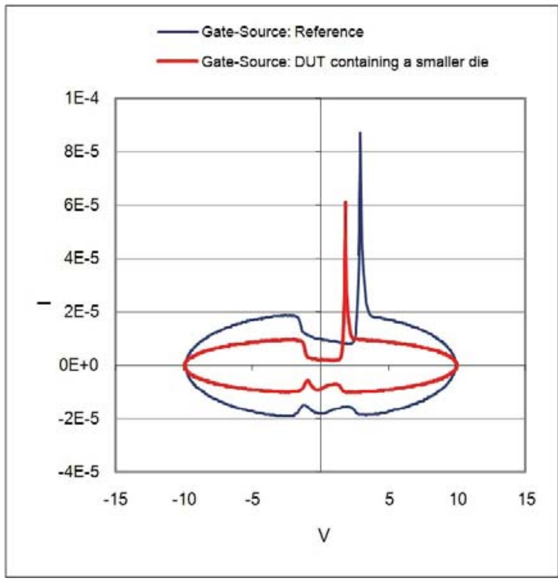


Figure 5: Comparison of gate-source Lissajous figures of different devices.

The tested device (red curve) in Figure 5 contains a smaller die of the same technology, resulting in lower capacities. This leads to a lower dynamic current. A lower threshold voltage causes a current peak at a lower gate voltage.

2.3 Lissajous Figures of Defective Devices

2.3.1 In Case of Drain-Source Short

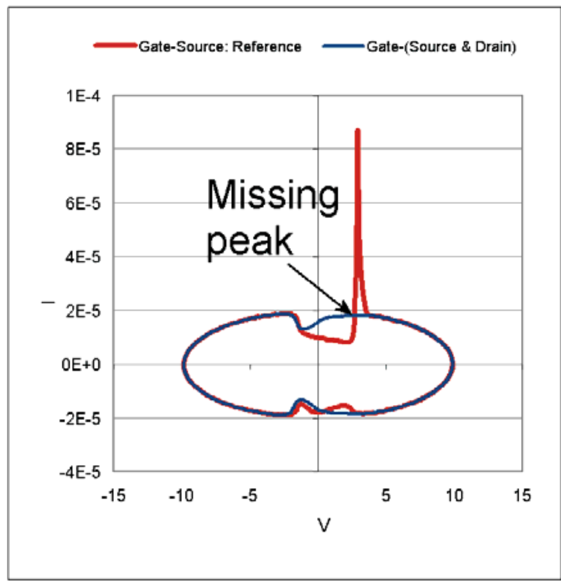


Figure 6: Gate-source Lissajous figure with drain shorted to source.

Due to the D/S short of the device in Figure 6 the onset of the channel has no effect (because D/S is already connected) and can't cause a peak as observed with the good sample (red dots).

2.3.2 In Case of Gate-Source Leakage

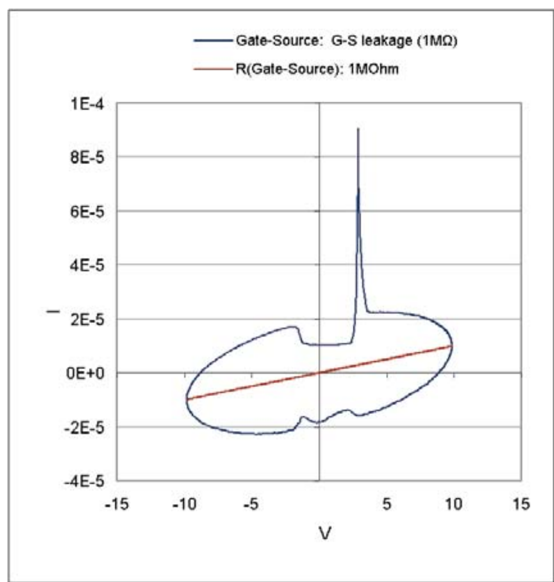


Figure 7: Gate-source Lissajous figure with an ohmic gate-source leakage present.

In Figure 7 the gate-source leakage current of the device leads to a rising gradient of the axes in the Lissajous figure.

2.3.3 In Case of Drain-Source Leakage

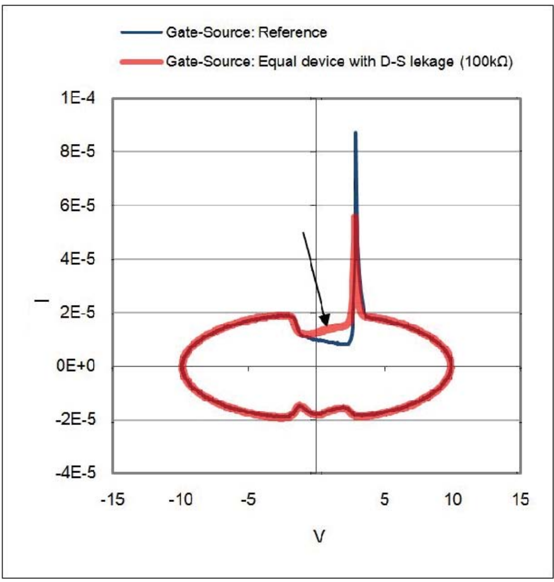


Figure 8: Gate-source Lissajous figure with a drain-source leakage present.

In the example of Figure 8 a 100kΩ resistor is connected in parallel with the D/S capacitance which leads to an increased current between 0V and the threshold voltage level (arrow). The difference of charge is lower when the channel switches ON which results in a lower peak.

2.3.3 In Case of Threshold Voltage Shift

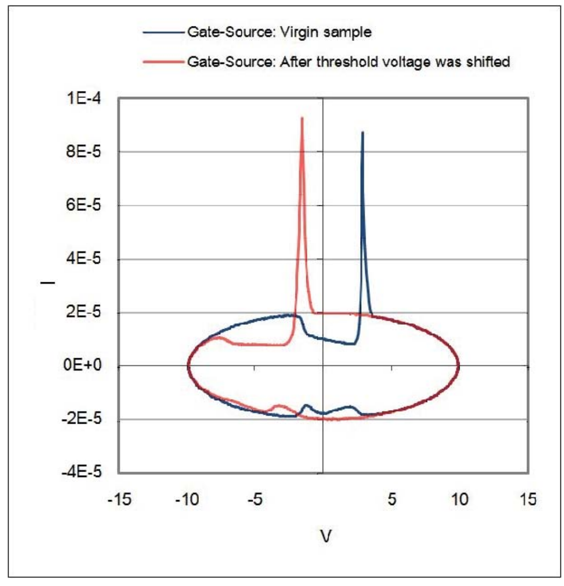


Figure 9: Gate-source Lissajous figure showing a shift of threshold voltage.

A threshold voltage shift as shown in Figure *9* can be caused by ionic contamination or gate-source over-voltage (with low energy). The latter was applied in this example. The overvoltage caused tunneling currents in the gate oxide but the breakdown energy was not high enough to inflict thermal damage. Instead charged traps occurred in the gate oxide which led to a shift of the $V_{GS\,(th)}$ peak to a lower value.

A threshold voltage drift can also be observed with alternative measuring methods as depicted in Figures 10 and 11.

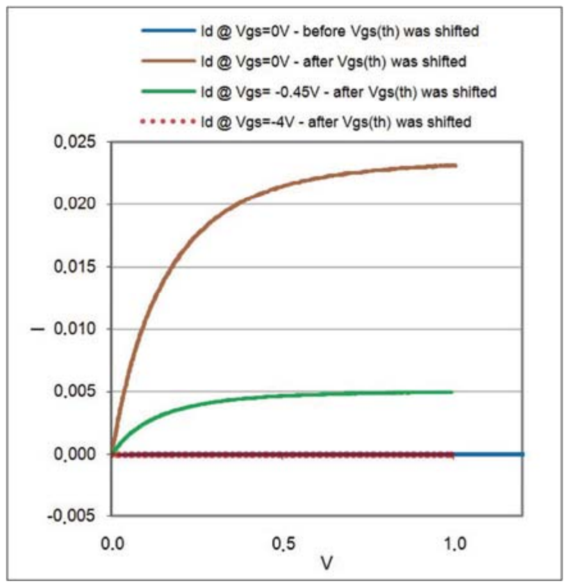


Figure 10: Drain-source leakage currents caused by threshold voltage shift.

Figure 10 shows another measurement of the device with the same $V_{GS\,(th)}$ shift as seen in Figure 9. Negative gate voltage is necessary to pinch off the drain-source leakage. In general, if it is possible to pinch off a D/S leakage, it can be concluded that the defect is a channel leakage. Other possible causes can be ruled out.

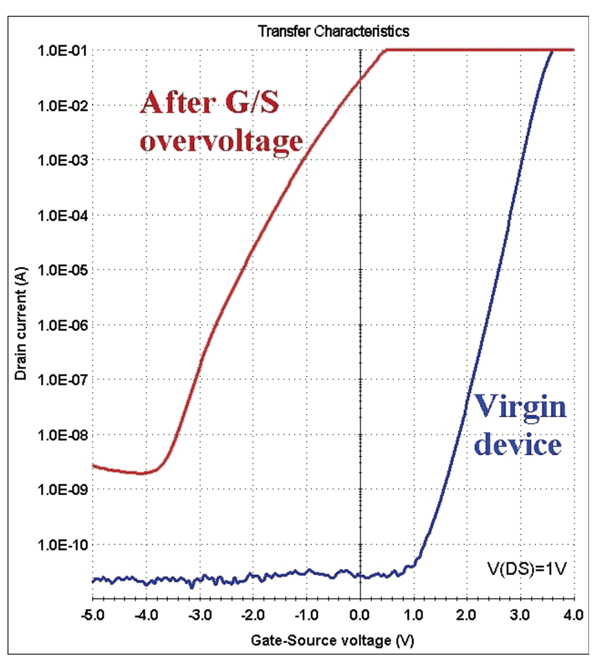


Figure 11: Transfer characteristics before and after G/S overvoltage

A useful alternative to the curve tracer for measuring of $V_{GS\,(th)}$ shifts is given in Figure 11. It shows the transfer characteristics of the same device which was discussed in Figures 9 and 10. These curves were measured with a parameter analyzer [3]

3. Typical Gate Defects and the Interpretation of there Characteristics

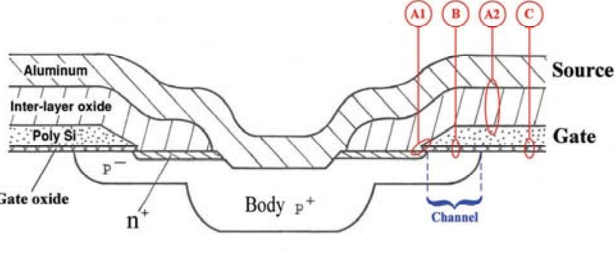


Figure 12: Illustration of principle locations of gate short circuits demonstrated at an N channel DMOS transistor *(source: Infineon).*

The positions of the shorts shown in Figure 12 are described as following:

A1: from gate (n^+ poly Si) to source contact (n^+ silicon) through gate oxide
A2: from gate to source metallization through inter-layer oxide
B: from gate to p^- Si of the channel region
C: from gate to n^- Si of the Epi Si (drain)

In the following Figure 13 these short positions are transferred to the principle image of an N channel transistor in trench technology for comparison.

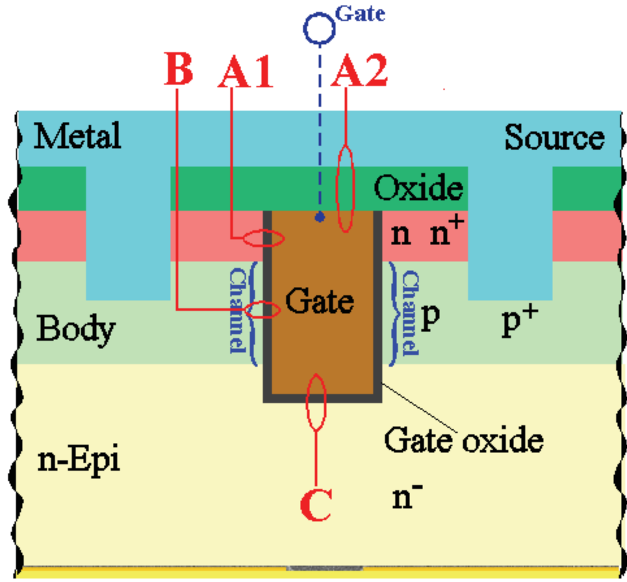


Figure 13: Illustration of the principle locations of the gate short circuits shown at an N channel transistor in trench technology (source: Infineon).

The Figures 14 through 17 show the measurements of devices with the short circuit as shown in Figures 12 and 13.

3.1 Short at Positions A1 or A2

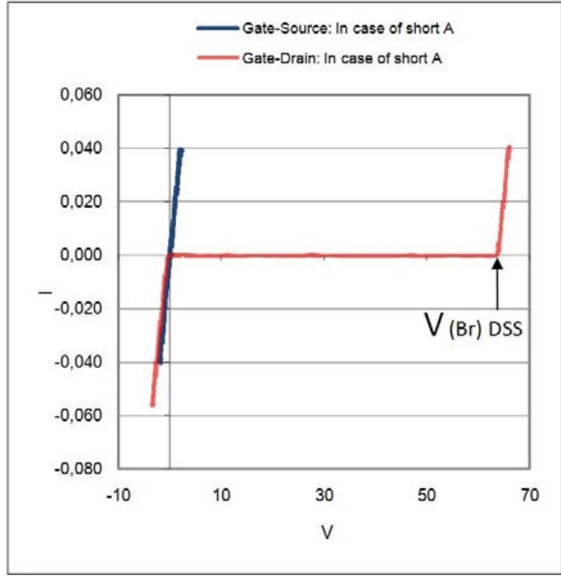


Figure 14: Examples of G/S (blue) and G/D (red) curves due to a short at site A1 or A2.

In Figure 14 the gate-source curve is nearly ohmic. Between gate and drain the forward and the blocking behavior of the drain-source (D/S) circuit can be observed (because gate is internally connected to the source by the short). If the body-Epi diode is not damaged and the remaining length of the channel is long enough the D/S breakdown voltage is not affected. The G/S short circuit resistance (blue) is added to both dynamic resistances.

3.2 Short at Position B

The occurrence of this case is rare, because commonly the channels are short. This reduces the possibility for such a particularly damage.

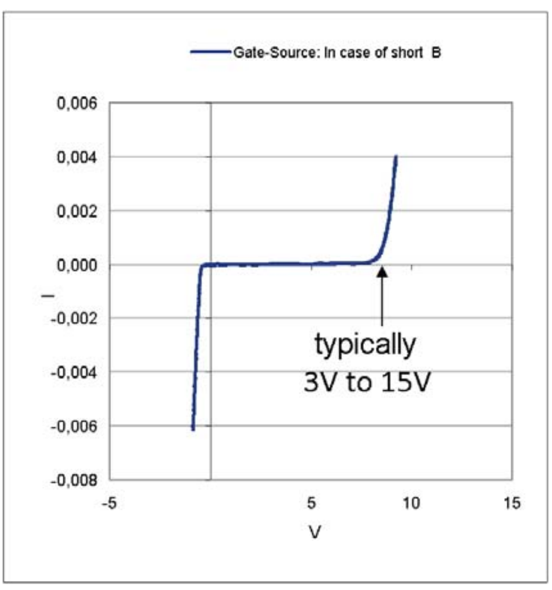


*Figure 15: Example of a G/S curve due to short at position **B**.*

As indicated in Figure 15 there is a high range of breakdown voltage in case of a small damage at position B. The melted contact between n^+ poly Si and p^- body does not deliver a defined blocking behavior. Typically it ranges from 3V to 15V. Further variations are possible depending on the doping and extension of the damage.

The following Figure 16 shows principle G/D curves resulting from damage at point B.

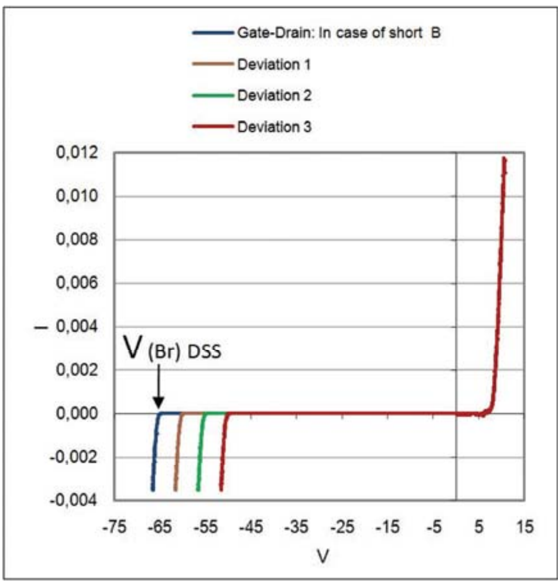


*Figure 16: Typical G/D curve due to a short at site **B**.*

If the damage is small enough, the channel is able to handle the full blocking. The blocking ability decreases with increasing damage.

3. 3 Short at Position C

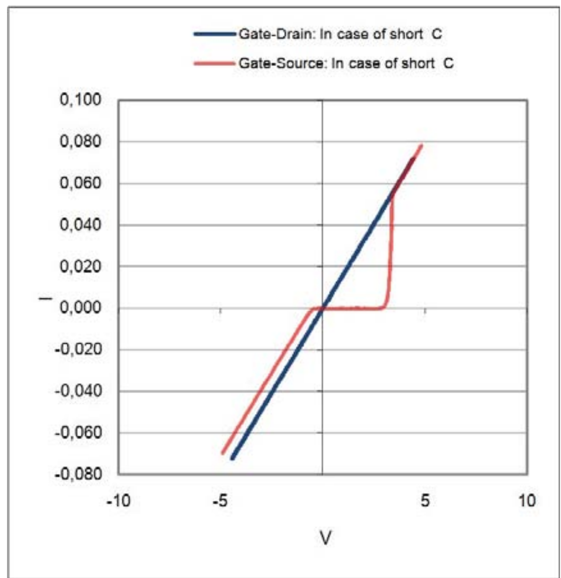


Figure 17: G/S and G/D curves due to a short at position C.

In Figure 17 the short circuit between gate and drain is nearly ohmic. The current from gate to source flows either via the inverse diode (n^- Epi to p^+ body if the gate is negative) or through the channel after the gate voltage reaches the threshold voltage level which turns ON the transistor. For voltages above the threshold voltage the current is limited by the serial resistors resulting from the poly Si, the bulk resistances, and the extension of the damage.

4. High Power Measurement With CTR

4.1 Measurement artifact

High power measurements like $R_{DS\,(on)}$ (drain-source ON resistance) and V_{SD} (inverse diode forward voltage) can heat up the device resulting in measurement artifacts if the wrong measuring mode is used (e. g. DC or half wave modes). Increasing the temperature results in an increase of the $R_{DS\,(on)}$ and in a decrease of the V_{SD}. These effects can be minimized by pulse measurement e. g. by using a High Power Curve Tracer which pulses the collector supply. In case of $R_{DS\,(on)}$ a pulsed gate voltage function on the curve tracer offers an improvement, but the result is not comparable to that of a High Power Curve Tracer.

Figure 18 shows an example of differences between pulse measurement and half wave mode.

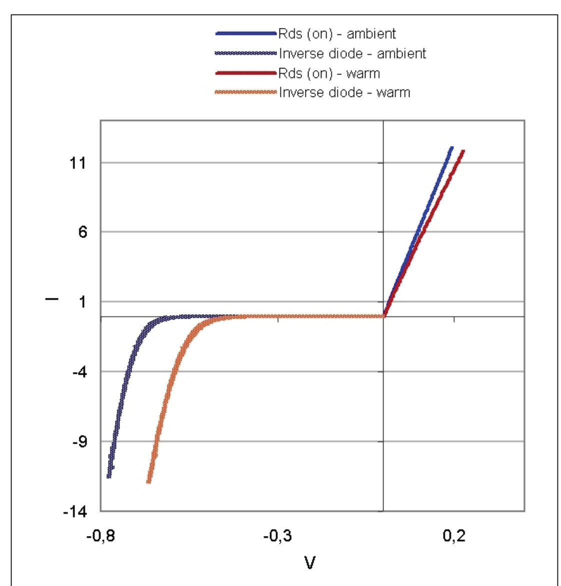


Figure 18: Typical RDS (on) and VSD curves showing their dependency from temperature.

In Figure 18, the blue curves were measured in pulse mode (T_j=25°C). During high power measurement in half wave mode the junction temperature of a transistor rose up. This resulted in an increased $R_{DS\,(on)}$ (red curve) and in a reduced V_{SD} (brown curve) which are measuring artifacts.

4.2 Investigation of devices with an Increased $R_{DS\,(on)}$

If an increased $R_{DS\,(on)}$ is caused by malfunction of the channels or by an ohmic component within the drain-source circuit, this can be verified using the differential resistance of the inverse diode. During this measurement nearly all ohmic components are active (except the contact between source metal and n^+ Si leading to the channel). If one of them is increased, it becomes visible in the differential resistance. Since the channel is not active during this measurement, its influence can be ruled out.

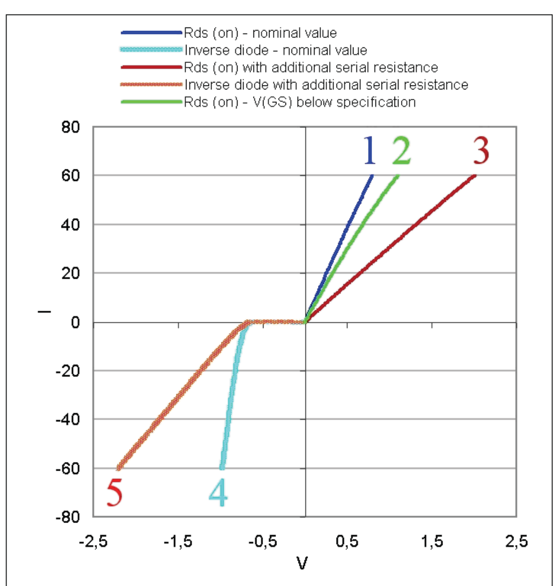


Figure 19: RDS (on) and VSD curves with different resistances (pulse mode measurements).

The curves in Figure 19 represent three cases.

Case 1: Reference. See curves no. 1 [$R_{DS\,(on)}$] and no. 4 [V_{SD}].

Case 2: Increased channel resistance, simulated by decreased V_{GS}; See curve 2 [$R_{DS\,(on)}$]. Curve 4 [V_{SD}] also fits to this case. The higher $R_{DS\,(on)}$ is caused by increasing the channel resistance. This has no effect on the V_{SD}.

Case 3: Increased ohmic resistance in the path between drain and source pins (e. g. due to an increased contact resistance), simulated by a serial resistor. See curve 3 [$R_{DS\,(on)}$] and curve 5 [V_{SD}]. This increase of resistance can be seen on both $R_{DS\,(on)}$ and on the differential resistance of the inverse diode.

Figure 19 shows that an increase of the resistance of the channel or the contact between channel and source metal only effects the $R_{DS\,(on)}$ curves. On the contrary, an increase of any of the other components between the drain and source pins can be seen on both the $R_{DS\,(on)}$ and the V_{SD} characteristics.

5. Conclusion

Basic information about a first electrical determination of defects at power DMOS transistors was described which allows an interpretation of the characteristics received by curve tracer measurement. Discussed were Lissajous figures which indicate leakages, short circuits, threshold voltage drifts, or just different dies inside the packages. Principle sites of gate shorts were illustrated which lead to typical failure characteristics. Typical curves of these defects were envisaged. It was described how measuring artifacts during high power measurement can be minimized and a possibility to distinguish causes of increased $R_{DS\,(on)}$.was discussed.

Considering the interpretations mentioned above, in many cases a first suspicion of the defect is given after a short curve tracer session. Therefore, beside the high precision measuring tools, curve tracing is still a useful tool for failure analysis.

6. References

The following list of references is recommended to further deepen the knowledge:

[1] **Curve Tracing Bipolar Devices:**
D.Wilson, Curve Tracer Data Interpretation for Failure Analysis Microelectronics Failure Analysis, Desk Reference Fifth Edition

[2] **DMOS Transistors:**
Jens Peer Stengl/Jenö Tihanyi, Leistungs MOS-FET Praxis, Pflaum Verlag München, ISBN 3-7905-0619-2

[3] **Oxide and Interface Trapped Charges:**
SEMOCONDUCTOR MATERIAL AND DEVICE CHARACTERISATION
Second Edition
Dieter K. Schroder
Arizona State University
Tempe, Arizona
A Wiley-Interscience Publication
John Wiley & Sons, Inc.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

High-Volume Scan Analysis: Methods to Avoid Failure Analysis

Darrell Carder, Steve Palosh, Rajesh Raina
Freescale Semiconductor Inc.
6501 William Cannon Drive West
Austin, TX 78735
Darrell.Carder@Freescale.com
Steve.Palosh@Freescale.com
Rajesh.Raina@Freescale.com

Abstract

Failure Analysis of IC's is an expensive process. This paper describes the benefits of deploying a Volume Scan Diagnostic Analysis flow to reduce the amount of die sent for failure analysis. After giving a description for the usage of volume analysis to reduce failure analysis, the paper reports a corresponding case study in the areas of Scan Test cleanup, and Correlation of failure modes across multiple devices. In closing, the paper discusses open challenges that require collaboration from several stakeholders in semiconductor industry.

1. Introduction

Scan based testing has become a standard for semiconductor designs [1]. This has been harnessed by Failure Analysis (FA) teams to locate failures and identify systematic trends in order to fix them and reduce fails on future wafer lots [2].

Diagnostics methods traditionally were based on single site stuck-at models. The callouts for a diagnostic run would give a logical net location, but the scores could vary significantly. The range of the score was influenced by a range of failure types that would be detected with the stuck-at based automatic test pattern generation (ATPG) patterns. Failures from opens to bridging, resistive vias, and speed related defects could not be accurately detected with stuck-at diagnosis.

Recent years have seen newer diagnostics technologies such as cell-internal [3], speed related diagnosis [4], and physical based diagnosis [5]. These technologies allow diagnosis down to the logical net and physical polygon. Detection between defects internal to the cell and interconnect defects are also achieved. The advances of these technologies have improved the callouts used for failure analysis.

Building on Failure Analysis success, a new flow was developed for achieving high-volume scan diagnostic analysis. Higher volume and statistical analysis enables identification of systemic failure trends that would not be seen at lower volumes. Many design-process marginalities can become visible. These new signatures can be utilized to reduce the number of die sent to failure analysis. A timely fix in either design or process or test - can increase yield and hence profit margin.

This paper discusses different usage models for volume scan diagnostics analysis and reports successful case studies on industrial designs. The use models cover Scan Pattern Cleanup, and correlating failure modes across multiple devices.

1.1 Organization of paper

Section 2 describes the setup and flow for collecting, storing and analyzing scan failures from a production environment.

Section 3 describes two use models for analysis implementations and cases studies on an industry ICs that are representative of a broad range of performance, power and cost. The real life case studies are used to prove out the use model and to accentuate the successes.

Section 3.1 describes how scan fails were determined to be false with no physical failure analysis required. The false fails were eliminated by regenerating tests using correct timing information.

Section 3.2 describes how a scan chain failure was correlated to scan logic failures using volume scan anlaysis. The correlation is then expanded to other devices using the same technology library.

Section 4 discusses common trends from the case studies and offers guidance on improvements. In Section 5 the paper concludes summarizing the key contributions of the paper along with the learning's and future direction.

2. Volume Data Analysis

2.1 Data Collection Requirements

Accurate collection and processing of scan failure data is critical for the successful completion of a volume diagnostics flow. A production wafer sort environment only allows one chance to capture the required data correctly before the devices move on to packaging.

To perform diagnosis there are certain data requirements that need to be met to achieve a usable flow. These requirements are outlined below along with an explanation of how each of them is challenged by traditional test development and production flows.

Knowledge of additional cycle counts added during test program development is required for diagnosis. This is to avoid mismatches between failing cycle number in the Automatic Test Equipment (ATE) data logs, and the corresponding cycle number in the automatic test pattern generation (ATPG) pattern.

Adding additional cycles is a common practice on devices that utilize an on-chip phased locked loop (PLL) for scan clocking (at-speed capture). If the PLL lock time is not set correctly in the ATPG test setup, then additional repeat-cycles are added during ATE pattern development.

Knowledge of any name changes is required so that pins can be re-mapped back into their original ATPG pin names prior to diagnosis.

Test development teams are often required to modify the pin names from the original ATPG pattern to be compatible with the syntax limitations of the target ATE pattern format.

Masking information for each ATE pattern must be documented and fed back to the diagnostic tool as part of the flow.

Masking of strobes is sometimes performed with pattern conversion flows or manually by a test engineer during silicon validation. Losing track of what has been masked will compromise diagnostic results.

Production scan fail data collection must continue after the first failing cycle until the required amount of failing data is collected for the given scan pattern.

This is contrary to traditional production test methodology that stops testing after the first failing cycle to save test time. The volume diagnostics flow must consider that each different ATE platform will have different limits as to the amount of failure data it can collect in its memory during a single pattern execution.

In the case of chain failure diagnosis, additional failure data will be required beyond the limits of most ATE. In these cases the pattern will need to be re-run with the next set of failure data collection starting at the point where the previous failure memory reached its limit. Test time will be increased as a result.

Once the scan data is collected on the ATE it needs to be stored in a standardized data format. Flows that involve multiple ATE vendors, separated business groups, and legacy products will often require an intermediary translation from the format of a particular ATE platform to a format that is compatible with the diagnosis flow [Figure 1].

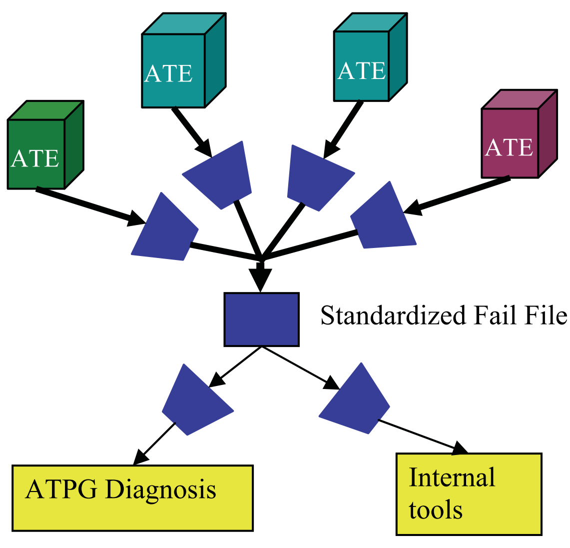


Figure 1: Fail data collection in a heterogeneous environment

2.2 Volume Diagnostics Flow

Taking into account the requirements of Section 2.1, the flow presented in this paper uses a legacy scan datalog format that has been standardized per ATE platform.

The flow converts each of these scan datalogs into an intermediate standard format to work with all tools in the remainder flow. Figure 2 shows how scan failures were taken from ATE at wafer probe and processed through the various diagnosis tools. The output of the volume analysis includes the option of sending failure data information straight to design or the FAB for corrective actions without going through FA.

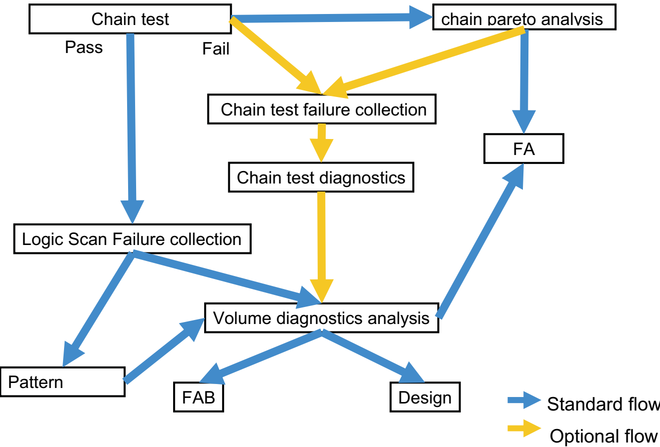


Figure 2: Volume Diagnosis Flow

3. Results

The yield ramp of the each device described in this case study involved the challenges of both a new product design and the introduction of a new FAB process technology.

3.1 Scan Pattern Cleanup

During silicon bring up a subset of the ATPG patterns passed at simulation but failed on silicon. The fail signature was not consistent across the wafermap even though process engineers were confident that the factory was meeting it's specifications for this technology. Since this was a new product, test pattern marginality was investigated.

The following case study confirms the use of volume diagnostics analysis to identify signatures associated with these scan failure issues.

Case Study I: Using volume diagnostics to correlate a failure mechanism to design flow issues

This case study involves the introduction of a high reliability low power 90nm process technology designed for the automotive safety market. The device in this example is a member of 32-bit microcontrollers based on Power Architecture™ with applications that include Electrical Hydraulic Power Steering (EHPS), low end Electrical Power Steering (EPS) and airbag control.

Spatial signatures of failing die on wafer level bin maps are commonly used to categorize FAB process-based failure mechanisms [6]. Grouping wafers by their spatial pattern, YMS (Yield Management Systems) can be used to data-mine various elements associated with the wafer as it was processed through the FAB.

In cases where a direct correlation to process variation cannot be made with reasonable effort, factory management must decide if an excursion effort is required that will pull resources away from other critical efforts. The timeline to resolve issues buried in the design flow can be extremely long which is very costly to the IC manufacturer.

The volume diagnostics flow was applied to a situation that involved the yield ramp on a new product that was experiencing minimum voltage transition scan test fallout at probe. Wafer level bin maps indicated a spatial signature at the center of the wafer for lots that were affected.

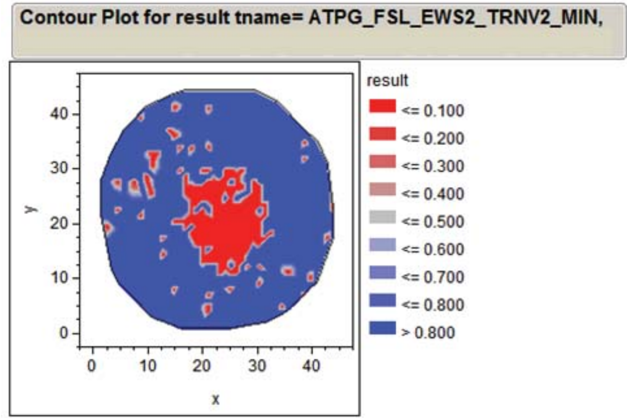


Figure 3: Wafer Bin Map (spatial)

The problem soon escalated when this material was packaged for final test. The same problem was seen with high scan test fallout during cold temperature testing.

After an initial investigation the FAB reported that although some process variation does occur at the center of the wafer it is well within the specifications window for that process technology.

If this variation was causing the speed sensitivity then the design team would need to get involved to understand why this is happening.

The scan fail data was loaded into the volume diagnostics and analysis flow. Failure data for the min vdd transition patterns used for this analysis was collected from 4 lots and 10 wafers totaling 252 failing die.

When analyzing the results a specific load/unload sequence during a scan test indicated a high failure rate at min vdd. This was identified via a pattern fail pareto.

Figure 4 shows that ATPG pattern 27 is at the top of the pareto. This single ATPG load/unload sequence accounts for 42% of the scan failures. The remaining pareto distribution across other patterns drops off very quickly.

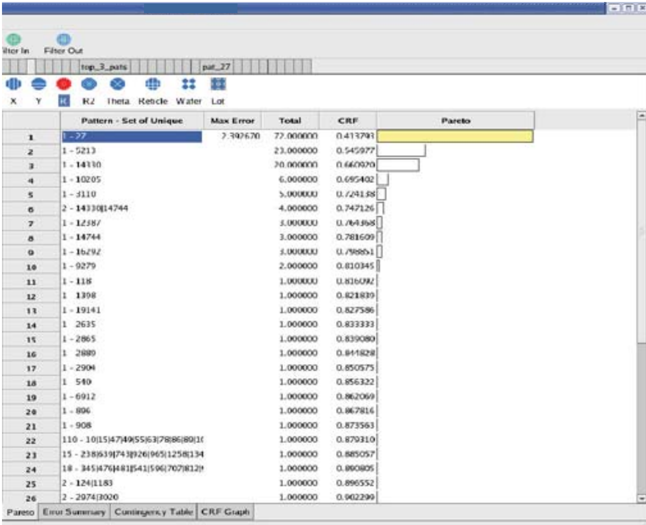


Figure 4: Pareto of failing test patterns

Filtering on this atpg pattern pareto a zonal signature of diagnostic callouts is shown in figure 5. The spatial signature of the zonal map for ATPG pattern # 27 looks similar to the original wafer level bin map seen in the production data (Figure 3).

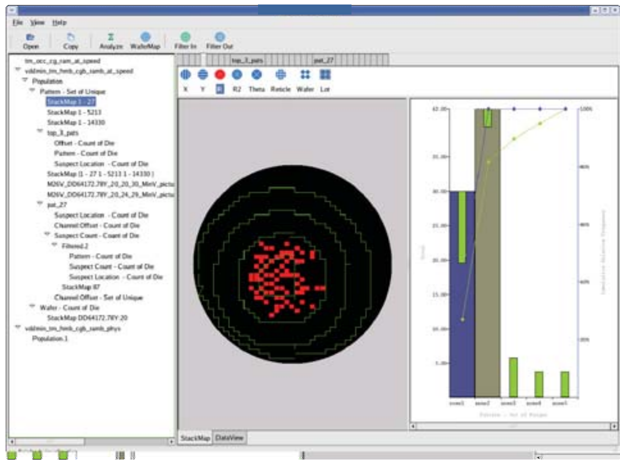


Figure 5: Spatial fail map for ATPG pattern # 27

The next step in the analysis was to look at the logic being tested with pattern 27. All failures for ATPG pattern 27 failed the same path. This path was identified using the diagnostics schematic viewer and the diagnostic report file. The failing paths for atpg pattern 27 were displayed and reported to a text file.

The path information was sent to the design team and was verified that in static timing analysis (STA) the path was declared a multi-cycle path. Further investigation identified that the path was missed in the scripts porting the multi-cycle paths to the ATPG setup.

Based on these findings, the multi-cycle path was included into the ATPG constraints and a new set of patterns was generated. Cold temperature testing was repeated with the new patterns to the same set of packaged devices. The systemic fails seen earlier, were now eliminated.

Feedback of "actionable information" from the FAB in the context of the design environment meant that the problem could be resolved quickly with no failure analysis resources utilized. Design was then able to improve their flow and the FAB was able to re-focus yield efforts towards other process related issues.

3.2 Correlating failure modes across multiple devices

When a failure is identified to be within a library cell there is the concern if this is a systematic failure and to what extent of an affect is on all devices using that technology library. When running volume scan diagnostics on multiple devices the data is available to compare failure rates of the libraries cells across these devices.

Case Study II: Using volume diagnostics to correlate failure mechanism across products

In parallel with the design flow issues identified in section 3.1, another failure mechanism was also appearing at the same time and required investigation.

Yield engineers reported scan chains to be failing near the edge of the wafer on certain lots (Figure 6).

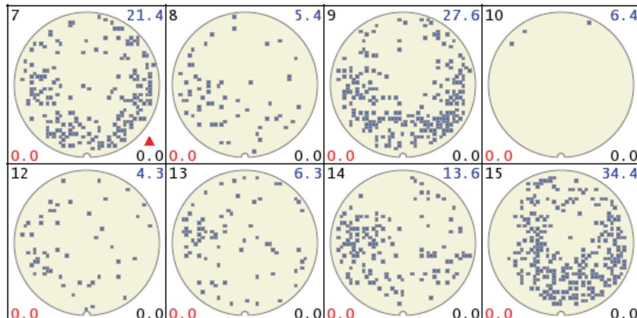


Figure 6: Wafer bin maps for chain test (Spatial)

The chain failure characteristics were timing and voltage sensitivities. Slower shift frequency and lower voltage would cause the failure.

Volume diagnosis was performed on the stuck-at scan failures that passed chain test and the highest scoring diagnostic callouts were selected for further analysis. Viewing the wafer map analysis of these failures indicated majority of the failures were on the edge of the wafers. A selection of die from the scan failures were packaged for FA along with die selected from scan chain failures.

When the packaged dies were received by FA some initial characterization of the failures was performed. A shmoo plot (Vdd vs. Time Period) of the chain failures on the left shows a similar signature to the shmoo of the stuck at failures shown on the right (Figure 7).

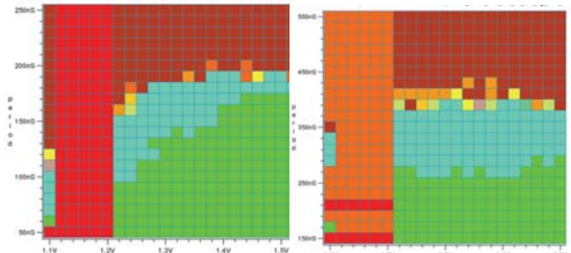


Figure 7: Chain Test (left) and Stuck-At Test (right)

The main difference between the two is that the chain tests will fail for periods greater than 100ns and the stuck-at tests will fail for periods greater than 250ns. This shows that the chain test patterns are more sensitive to this mechanism than the stuck at.

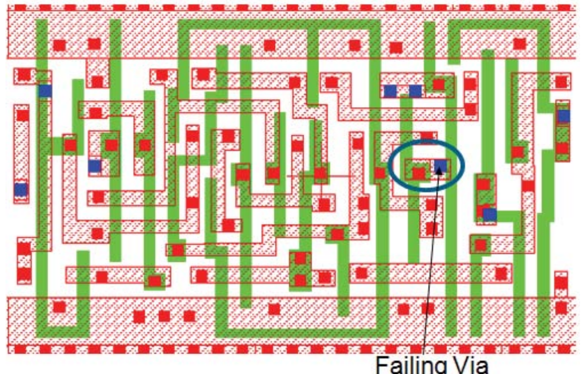
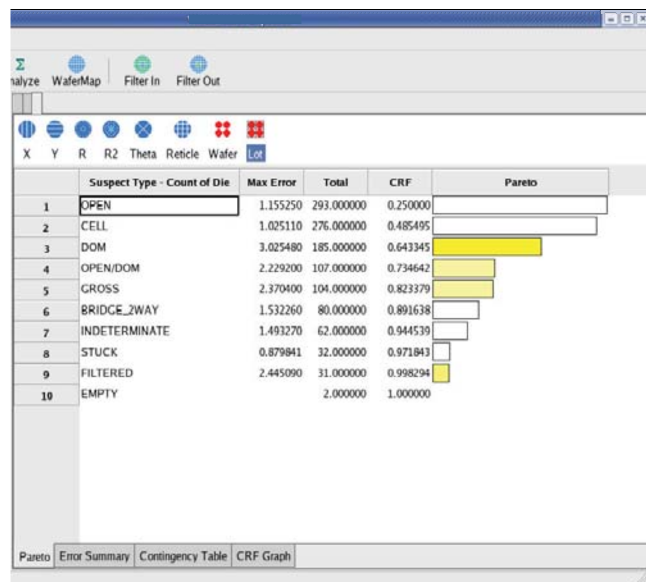


Figure 8: Cell clock input failure point

Electrical Failure analysis was performed on both a scan chain and scan logic failure. EmiScope data obtained correlated the exact same cell failure location for the chain failure and scan logic failures. Figure 8 shows the via1 failure point within the cell on the clock input.

With an initial indication that the failure mechanism identified in the chain test can be seen in the stuck-at failures, volume diagnosis across multiple wafer lots was performed on a larger number of failing die.

Layout aware diagnosis was used for this larger sampling of failures.



	Suspect Type - Count of Die	Max Error	Total	CRF	Pareto
1	OPEN	1.155250	293.000000	0.250000	
2	CELL	1.025110	276.000000	0.485495	
3	DOM	3.025480	185.000000	0.643345	
4	OPEN/DOM	2.229200	107.000000	0.734642	
5	GROSS	2.370400	104.000000	0.823379	
6	BRIDGE_2WAY	1.532260	80.000000	0.891638	
7	INDETERMINATE	1.493270	62.000000	0.944539	
8	STUCK	0.879841	32.000000	0.971843	
9	FILTERED	2.445090	31.000000	0.998294	
10	EMPTY		2.000000	1.000000	

Figure 9: Pareto of diagnostics suspect type

The diagnosis suspect type pareto (Figure 9), shows that Cell failures are the second highest failure for this set of data (276 failing die).

Further analysis for the Cell callouts identify 172 failures with the same Register cell type and clk input pin as the failures sent to FA.

- For the scan data received the total # of scan failures for these lots is 1402
- This failure mode is 12% of overall scan logic failures.

These results provide correlation between the scan chain test failures and the scan logic failures. A volume analysis signature is identified that can be used to determine the affect of this failure mechanism across all device using the same technology library.

With the knowledge of a specific correlating signature the volume diagnostics analysis was expanded to include additional devices using the same technology library.

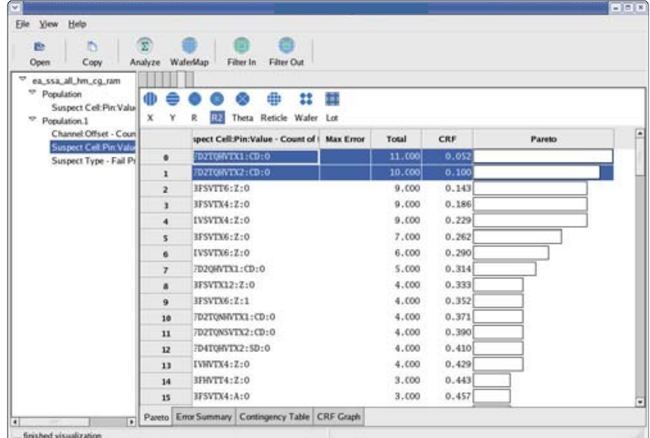


Figure 10a: Second device showing cell/pin signature.

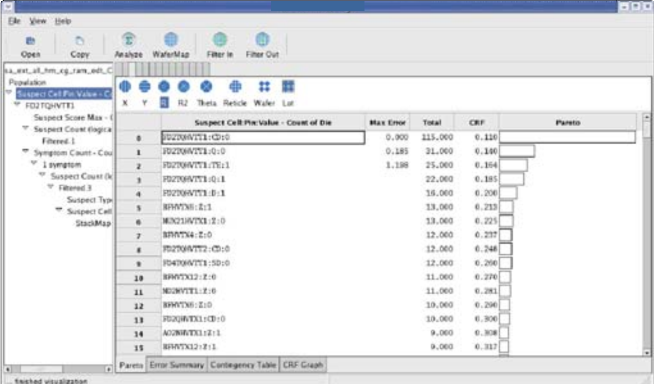


Figure 10b: Third device with cell/pin signature.

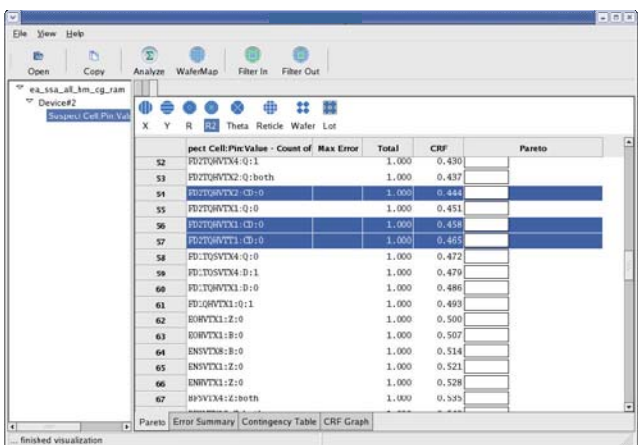


Figure 10c: Forth device with cell/pin signature.

Figures 10 a, b, and c show three additional device paretos for cell and pin failures. All three devices have the common signature for the register clock input via1 failure. Knowing that this signature is appearing on different products means that the yield loss was related to the new process technology and not one specific design.

As a result of this learning the appropriate level of priority and engineering resources could be assigned to this issue and a process engineering task force was formed. Having the diagnostics signature allows monitoring of the failures during the process fix to validate corrective actions without requiring failure analysis.

4. Discussion

The motivation for implementing a diagnosis-driven yield analysis flow in volume is to identify subtle design margin violations. These marginalities may not be visible to design or manufacturing using traditional methods. With the usage of the flow it became evident that the volume scan diagnostics analysis had the power to minimize and in some cases eliminate the need for failure analysis. The expectation of the flow was to utilize the volume analysis to identify the best candidates for failure analysis. These candidates would be selected on score and number of suspects. During the course of our study we were surprised by the secondary benefits of high-volume scan fail analysis as described in section 3.

An improved design signoff flow has been mandated as a result of this experience. Despite the lack of full automation, the case studies illustrate a couple of common points: (a) The capability to collect and analyze scan fails through various data filtering and datamining mechanisms is a useful supplement to the existing yield improvement methods. (b) Subtle fails are not limited to design, or test or process for a cause (or a fix) – but rather a complex sequence of events leads to these subtle fails.

In the past these failures may have well gone undetected and unresolved through the lifetime of a product. However, with the increasing automation and datamining capability, IC makers expect to overcome the yield challenge associated with deep submicron technologies.

The open challenges for our industry are summarized below:

(a) All ATE must develop, agree and use a standard scan fail log format.

(b) All ATPG tools must provide standard test generation format [STIL] and accept standard fail format

(c) On-site ATE manipulation of test patterns is a necessity (nothing works perfectly the 1st time). This includes adding repeat cycles, masking certain bits, etc. Development of a standard framework that ATPG and ATE understand such that the last minute manipulations can be fed back to test generation source is required.

(d) Pin formats and naming must be standardized.

(e) A standard must be developed to pass timing information (including multi-cycle paths) between Design and DFT

5. Conclusions

A high-volume scan analysis flow is easy to understand. However, deploying it in an industrial environment presents practical challenges. This paper provides a peek into these challenges starting with the data collection requirements and the flow for diagnostics analysis. Despite the challenges, the use of the flow is demonstrated in the case studies. The case studies highlight how various functions of the IC industry – from design to test to FAB to yield to FA must all work together in order to locate the root cause for a class of defects in a timely manner. Many are subtle defects that likely went undetected through the lifetime of earlier generation products. The main contribution of the paper is to share the practical challenges and methods for solving them in hopes of rallying IC development ecosystem towards greater automation in high-volume scan fail analysis and diagnosis.

6. Acknowledgements

The author's would like to acknowledge the support of the various teams at Freescale who were involved in this effort that include the FA and YE organizations of the MOS13 FAB, ATX Product Analysis Lab, as well as the individual efforts driven by Stephen Traynor, Cliff Howard, Helmut Lang, and Ernst Aderholz.

8. References

[1] D. Appello, A. Fudoli, K. Giarda, E. Gizdarski, B. Mathew, and V. Tancorre, "Yield analysis of logic circuits", Proc. of VTS, pp. 103-108, 2004.

[2] M. Sharma, B. Benware, L. Ling, D. Abercrombie, L. Lee, M. Keim, T. Huaxing, W. Cheng, T. Ting-Pu, C. Yi-Jung, R. Lin, and A. Man, "Efficiently Performing Yield Enhancements by Identifying Dominant Physical Root Cause from Test Fail Data", Proc. of ITC, pp. 1-9, 2008.

[3] Guo, R., et al., "Detection and Diagnosis of Static Scan Cell Internal Defect," Proceedings of the IEEE International Test Conference, 2007.

[4] Mehta, V.J., et al., "Timing Defect Diagnosis in Presence of Crosstalk for Nanometer Technology," Proceedings of the IEEE International Test Conference, 2006.

[5] Keim, M., "Layout-Aware Diagnosis of IC Failures," *IC Design* and Verification Journal, January 2009.

[6] R, Turakhia, M. Ward, S. Goel, B. Benware, "Bridging DFM Analysis and Volume Diagnostics for Yield Learning – A Case Study", Proc. of VTS, pp. 167-172, 2009

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Differentiating between EOS and ESD Failures for ICs

Leo G. Henry, Ph.D.
ESD/TLP Engineering, LLC
Fremont, CA
leogesd@pacbell.net

INTRODUCTION:

Distinguishing between EOS and ESD failures and differentiating the subtle differences between damage due to the several distinct ESD stress "Models" continues to challenge failure analysis capabilities as device dimensions shrink and critical defect sizes are reduced. Most of the ESD damage sites are not visible to optical microscopy and de-processing together with very high magnification examination using the SEM is most often necessary. However, the use of stress model simulators can most often replicate a Failure Signature, i.e., a unique die location and morphology associated with the ESD model [1, 2].

In the Semiconductor I/C industry, it has been well documented that the proportion of factory and customer field returns attributed to device damage resulting from Electrical Over-Stress (EOS) and Electro-Static Discharge (ESD) can amount to 40 to 50 % [3, 4]. These ESD events, which are the subset of EOS events, are associated with high voltages and can be replicated using one of several ESD Failure Model Simulators. It is to be noted that additional hard (functional) and soft (leakage) failures which also occur in the factory are normally detected by effective test programs. It is therefore still necessary to determine the probable cause of failure by Laboratory Simulation and Failure Analysis before effective corrective action can be accomplished.

This tutorial utilizes the results obtained from the evaluation performed on hundreds of devices over several years (mainly CMOS technologies- micron to submicron). The study entailed EOS and ESD simulation using a variety of models, conducting detailed electrical and physical failure analysis and then comparing the results with documented analyses performed on customer field returns and factory failures. The immunity to EOS and/or ESD is indicated by parametric measurement and/or functional performance after exposure to the stress.

As a result of the differences in the EOS and ESD current stress magnitude and the associated time domain, we can determine the location, type, variation and magnitude of damage at the failure site. This is then used to establish a relationship between the electrical signature, physical damage type, and location. The actual physical FA procedure will not be described here as this topic is addressed in other sections of this desk reference [5].

Physics of Junction Failure:

It has been shown that there is similarity in damage resulting from ESD due to HBM or MM [6]. This is due to the common underlying physics for the thermal failure of P/N junctions. The dependence of failure power upon pulse length for high current square waveform pulses is illustrated in Fig. 1 and described by the following one-dimensional electro-thermal equation known as the Wunsch-Bell Model [6].

This relationship describes the behavior of the instantaneous failure power density in a device plotted against the logarithm of the input pulse width (Fig. 1).

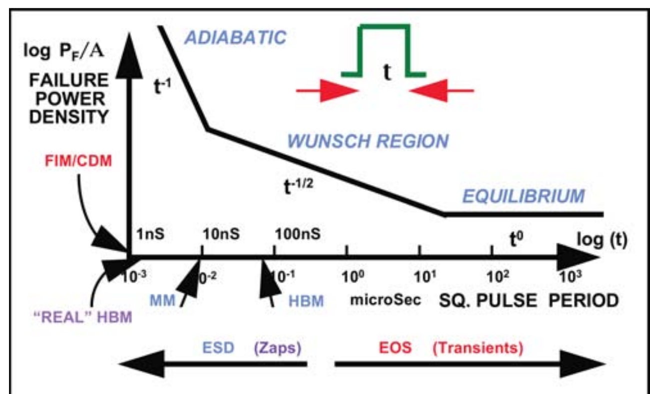


Figure 1 *- The Wunsch-Bell Model for Junction Failure.*

$$P_F/A = K_1 \times t^{-1} + K_2 \times t^{-1/2} + K_3 \times t^0$$

Where P_F is the failure power, A is the junction area, t is the pulse period and K_1, K_2 and K_3 are proportionality constants.

This shows the characteristic "Wunsch" region which describes most ESD failures due to an external event, i.e., discharge to a package pin from the human finger when a different (another) pin (or pins) is (are) grounded. Also shown is the "equilibrium" region for EOS events occurring for pulse widths greater than a few microseconds.

The time duration for typical EOS events is shown to be approximately 1 to 100 milliseconds, but for ESD events (from charged metallic objects and charged personnel) to the package, the duration will be approximately 10 to 150 nanoseconds. However, the failures due to a Charged Package are below the 5-nsec pulse width region [7] and is at the fastest time domain region (< 500 psec rise time). Note that the region marked "real HBM" refers to the time domain below approximately one nanosec, and relates to the electrical disturbances likely to cause system upset due to discharge from the human holding a metallic object [8].

EOS and ESD Failures:

While obvious EOS may be physically evidenced by cracked packages, carbonized plastic, burnt-out bond wires and massive visual damage to the metal on the die surface, the more subtle ESD failure is not visible at low magnification. However, in most instances, the observation of discoloration will provide a means of differentiating between the EOS and the ESD events [1, 2], since the EOS may result in a discoloration radius greater than 4m [9]. Thus EOS damage is visible at low magnification under an optical microscope, and the fail site will be surrounded by a region of discoloration; (note that the use of color Polaroid film facilitates the recording of any discoloration visible after de-capsulation). However, the relative absence of discoloration together with the need to de-process and to use high magnification (including the Scanning Electron Microscope) to locate the failure site indicates a type of ESD failure.

May and Guravage [9] have also shown that the region of apparent discoloration which surrounds the damage site due to EOS, visible at low magnification at the top glass to metal interface, has a radial extent that directly relates to the pulse length in the range of 100 nanoseconds to perhaps 30 milliseconds. This occurs because the thermal transient has enough time to diffuse laterally away from the site of EOS damage across the die surface.

Dielectric Type Failure:

Dielectric failure resulting from the Electric Field dependent Charged Device Model (CDM) is indicated in Figure 1 at the fastest end of the time domain. This type failure is due to discharge from the package to ground via a single pin with effective pulse duration of much less than 5 nanoseconds. The severity of the failure in this case will largely depend upon the package type used. Susceptibility is increased in SMT packages with very low lead inductance and with a large fraction of the lead-frame body capacitance upon the die paddle [10].

There are two charge modes for CDM [11]: the field induced mode (FIM) and the direct charged package mode (CPM). The distinction between these two electric field dependent failures is as follows:

The FI mode type of failure occurs when a single pin is grounded while the package lead-frame is at elevated potential. Charges are induced on the lead frame while the device is located within an electric field. Note that the failure occurs only after one of the pins has been grounded.

The CP mode type of failure also occurs when a single pin is grounded, but here induction on the lead frame is due to the electrostatic charge already present on the package surface(s). The charged package may be due to contact electrification or triboelectric generation. Note again that the failure occurs only after one of the pins has been grounded.

The actual failure mode for both types is identical for the potential imposed upon the lead-frame, and normally results in a unique failure signature far internal to the die consisting of gate oxide damage at the first Input buffer. This location is beyond the HBM and MM protective structure at the bond pad. Most often the oxide failure is located beneath the poly gate in the gate oxide, or is located at the poly gate edge but still in the gate oxide.

In this tutorial, reference shall hereafter be made to both the FI mode and CP mode as CDM. The three major packaged device level ESD models were used; namely the Human Body Model (HBM), the Charged Device Model (CDM) and the so-called Machine Model (MM). Four EOS simulations were developed and applied: (1) programming over-stress (variation of the programming voltage applied to special high voltage pins); (2) forced I/O reverse breakdown (actually BV_{DSS}); (3a) reverse insertion for dual in-line packages (DIP); and (3b) misorientation for surface mount packages (an example is PLCC); and (4) transient latch-up [TLU] using a capacitive discharge [12, 13]. TLU will not be discussed here as it is the subject of a more advanced tutorial.

For the ESD Simulation, use was made of state-of-the-art commercial simulators, but for the EOS stresses, in-house apparatus was used as required. Stress procedures for ESD simulation conformed to those established by the ESD Association Standards: HBM-S-5.1 [14], MM-DS-5.2 [15], CDM-DS-5.3 [16], and where appropriate, the MIL STD-883 [17].

ESD Laboratory Simulations:

The sample sizes for HBM ESD and MM ESD are as specified [{2N+1}xVxL] in the standards. N is the number of independent power supplies, V is the number of voltage levels to be stressed and L is the number of different lots from which the samples are taken. This is typically three. For the HBM (sample size used was 60), the stress level ranged from +/- 500 Volts to +/- 6000 Volts. For MM (sample size

used was 32), the stress ranged from +/-100 Volt to +/- 1000 Volt. For CDM only, the sample size (21) is simply VxL with a stress range from +/- 250 Volt to +/- 3000 Volt. All units were data-logged both pre- and post- test such that any leakage current equal or greater than 500 nano-Amps was "flagged" for analysis. Failing devices were analyzed using conventional electrical, chemical and electrochemical techniques. High magnification SEM examination had to be used to distinguish the physical damages.

HBM ESD Discharge:

For the HBM simulation, the physical failures occurred mainly in the ESD protective structure at the input pads and within the "self protecting" output circuits, but may also occur in the core of the device if the failure is functional or Icc between the Vdd (or Vcc) and Vss pins. The failing devices can have a variety of electrical characteristics at different pins: resistive short-circuit, Icc failures, functional, leakage currents (Fig. 2) as low as 700 nano-Amps, and breakdown voltages down to 2–3 volts. These devices were all stressed to failure.

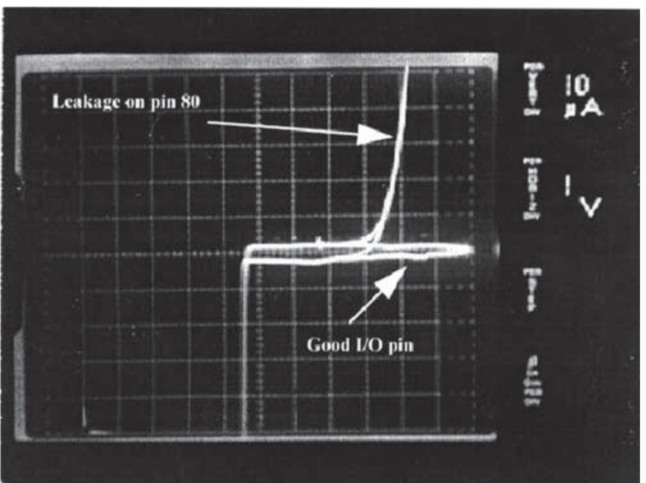


Figure 2 - *I-V characteristic curve comparing good (horizontal line) and leakage pins (curved vertical line).*

Figure 3 shows an optical photograph (<1000X) of the ESD failure. The failure site is revealed as a hot spot location (circled) revealed by liquid crystal. This is for an apparent electrical short circuit resulting from the I/O to V_{SS} pin combination subject to +/–3000 Volt HBM stress. Note that no physical damage is visible at this low magnification and there is no discoloration at the hot spot location. The physical damage is assumed to be below the surface at this stage. This figure shows the pad and the multiple fingered ESD protection structure connected also to the Vss bus line (left side looking at picture).

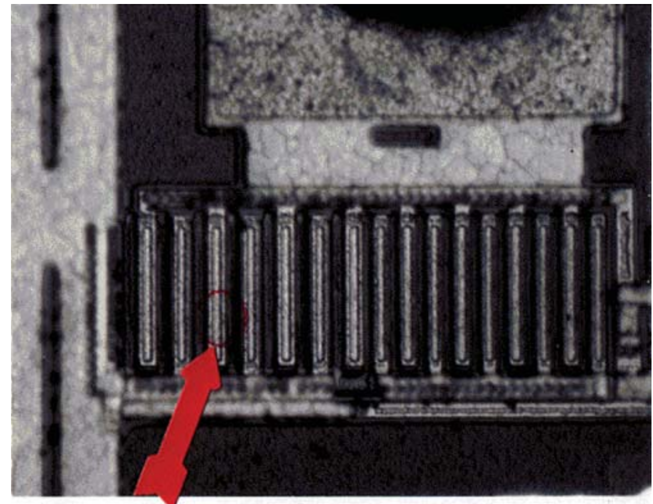


Figure 3 *- Hot spot location (circled) revealed by liquid crystal for a short circuit in the ESD protective circuit (transistors) at the pad.*

Note that for the I/O to Vcc pin combination stress, the ESD failure would occur (and did occur) on the other side (right side looking at the picture) of the ESD protection structure as this side is connected to the Vcc bus.

The ESD damage shown in Fig. 4 is in the contact area and at the edge of one of the fingers. This SEM (20kX) photo represents the failure site from Fig. 3 above and is taken after deprocessing down to the poly gate level. The physical failure then is below the surface, a distinction which will be used to distinguish between ESD and EOS failures.

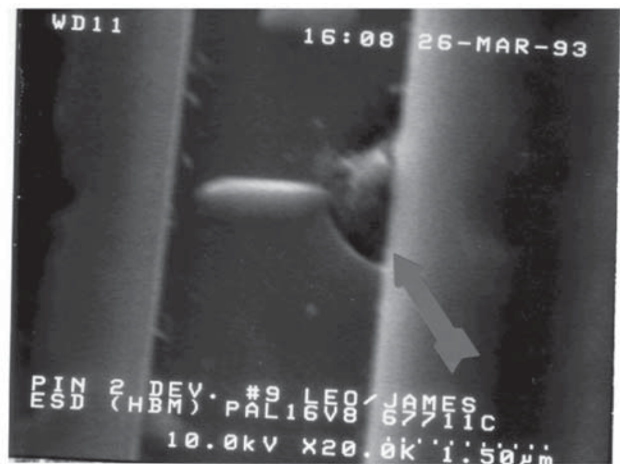


Figure 4 *- Subsurface ESD damage to the contact at the failure site in the ESD protection circuit. There is no discoloration.*

We note also that there was no discoloration at the ESD failure site, another distinction which will be used throughout this article to differentiate the ESD failures from the EOS failures. Although the ESD failure is the result of I/O to Vss stress, the I/O to Vcc HBM stressing gives the same ESD failure signature.

Emission microscopy [5] is particularly useful to detect anomalous regions or damage, which results in the generation of an emission at wavelengths of 400–1100 nm. The bright spot in Fig. 5 below taken after decapsulation only indicates damage within the ESD protective structure. This is revealed by the EMMI for a leakage failure resulting from +/– 3000 volt HBM stress applied to an I/O to V_{CC} pin combination. At this the die surface, there is no discoloration, no visible physical damage at the failure site.

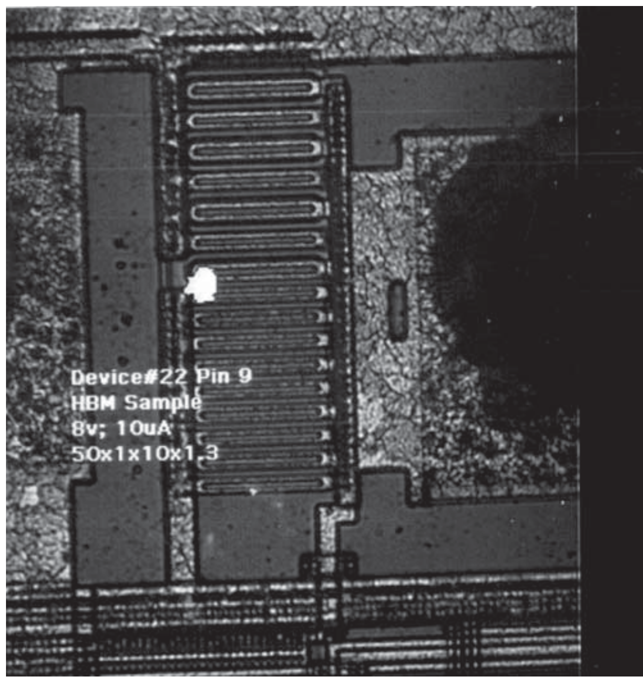


Figure 5 - *Site of leakage failure revealed by emission carriers.*

The SEM (magnification 7k×) photo shown in Fig. 6 indicates contact damage after deprocessing to the poly silicon level. Note the absence of discoloration, and the fact that the physical damage is revealed only after deprocessing.

HBM Summary:

In all cases no physical anomaly was observed after de-capsulation. SEM examination was required in conjunction with de-processing to establish physical failure site morphology and final location of the physical damage. These simulations were compared with similar damage on customer field failures (Fig. 7). The contact to contact damage shown in Fig. 7 is probably due to Icc failure after I/O to Vss or I/O to Vdd stressing.

MM ESD Discharge:

The MM or "Japanese model" described by EIAJ was in fact intended as a worst case HBM [14,15] but was found useful and developed in the US for failures which occurred due to the direct contact between the metallic leads of the device and the conducting metallic arm of robotic type equipment. Hence, the MM ESD simulations described here were conducted using a calibrated model compliant with ESD DS-5.2. Most failures occurred within the ESD protective structures (similar to HBM), but the physical damage revealed after de-processing to the poly silicon gate level was more severe as reported elsewhere [1].

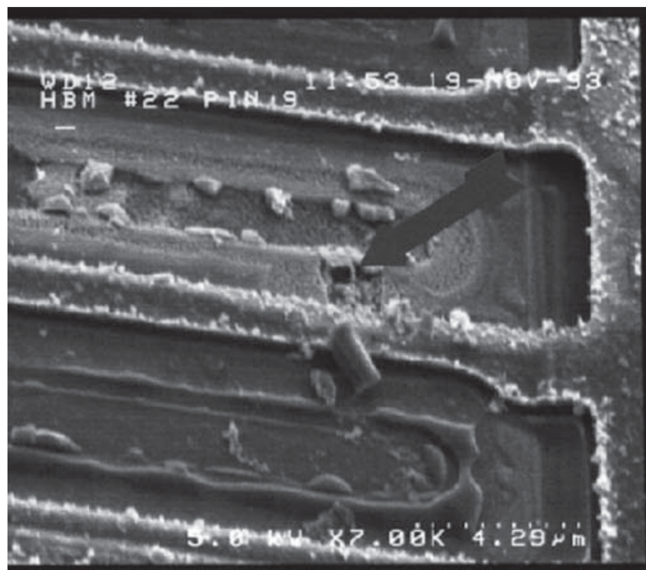


Figure 6 - *SEM of contact damage site seen in Figure 5 after de-processing to the poly silicon level.*

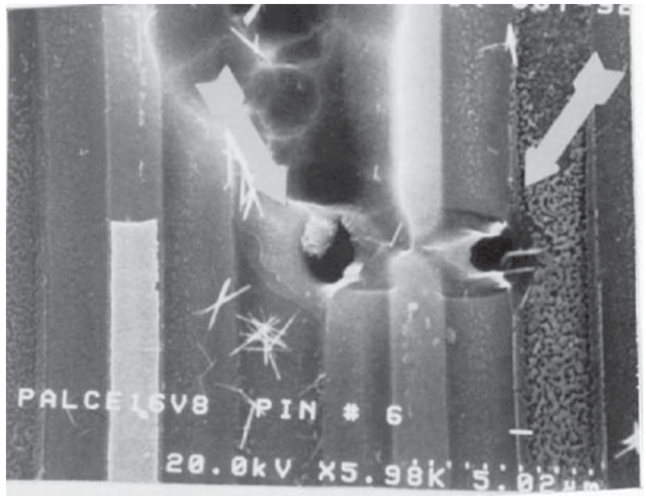


Figure 7 - *HBM. A field return showing arc-over failure between contacts in the ESD protection structure.*

Here also, the physical failures could also occur in the core of the device if the failure is functional or Icc between the Vdd (or Vcc) and Vss pins. The failing devices can also have a variety of electrical characteristics at different pins; resistive short circuit, Icc failures, functional, leakage currents and low breakdown voltages. Similar to HBM, no discoloration was observed after decapsulation, and the physical anomaly was observed after deprocessing to the polysilicon level.

These devices were also all stressed to failure to thoroughly identify the physical damage.

Although the damages observed for MM were more severe than for HBM, similar degradation of electrical characteristics were found at failing pins, with leakage currents of 1.0 to 4.0 micro-amp and very low apparent breakdown voltages of 1-2 volts. Figure 8 shows typical leakage failure current induced by MM and Fig. 9 illustrates the more severe damage where large deep pits occur at the contact/s suggesting high current parasitic bipolar action deep in the substrate.

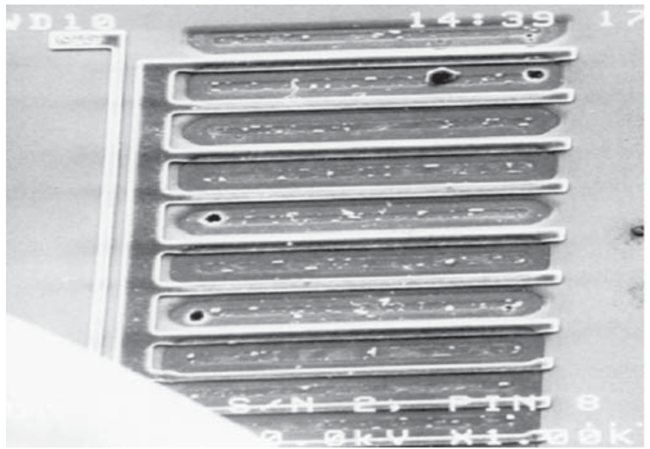


Figure 8 - MM. Severe damage at the contacts of the ESD protection structure.

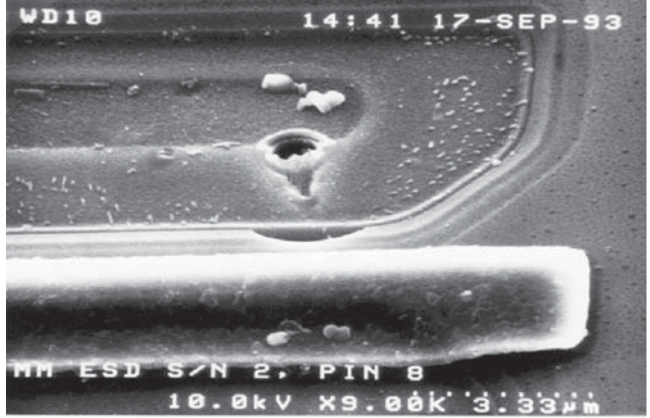


***Figure 9 -** is a close up magnified view of a typical pitted contact where the hole is very deep compared to the relatively shallow holes found for HBM.*

In Fig. 9, we see the magnified view of an ESD Machine Model (MM) failure, also from an electrical leakage signature, where the failure represents a metal to metal contact discharge. It also represents a discharge from an external source to an internal location in the device. Since the impedance of the device is high compared to the impedance of the input pulse circuitry (zero DC resistance, but single digit impedance), the event is regarded as coming from a voltage source. In contrast, the input circuitry of the HBM ESD Event has a higher resistance (1500 ohms) compared to the resistance for the ICs (below 10 ohms in most cases), and so, the HBM event is regarded as coming from a current source.

This Failure Signature was characteristic of all MM induced damage irrespective of the stress magnitude used to cause failure. Corresponding signature has been reported [18] mainly by the auto industry.

MM Summary:

The physical failure locations for MM failures are the same as for the HBM failures, and also for the same pin combinations even though the failure thresholds are different. The shape and the depth of the physical failures could be different because the energy in the MM pulse is many times that of HBM for the same voltage level.

Pin Combination Differences:

Failures can occur for all the pin combinations [19]: I/O to $V_{SS,}$ I/O to $V_{DD,}$ I/O to I/O, Vss to Vdd, and Vdd to Vss. It is established that ESD damage will occur at the core of the chip when these pin combinations are stressed. The HBM and MM stresses produced failure in the same physical location, and had similar physical shapes, but there were clear distinctions in the failure site for the different pin combination stress mode when combined with the electrical characteristics. The Input/Output I-V electrical characteristics for leakage and Icc failures showed distinct differences between the various physical failure signatures. This enabled easy identification of the stress mode initiating the failure.

I/O to Vss: HBM and MM Stresses:

Current is expected to flow through the internal circuitry (between Vss and Vdd) when I/O pins are stressed with respect to Vdd or Vss [20]. The physical damage below the die surface will occur to the NMOS pull-down output transistor devices during I/O to Vss stress. Both pin leakage and Icc failures can occur. This results in Source to Drain arcing or contact failure depending on the electrical characteristics. These two physical failure types were distinguishable because the source to drain arc-over was due to the Icc failures and the contact damage occurs for the pin leakages. This was consistent for many devices and several technologies

I/O to Vcc: HBM and MM Stresses:

The I/O to Vcc failures occurred in the PMOS pull-up output transistor devices with the same physical damage (as the I/O to Vss mentioned earlier), reflecting the distinction between leakage and the Icc failures.

Pin Combination Discussion:

Note however that the I/O to I/O leakage failures produces physical damage in the contacts for both the pull-up and pull-down transistors. No Icc failures have been observed for I/O to I/O stressing for any of the devices stressed and analyzed.

The Vss to Vdd (or Vdd to Vss) stresses predominantly caused failures in the core logic of the device in the form of Icc failures which showed arc-over damage (Fig. 10) from Source (square structure) to Drain (rectangular structure) at an internal location far from the ESD protection structure and far from the input buffer circuitry. The latter location will be shown (later in this text) to be associated with CDM failures from the CDM events. For the devices analyzed here, the HBM and MM failure sites in the core were the parasitic lateral npn "structures" formed between the Drain of the NMOS device connected to Vdd BUS and the Source of another NMOS device connected to the Vss BUS. The completely deprocessed device in the figure all the way down to the silicon level shows the arc-over physical damage quite clearly. We note that a good power supply clamp would reduce the failures that would occur for the pin combinations: I/O to V_{CC}, I/O to $V_{SS.}$, Vcc to Vss and Vss to Vcc.

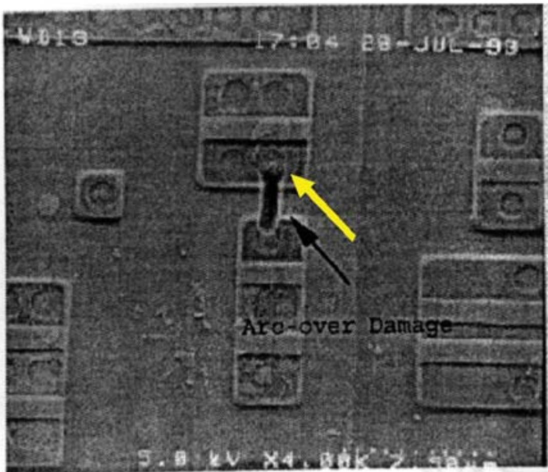


Figure 10 - ESD Arc-over at the core of the device and occurring between Source and Drain for Vss to Vdd stress.

Table 1 is a summary of the electrical and physical morphological signatures associated with the different pin combination stress modes. Note for example that the electrical signatures of leakage and Icc failures can occur for the pin combination of I/O vs Vss but the physical signatures are quite different. Note also that the electrical signatures of leakage for I/O vs Vss produce contact/junction damage in the pull-down transistor, but occurs in the pull-up transistor for I/O vs Vcc. Note finally, that the electrical failure if Icc occurs for Vss to Vcc, and also for the reversed Vss to Vcc. The physical failure location (the core) is the same for both, but the beginning and end of the physical failure can be distinguished- S to D from Vss to Vcc and D to S from Vcc to Vss.

Pin combination	Failure type	Damage/location
1. I/O to Vss	Leakage	Contact/Junction in the pull-down transistor
2. I/O to Vss	Icc	Arc-over at pad to contact in pull-down transistor
3. I/O to Vcc	Leakage	Contact/Junction – pull-up transistor
4. I/O to Vcc	Icc	Arc-over at pad to contact - Pull-up transistor
5. I/O to I/O	Leakage	Contact /Junction – pull-up or pull-down
6. Vss to Vcc	Icc/func	Source to Drain arc-over in the core of the device
7. Vcc to Vss	Icc/func	Drain to Source arc-over in the core of the device

CDM Discharge. Exposure to CDM stress is carried out using a "dead-bug" socket-less configuration with air discharge. The setup results in charging by field induction as described by Renninger et al. [21] and conforming to ESD DS-5.3. The CDM method is greatly influenced by parasitics in the tester setup even though the stray capacitance of the socket is absent and series inductance to the RF ground plane is minimized because a relay is not used. This also allows for discrimination in the influence from the different package types (CDIP, PDIP, PLCC, LCC, PGA, BGA, TSOP and SSOP etc.) with respect to the parasitics in the packages. This is quite different from SDM (Socketed Device Model) where the process is greatly influenced not only by the parasitics in sockets but also by the relay network etc of the simulator [22].

The effect of variation in air arc resistance is only partly minimized by controlling the local humidity and using multiple zaps (<5 per polarity) per pin, but even then .the humidity changes have a profound effect on the variation in the resistance of the arc. Values between 25–50 ohms for the air-arc resistance have been quoted and used. The peak current, Ip, is dependent not only on the ohms law (V = IR), where V is Voltage, and R is the total resistance in the circuit, but is also affected by the parasitic capacitance and inductance of the tester.

The need to minimize these parasitics is a major issue, since they also cause the overshoot and ringing in the shape of the CDM waveform. In all cases, the CDM failures occurred in the gate oxide at the first input buffer. This buffer circuit is located internally beyond the (HBM & MM) ESD protective structures (see Fig. 11). This is characteristic of the CDM failures. Note that while the damage site occurs most at the NMOS pull-down, the PMOS pull-up is not immune to the CDM damage. It is to be noted that this differentiation between the CDM and the HBM/MM failures is significant.

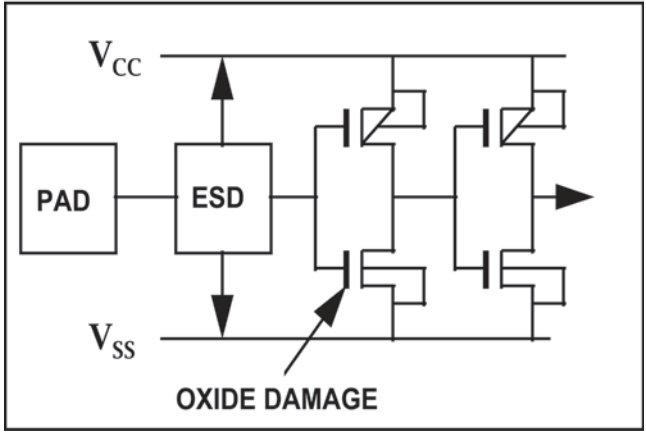


Figure 11 *- Location of CDM Damage in Input Buffer Circuit.*

Failed pins can show enhanced leakage currents from 500nA to 100 micro-Amps, severely degraded breakdown voltage, but could still be functional when hand tested (taking full ESD preventative measures). Figure 12 is an optical photo (before CDM stressing) indicating the location of multiple data lines feeding into input buffers immediately below (see arrows in Fig. 12).

For Fig. 13 at very low optical magnification, this characteristic photoemission is seen at an input buffer after exposure to multiple ESD discharges to the pin with the package leadframe charged to +/– 1000 volts with the CDM ESD simulator. All CDM ESD stressed input buffers showed an emission at this same location for the same pin designation. No physical damage is visible at this low magnification and there is no discoloration, as would be expected if the damage was the result of an EOS failure.

After deprocessing (removing the several top layers of metal and interlayer dielectric glass) to the polysilicon gate level, some perforation could just be discerned. This is shown in Fig. 14 at the gate oxide edge and possibly in the gate oxide below, which is not yet removed in the Fig.14 image. There is no discoloration, and as you can see, this subsurface damage was not visible in the photo of Fig. 13 nor in Fig. 12. This is to be expected. In general, ESD damage far below the surface and down to the silicon level is not visible. This is in contrast to the EOS failures which in almost all cases occur at the surface of the die and is discoloured. The two arrows point to the "round" holes in the oxide below the gate. This input buffer location is definitely the result of the CDM ESD type failure, resulting from charge build up in the device and discharging to an external lower potential point. Typically, this is a metal to metal discharge.

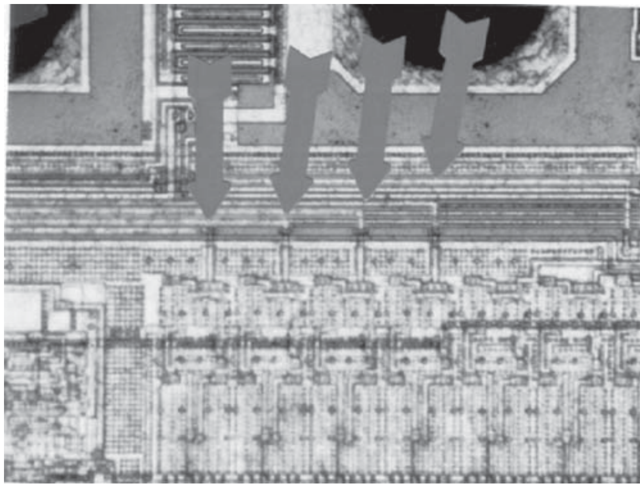


Figure 12 *- Optical photo showing location of the multiple data lines (horizontal) feeding into buffers below.*

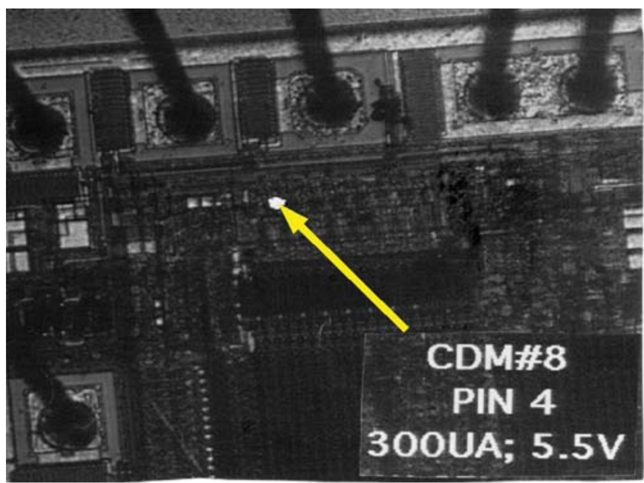


Figure 13 *- Photoemission from beneath the metallization at the input buffer damaged by CDM stress. There is no visible damage from a low magnification optical view.*

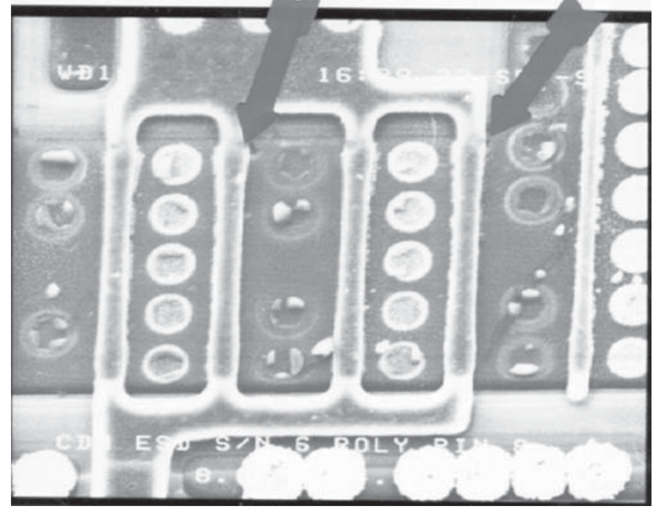


Figure 14 *-CDM. Perforations at gate oxide edge just visible after deprocessing to polysilicon level*

After further stripping to the substrate, pinholes are clearly indicated by pits in the silicon surface (Fig. 15), all of which are below the gate oxide. Most of the gate oxide has been stripped during the many deprocessing steps. This is clearly different from the HBM and MM failures which occurred in the contacts (see Fig. 6 and 7). These CDM failures had an internal to external mechanism.

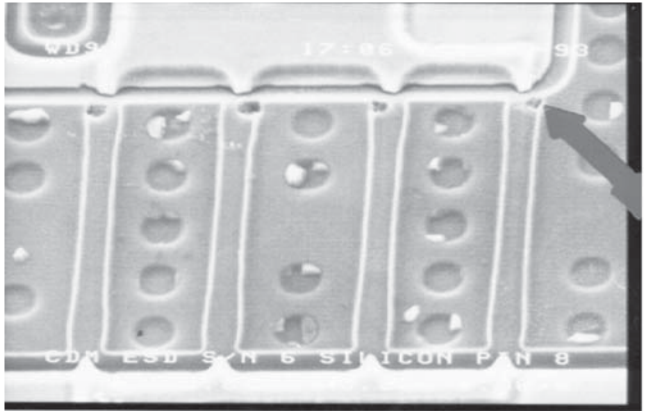


Figure 15 *- CDM. Pits in the silicon substrate showing clear evidence of gate oxide damage/rupture in the input buffer circuitry.*

Figure 16 shows a typical CDM failure corresponding to a customer field failure after de-processing to the poly silicon level. Again, there is no discoloration which is typical for ESD type failures.

Reports of leakages as low as 100 nano-Amp for single zap events have been reported [23]. In our laboratory simulations the oxides show multiple perforations or are made up of more complex shape probably due to the use of the prescribed multiple zaps. The CDM ESD event for field or factory failure can be described by a fingerprint consisting of enhanced leakage (at well below specification levels of 1–10 microamps). Such leakage is confirmed when the data-in line is laser cut just beyond the HBM protective structures at the bond pad. This confirms that the leakage is at the input buffer circuitry.

Figure 16 *- Typical customer return showing a single small hole in the gate oxide.*

CDM Summary:

We note in summary that there is obvious difference between CDM ESD and the other two (HBM and MM) ESD models but the distinction between HBM and MM is not clear cut so simulation must be done. The CDM physical failure location is at the input buffer and in the gate oxide, where as both HBM and MM failures occur mostly in the contacts at the input protection structures. Neither do CDM failures exhibit arc-over type damages.

Is it an ESD, EOS or ESD -induced EOS Failure?

These failures can occur at anytime and are especially prevalent at burn-in. An EOS failure that cannot be directly duplicated using one of the "in-house" EOS models or by SLU or by TLU is prime suspect. However, should all EOS failures be checked as a possible ESD induced failure? This may be a judgement call.

There is some distinction. The suspect ESD induced EOS failure which should be clearly visible at low magnification is very carefully deprocessed with visual and/or SEM examination at each step. If it is a pure EOS failure, the subsurface will be devoid of damage [24]. The extent of any physical damage beyond the metal layers will be at the dielectric layer below. The thermal nature of the very slow EOS pulse does not extend to the subsurface. Recall that a pure ESD failure has no visible damage at the top surface and there is no discoloration [1, 24].

If it is an ESD induced failure, then further deprocessing will reveal damage down to the silicon level. The ESD damage would have already been present in the subsurface. Due to the very short pulse width (< 10 nsec) and very fast rise-time (< 500 psec) of the ESD pulse, the ESD event ruptures the oxide first. Additional stress can cause or induce an EOS failure in the layers above. SEM examination and some cross-section will reveal additional damage below the surface. This has been demonstrated and also published by several authors including [25].

EOS Simulations:

A number of laboratory simulations can be devised and executed to try and replicate the major classes of EOS damage seen for customer returns and in factory and reliability rejects. Successful replication is described here for gross EOS damage evidenced by package distortion. This allows ready distinction between incorrect insertion (for both PDIP and PLCC) and latch-up related failures, and for failures occurring most often on high voltage input programming pins.

Electrical Failures:

These failures ranged from simple leakage to shorts (Fig. 17), low voltage breakdown, Icc and functional.

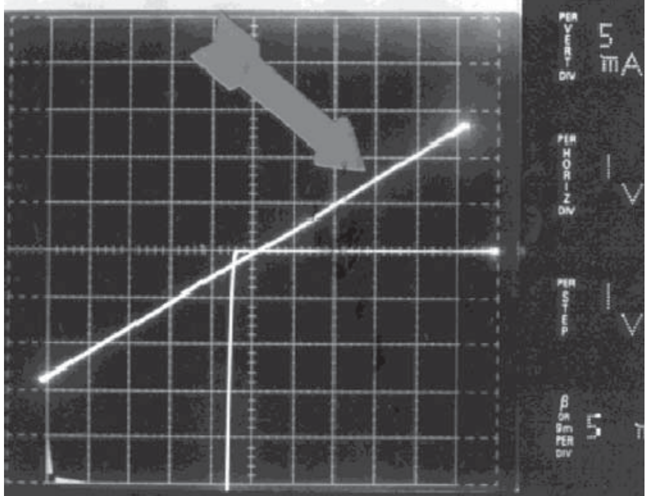


Figure 17 - *Degraded I-V curves showing sloped line for short failure (arrow) and good breakdown (horizontal line).*

Physical Failures:

“Burnt” Plastic. A class of failures showing damage to the package was characterized by “burnt” or carbonized plastic adhering either to the die surface shown in Fig. 18 or to the bond wires shown in Fig. 19, suggesting perhaps two different failure modes.

Bond Wires. This is the first and simplest class of EOS returning from the field or from the factory, the test area, or from the reliability lab. A curve tracer can be used to supply over-voltage or over-current to the pins resulting in EOS (Fig.18).

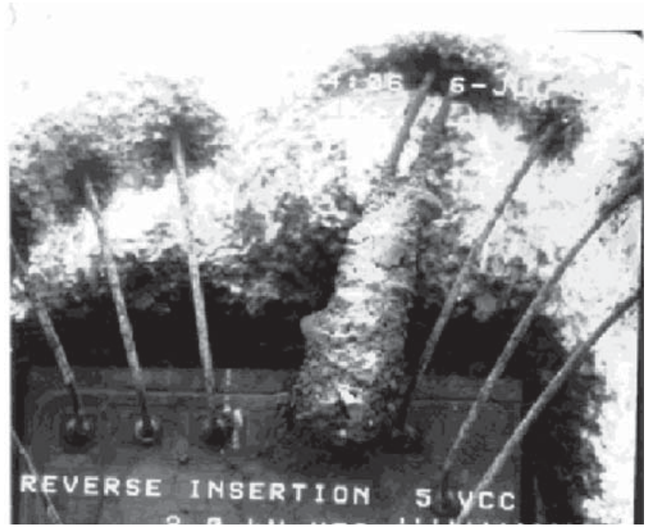


Figure 18 - *Carbonized plastic surrounding the Vcc dual bond wires and adhering to the die surface of a PDIP packaged device.*

Figure 19 - *Fused open Vcc bond wire from PDIP misinsertion (RVI).*

Incorrect Insertion. Use is made of dynamic burn-in boards to simulate possible misinsertion. Here the device is powered up and the outputs are pulled high by a resistor to V_{CC}. This EOS simulation of reverse insertion (RVI) for PDIPs and of wrong orientation (WRO) for PLCCs revealed a Failure Signature of fused open V_{CC} power and/or V_{SS} ground bond wires with “Burnt” plastic around the damaged bond wires and/or on the die surface.

For PDIP reverse insertion (180 degrees only), I_{CC} failures occurred in times as short as 10 minutes at 25° C. However, in all cases the die was still functional when micro-probed at the bond pads and no metallization damage was observed. Under these conditions, this die functionality is a distinguishing feature and serves to differentiate between this regular and well-known EOS failure from the latch-up type failure. If probing the pad had resulted in an Icc current anomaly then further and a different simulation would have been necessary to confirm the likely hood of it being a latch-up type failure.

The photo in Fig. 18 illustrates the typical example of carbonized plastic adhering to the Vcc (Vdd) or Vss bond wires and to parts of the die surface.

For Fig. 19, it illustrates the typical fused open Vcc (Vdd) or Vss bond wires with some carbonized plastic on the bond wires. The PDIP mis-insertion caused the location of the Vdd and the Vss pins to be in the wrong place. This is especially problematic when we consider the use of higher voltage and the current requirements for the power supply pins.

For the PLCC packaged device, incorrect insertion can result in misorientation or wrong orientation (WRO) by 90°, 180° or 270°. For the 180° position for the PLCC device , the Failure Signature is identical to the 180°, position for the

PDIP package, but there were no electrical failures including functional testing after 10 minutes at room temperature (25 °C). There were no visible failures either using the optical microscope. All the PLCC units failed however after 24 hours at 125° C. Presumably, the difference is explained by the difference in thermal impedance of the package/socket combination.

Figure 20 shows fused and open V_{SS} bond wires for the PLCC package. The temperature of the device and package was so high that the wires burnt until they became open. It is also possible that the longer time associated with the Icc flowing in the wrong/reverse direction in the Vss and Vcc pins exacerbated the situation.

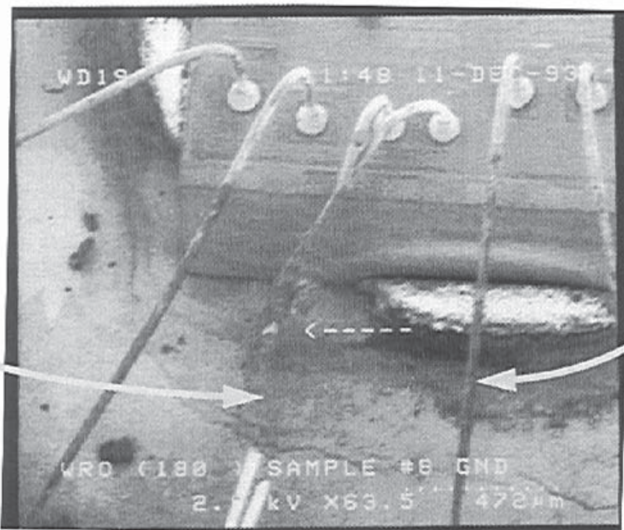


Figure 20 *- Fused open Vss bond wires for PLCC.*

For the 90 and 270° WRO positions, none of the samples failed and negligible Icc current (< one micro amp) was drawn from the supply. This negligible current may not normally be detected during burn-in and could lead to erroneous conclusions. Similar type failures (occurring at Burn-in at the customer site) had returned from the field. These devices had passed the JEDEC type EOS stress, called static latch-up (SLU). They passed/survived +/- 100 milliamps at 125 °C using the slow usec type square pulses. It is surmised that during burn-in, or while the devices are powered, they experienced a transient type pulse (nsec pulse width ranges).

The transient pulse simulations do show similar type failures in addition to internal core EOS failures (Fig. 21) which are not replicated using the incorrect insertion simulation. Note that this EOS failure occurred on the surface of the die, it is visible at low magnification and the physical failure shows discoloration. This is a clear distinction between EOS and ESD failures. However, Static and Transient Latchup simulation is beyond the scope of this tutorial and is detailed elsewhere [26].

Figure 22 represents a burn-in failure returned from the field. Here the carbonized plastic is adhering to the die surface, covering all of the Vcc or Vss bond wires and bonding pads. This carbonized material is very difficult if not impossible to remove, so the determination of the exact nature (root cause) of failure is hard to determine. As you can see, the Vss and/or the Vdd pads are not accessible. Hence the Icc (as measured on the Vcc and Vss bond pads) cannot easily be determined. This visible top layer/level damage to the plastic package and die surface.

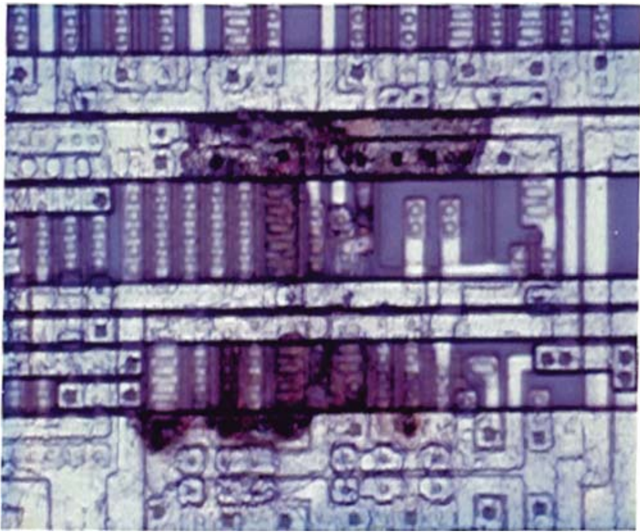


Figure 21 *- Internal core EOS failure showing melted metal after transient (nsec pulse width) latch-up. Note the discolouration of the failed area (if photo is not black and white).*

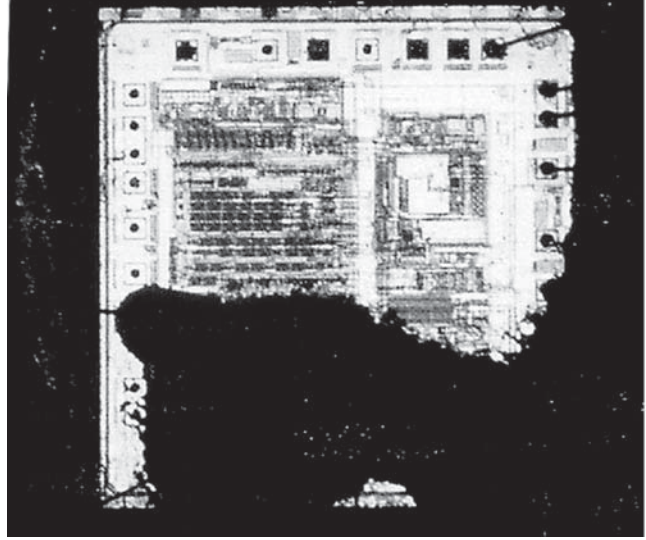


Figure 22 *- EOS. Heavy carbonization on bond wires for failure during actual burn-in.*

Discolored Metallization:

The second class of EOS customer returns showed spiked and discolored metallization at the source and/or drain of the protective transistors adjacent to the bond pad. This is particularly prevalent for the HV programming pins suggesting over-voltage, which can be detected during the routine Monitor Program in the factory or reliability lab.

Programming Over-Stress:

A Data I/O Unisite programmer can be used to investigate variation of programming parameters for both PDIP and PLCC packages. Varying the programming pulse width did not cause failure for these devices (but should not be ruled out). However, increasing the Vpp (above the nominal 9.0V) produced failures at between 11.5 and 15.0 volt caused pin failures for all units and the circuitry of adjacent pins was also damaged in some cases.

De-capsulation revealed that during simulation the dominant failure signature was replicated, that is, spiked and discolored metallization (Fig. 23).

Figure 24 is an optical microscope picture which is less that 1000× magnification and shows a field return where the EOS damage are in the transistors and at the die surface. Discoloration is obvious (color photo necessary) and the failure is observed under low magnification.

Figure 24 shows the EOS damage with clearly visible discoloration at the ends of the drain fingers of the input protection transistors adjacent to the bond pad. These EOS type failures have been the dominant type failures returning from the field.

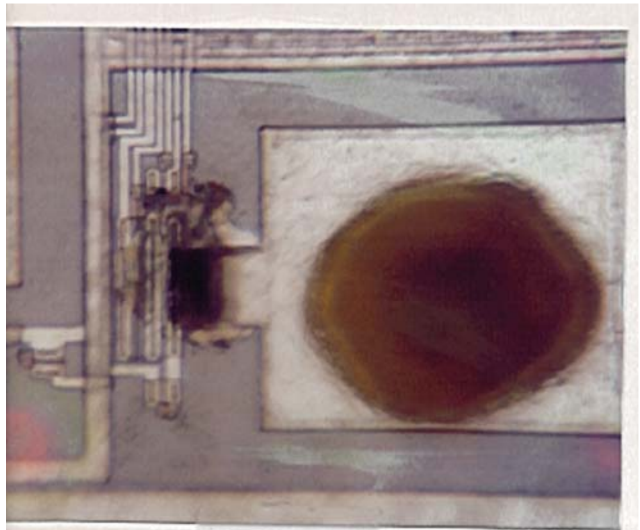


Figure 23 *- Simulated EOS damages with clearly visible discoloration (color photo required) on the transistor metal lines.*

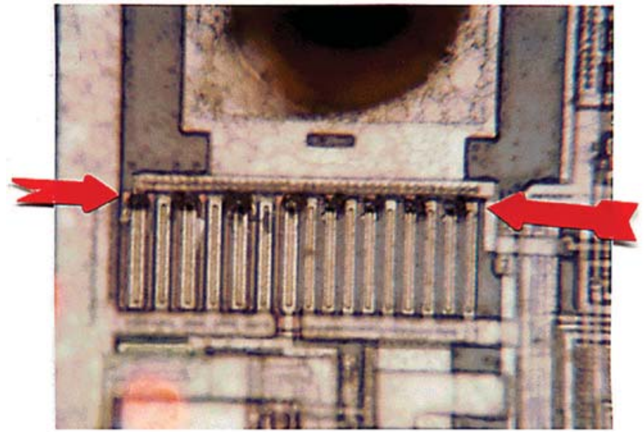


Figure 24 *- Field return where the damage is in the transistors and at the die surface.*

Reverse Breakdown (BVDSS):

The input ESD protection structure shown in Fig. 25 below consists of two large NMOS transistors (Weff=45micron).

One transistor in Fig. 25 is connected to VSS and the other to VCC with the gates connected to the substrate ground via a high by a value resistor in order to initiate bipolar action by dynamic gate to substrate voltage. A standard curve tracer (Tektronix 575) and a manual power supply can both be used to apply an over-voltage to the input pad to force the input transistors into BVDSS mode breakdown, which occurred at 12-13 Volts. Failure was immediately apparent; the electrical characteristics showed opens, gross leakage and shorts.

The characteristic spiking and discoloration visible after decapsulation is shown in Fig. 26 for both terminated NMOS pull-up and pull-down transistors with no apparent damage to the adjacent pins.

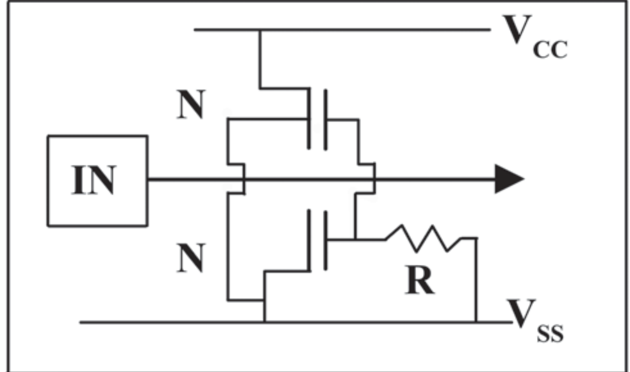


Figure 25 *- Input –only circuitry using a terminated “dummy buffer”.*

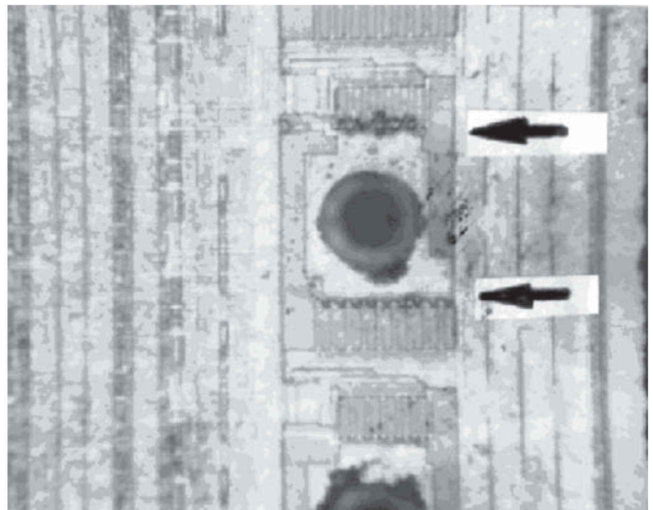


***Figure 26** - characteristic spiking and discoloration (discolouration is visible only if coloured photographing is used) for both terminated NMOS pull-up and pull-down transistors.*

It is apparent that the failure mode dominating these failures is due to over-voltage of sufficient magnitude that the protective transistors at the pad are damaged by being forced into BVDSS mode with no current compliance limit.

Since this failure mode is absent as a return for product which are programmed in the factory, it is evident that many customer sites have excessive voltage on their programmers (or programming equipment). This can be minimized by careful selection of a power line conditioner, routine calibration of voltage levels, verification of the correct programming algorithm, and by maintaining the condition of the programmer sockets. Routine periodic examination of the waveforms during programming using a wide band width oscilloscope to detect over-shoot will also prove beneficial.

Curve Tracing for Simulating EOS:

A simple simulation utilized the Curve Tracer to apply a high current or voltage to the input pin. At a voltage or current which is beyond the manufacturers rating of the device, the device failed. This is revealed by the melted metallization (Fig. 27) which is observable at low optical magnification and which shows discoloration in the damaged glass and metal. The EOS discoloration is an important distinction and separates the EOS failure types from an ESD type failures.

Discussion:

Failure signatures between HBM and MM show some variation in morphology but they are located in the same ESD protection structure. The failure for MM is more severe as is evidenced by the much lower (5-20X lower) voltage required for the same failure threshold. A clear distinction is achievable however after a combination of simulation

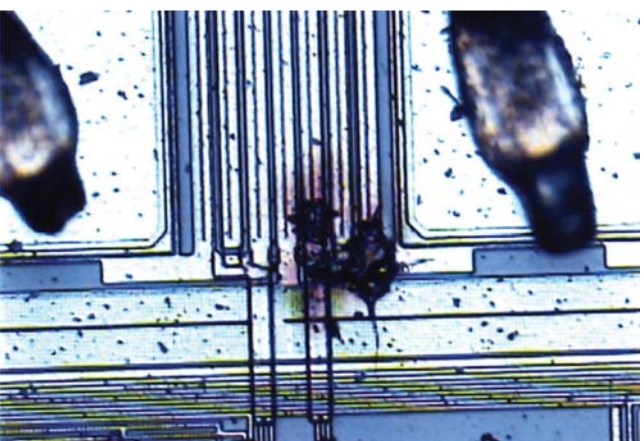


***Figure. 27** - EOS failure using the curve tracer to supply over voltage or over current*

and failure analysis because the waveform from the input pulse (which contains rise time, voltage, peak current, and pulse width) used is quite different for each ESD event. The energy (power) of the MM pulse is much larger than that for HBM using the same input voltage for both so the failure is more severe. .

The physical failures for CDM occur at the input buffer circuit and are therefore quite different in location from that of HBM and MM. Even though both FIM and CPM type failures have exactly the same signature/fingerprint and are just subsets of the CDM event, it is necessary to distinguish between FIM and CPM as follows: For Ceramic Packages: FIM failures are the most likely due to the fact that tribocharging these ceramic surfaces are quite difficult, but possible. It is much easier to induce a charge on the metallic lead frame while the ceramic packaged device is in a field. This can occur quite easily when the devices are in an uncontrolled automatic piece of equipment.

For Plastic Packages. CPM is most likely because of the propensity of the (insulating) plastic package to accumulate charge upon its surface from tribocharging or contact electrification, resulting in induced charges on the lead frame of the device. However, a device in plastic packages can also fail due to FIM as a result of the package being in an electric field. Here, the charges will be induced on the leadframe. Recall that charges on insulators (plastic packages) are not mobile. It is seen then that the nature of the failure environment is (always) of utmost importance.

Summary:

The various test "models" outlined here result in characteristic Failure Signatures that we have described. They are shown to replicate the morphology of actual factory and customer returns.

Clear correlation is shown for FIM and CPM, which also occurs in the field as well as in the factory. A sound low leakage screen in the test program will keep such damaged product from leaving the test floor and alerts staff to a possible problem.

Simulation then shows that some field returns are due to misinsertion with a Failure Signature that includes fused or open Vdd/Vss bond wires, burnt plastic on the damaged wires, but no metallization damage on the die. Transient Latch-Up simulation using capacitive discharge allows differentiation of a similar Failure Signature, but where burnt plastic adheres to the die in the region of the VCC/VSS bond pads. Damage is found to the metallization upon decapsulation and to the underlying structure upon further de-processing. Mis-insertion suggests the need for more care in customer handling methods, whereas the evidence for latch-up suggests the existence of power system faults of a transient nature that can be eliminated using suitable line conditioners and filters.

Conclusion:

It is shown that a distinction can be made between EOS (discoloration) and ESD (no color) failures and between the characteristic Failure Signatures produced by the various ESD models. This can then be used to correlate between actual factory and customer field failures. It is imperative then that the major EOS and ESD models be available to a Device Analysis facility since replication of Failure Signature is seen to be a powerful tool in determining probable cause of failure.

ACKNOWLEDGEMENTS

The author would like to extend special thanks to the authors of the original 1994 paper and to the rest of the engineering and technician staff. A lot has happened since then, but they were part of an original team trying to establish a clear distinction between ESD and EOS.

REFERENCES

1. Leo G. Henry, T. Raymond, Mahanpour, I.H. Morgan, “EOS & ESD Lab. Simulations and Signature Analysis of a CMOS Prommable Logic Product”. ISTFA, 1994. LA, CA., p117.
2. I.H. Morgan, Proceedings of the 3rd ESD Forum, Grainau, Germany, Dec. 1993. P27-33.
3. C. Cook, and S. Daniel, Proceedings EOS/ESD Sympo. Dallas, Texas, Sept., 1992, p 149-157.
4. B.L. Euzent, T.J. Maloney and Donner II, J.C., Proceedings EOS/ESD Symposium, Las Vegas, Nevada, Sept., 1991, p 59-64.
5. Microelectronics Failure Analysis Desk Reference, 3rd Edition, 1996. Editors- T.N. Lee and S.W. Pabisetty.
6. D.C. Wunsch. and Bell, R.R,. IEEE Trans, Nucl. Sci, 1970.
7. Leo G. Henry, Hyatt H., Barth, J; Stevens, M. and Diep, T., Charged Device Model Metrology. Proc. 18th EOS/ESD Symposium, Orlando, Fl., Sept., 1996, p167-179.
8. International Standard: IEC-1000-4-2. 1996. Electromagnetic Compatibility. Part 4. Testing and measurement. Section 2. ESD Testing. Note: Previously named- IEC-801-2.
9. J. T. May, and J.F. Guravage, Proceedings ISTFA, Los Angeles, CA, Oct., 1990, p 143-147.
10. L.R. Avery, EOS/ESD Symposium, Orlando, Florida, Sept., 1987, p 186-191; & p 88-92.
11. P.R. Bossard., R.G. Chemelli, and B.A. Unger, Proc. EOS/ESD Symposium, San Diego, CA, Sept., 1980, p17-22.
12. T. Raymond, K.L. Chang, and Leo G. Henry AMD 3rd Engineering Conference, Marriot Hotel, Santa Clara, CA, Feb., 1994
13. E.J. Chwastek, Proceedings of EOS/ESDS Symposium, New Orleans, La Sep, 1989, p149.
14. ANSI/ESDA/JEDEC, Joint HBM Standard JS-001,2010.
15. ESD Assoc. Standard DS-5.2, 1994, revised 2011.
16. ESD Assoc. Standard DS-5.3.1, 1994, revised 2010.
17. MIL-STD 883, Method 3015.8, March,1989, revised 2010
18. 18. A. Kelly, G. Servais, T. Diep, D. Lin, S. Twerefour, G. Shah, “A Comparison of ESD Models and Failure Signa for CMOS ICs.” Proc. EOS/ESD Sympo, Phoenix, AZ, 1995. P175.
19. Leo G. Henry, Failure Signatures Associated Pin Combinations and the Core of the CMOS Device. AMD Internal Reliability report, 1996.
20. A. Amerasekera, C. Duvvury. “ESD in Silicon Integrated Circuits”. John Wiley, 1 London, 1995.
21. R.G. Renninger, Jon, M.C., Lin, D.L., Diep, T., and Welsher, T.L., Proceedings EOS/ESD Symposium, New Orleans, Louisiana, Sept. 89, p 59-71.
22. M. Chaine, L. Avery, K. Verhaege, Leo G. Henry, H. Geiser, M. Farris, T. Bodbeck, T. Meuse, K. Bock. “Investigation into Socketed CDM Tester Trans. Line Characteristics”, Proc. EOS/ESD Symposium, Reno, Nevada, Oct.98.
23. H.A. Gieser, P. Egger, .M.R. Herrmann, J.C. Reiner, and A. Birolini, ESREF Proceedings, Bordeaux, France, 1993.
24. Leo G. Henry and J.M. Majur, “Basic Physics of Color Coded EOS Metallization Failures”

25. D.S. Kiefer, R.T. Milburn, and K. Rackley, “FA of EOS/ESD Damage to HCMOS Gate Arrays,” Proc. Of 15th ISTFA Conf, LA, CA p201, 1989.

26. Leo G. Henry “EOS” and ESD Laboratory Simulations for Failure Signature Analysis”. Tutorial-R at EOS/ESD Symposium, Reno, Nevada. October, 1998.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

The Power of Semiconductor Memory Failure Signature Analysis

Cary A. Gloor
LSI Logic Corporation, Ft. Collins, Colorado, USA

Abstract

It has long been recognized that semiconductor memories are superb drivers for process yield and reliability improvement because of their highly structured architecture and use of aggressive layout rules. [1] This combination provides outstanding failure signature analysis possibilities for the entire design, manufacturing, and test process. This includes areas of study such as design and layout robustness, baseline yield analysis, maverick lot analysis, process and product reliability qualification, lot acceptance testing, and customer return failure analysis (FA). These areas of study are important whether the memories are stand-alone devices or deeply embedded within an Application Specific Integrated Circuit (ASIC).

Introduction

The signature analysis process includes five key disciplines that need to be orchestrated within the organization:

1. Design For Test (DFT) practices
2. Test floor data collection methodology
3. Post-test data analysis tools
4. Root cause theorization
5. Physical Failure Analysis strategies

These areas require cross-functional cooperation and management support in order to accomplish the desired objectives of improved time to root cause with high success rates.

Design For Test Practices

Overview

Modern circuit designs contain many different DFT solutions. It is critical to have a well defined roadmap for DFT practices. Good planning, communication, and buy-in within the organization is essential to it's success so that down stream engineering groups can manage the technology changes that will be presented to them.

Evolution of DFT

For embedded memory applications, the industry has seen an evolution that included direct access of memories using parallel vectors, to testing with embedded microprocessors, to scan-based testing using serial patterns, to Built-In-Self-Test (BIST) solutions, and more recently to programmable BIST. This evolution required many changes in the data collection and analysis methodology along the way. Each DFT solution presents challenges for controlling and observing the memories. The use of redundancy, Built-In-Self-Repair (BISR), and error correction also adds to the complexity of observability and requires special attention during the development phase. BIST is now the norm for embedded memories and can provide an important standardization to the testing and FA strategy.

A classical memory fault model

As a minimum, a classical memory fault model should be used to determine the fault coverage obtained. [2] This should include the following:

- Address fault
- Stuck-at fault
- Bridging fault
- Transition fault
- n-cell coupling fault
- Delay fault
- Retention fault

There are many tests on the order of N or $N^{3/2}$ in length, where N is the number of logical addresses of the memory, that provide good fault coverage to this model.

Test pattern nomenclature

There are three independent pieces of the test pattern description, namely, the test algorithm, the background data pattern, and the address sequence. One test that satisfies the demand for very high fault coverage and minimal test time is a march algorithm, and will be used by way of example. The definition of a march is this: A test of length order N, usually between 10N and 17N, that detects all victim bit cells but does not necessarily locate aggressor operations, for all address, stuck-at, bridging, transition, and 2-cell coupling faults. An aggressor operation is a memory access that activates one or more bit cells into a failing state. The trade off for having a

shorter test is the lack of ability to locate aggressor operations that other $N^{3/2}$ or N^2 algorithms can achieve. There are many variations of march algorithms and one example is:

+(W) +(Rwr) +(rWR) -(Rwr) -(rWR) +(R)

Algorithms are made up of sections called phases. This example contains six phases denoted by "( )". Each phase is a set of memory access operations that are done address by address over the specified address space. The address sequence is loosely specified by an increasing or decreasing symbol such as "+" and "-". This address order is independently specified and is typically either a binary row-fast or column-fast sequence, but can take on many other variations. Row-fast, for example, means that the portion of address bus dedicated to the row decoding is completely cycled through before any of the column decoding portion of the address bus is switched.

The second phase in this example is +(Rwr) which means that a read operation of true data is followed by a write operation of complement data, and then followed by a read operation of complement data. Thus, upper case denotes true data, and lower case denotes complement data.

This leads to the third and final part of the test pattern definition, namely, the definition of the background data pattern. Notice that the algorithm simply specifies data as true or complement, and not as logical ones and zeros. Data can generally be defined as a function of the logical or physical address, and Boolean equations are used to express it. A checkerboard data pattern might follow the Boolean equation $D=A3 \oplus A0$, where $\oplus$ represents the exclusive OR logical operator. In this way logical states can now be specified for any bit cell position at any point in the test algorithm.

This march algorithm can also inherently accommodate retention stalls because of it's design. Stalls can be inserted after the second and third phases where the background data is complement and true respectively.

Bitmaps and scramble specification

In order to bitmap a memory, a transformation of the logical address and I/O information to the physical row and column is needed. A mapping of the logical address bus to the tester address resources is first made. For example, A[9:0] may map to X[6:0]Y[2:0] where X represents the row decoder bits and Y represents the column decoder bits. Then, the address scramble Boolean equations can be written to transform X and Y into physical row and column locations. In a similar way, the data scramble Boolean equations can be specified to account for architecture issues like bit line twisting or bit cell rotations, depending on the desired frame of reference. This may be with respect to the overall bit cell, or with respect to various layout nodes within the cell.

Design aids

There are a host of other features that can be added to the layout to order to aid in the FA effort. These may include top-layer mechanical probe pads on key nodes. Also, test structures fit into the scribe streets can provide important clues especially early on in the process development phase. Various test modes for circuit characterization, test time reduction, and burn-in test modes may be included. Design Of Experiment (DOE) such as varying instance sizes and layout orientations may be evaluated. It is desirable to have all eight orthogonal instance orientations available on memory test chips. In an ASIC product, some of these DOE's are inherent in the layout.

Test Floor Data Collection Methodology

Overview

Test floor data collection must comprehend the variety of tester platforms used, data file formats, data sampling plans, and file management concerns. Integration includes both in-house and third party hardware and software aspects. The Test Engineering groups must efficiently handle the test program development and code insertion requirements for this data collection.

Tester platforms

On memory testers, bitmap dumps can be captured and saved for later processing. Some of these can be quite large and file compression techniques may be needed. On logic testers, native data logs are easily captured and require minimal test program code development. Every platform will have different issues regarding capture memory architecture, data format, and the speed of data capture realized.

File formats

Custom file formats may be chosen to help standardize data files coming off of the test floor's various platforms. An ASCII comma separated values (csv) format works well because of it's readability and structure, but a binary format may be chosen as well. File content starts with meta data like test hardware setup information. Then specific information about the lot, wafer, and die are added. Finally, the raw data for each test condition is written to the file. Each of these data types can be structured as a uniquely formatted record or object. Data converters may be necessary to be able to read the file as input to the various data analysis tools.

Data sampling plan

A good data sampling plan is very important in the data collection methodology. Interactive control of at least three key aspects is desirable. First, the ability to turn data collection on and off for a production run is important. Second, a plan to data log specific die or memories within each wafer is needed, and this could be every third failing die for example. Finally, control of the amount of data collected per test pattern is needed. This is especially important if the data is coming off of a logic tester rather than a memory tester. In this way a conscience decision is made to add a limited amount of cost to production test in order to make yield enhancement more efficient through statistical data collection.

Several other areas of data sampling may be added. For example, a limit on the total test time overhead added by the data collection is useful. The data collection may be turned off if more than a specified time per wafer has been added. Further control can be added to customize the data collection for experimental lots coming from the fabrication process, or for specific wafers targeted with additional in-line process inspection data.

File management
File management consists of four issues: distribution, compression, archival, and conversion. As the data comes off of the tester, a central data repository is needed to hold it. From there, users may pull files as needed, or scripts may push the data to specific servers in the user community. File compression should be used to help manage the often large files produced. Data archival should be triggered, for example, when the file is four weeks old. Scripts to provide data restore may be provided to the users as well. In some cases data conversion is necessary for the various formats required by the analysis tools that have been chosen.

Test program code
Code insertion scripts can be very useful if data logging is routinely added to test programs for new devices. Perl scripts provide an efficient way to insert source code and library calls to these functions. This should include not only the data collection code, but also the test program flow options. For example, data collection may be done both before and after a Vdd stress is applied to the unit, or during sensitivity analysis such as minimum and maximum applied Vdd test conditions.

Post-Test Data Analysis Tools

Overview
Post-test bitmap tools must be sophisticated enough to provide high quality data analysis and mining. They need to include automatic failure mode recognition software for defect classification [3], data analysis tools based upon statistical methods, and powerful interactive data mining features. (See Figure 1). The wealth of signature information available is at the heart of the FA effort, and is often underutilized or misunderstood. A combination of manual and automatic data mining must sort through this information to improve the path to root cause identification.

Bitmap tool evaluation criteria
The bitmap visualization and analysis tools should be evaluated with the following criteria in mind:

- Cost – initial, recurring, support and maintenance, i.e. the total cost of ownership
- User friendly - Ease of product setup and application use
- Features – bitmap visualization, overlay to layout for in-line captured defects, failure signature analysis, defect modeling, layout link to schematics and netlist names
- Powerful data mining – automatic failure mode classification (AFC), statistical data analysis, root cause theorization
- Comprehensive report generation – wafer maps with thumbnail bitmap images, trend charts, pareto charts
- Integration with business practices – seamless interaction with manufacturing operations, file formats and data structures
- Data sharing - import data from, or export data to other yield analysis tools
- Flexibility – open code, code hooks, software customization

The choice of developing internal software tools over purchasing readily available products from software vendors is a difficult one. The pros and cons may include open source code vs. maintenance for in-house tools, and rich data analysis features vs. integration issues for vendor tools.

Failure mode classification
Failure mode classification is best done from a design and layout point of view. Although it is desirable to have a perspective that is closer to the root cause defect, trying to incorporate defect nomenclature into the classification should be avoided. This makes the automatic failure mode classification software much more reusable across products, memory types, and technology nodes. Making the jump to defect and process based nomenclature is solved during the root cause theorization process explained below.

A typical list of failure mode classifications is quite short. It can first be defined by three types of “elements”: rows, columns, and bits. Each of these is augmented with a span attribute of single, double, triple, quadruple, multiple, or other. A few other classifications are added: “cross” which refers to a row and a column that intersect, and several other non-bit cell periphery failure modes. These include one group of periphery logic circuits such as address buffers, address decoders, sense amplifiers, write drivers, and precharge circuit defects. Another group is contiguous blocks such as complete I/O failures or sub-arrays. A final group is systemic defects, or those global process issues which cause many of the same failure mode to occur within one memory instance. Other classifications may be added if necessary, but most of the other variations are handled in the attributes assigned to these 22 primary classifications. They are summarized with the following list:

- Single Bit
- Double Bit
- Triple Bit
- Quadruple Bit
- Multiple Bit
- Other Bit
- Single Column
- Double Column
- Triple Column
- Quadruple Column

- Multiple Column
- Other Column
- Single Row
- Double Row
- Triple Row
- Quadruple Row
- Multiple Row
- Other Row
- Cross
- Periphery Logic
- Contiguous Block
- Systemic Defect

Classification rules

Several important rules are used to classify defects. If a die fails more than 95% of the bits it is considered a gross failure, or sometimes referred to as a “chip kill”. This is tallied as part of the contiguous block classification.

If more than 2% of the total bits fail, then a fail bit density is calculated. First, a contiguous rectangular block of bits that spans all of the failing bit cells is delineated. The total number of failing bits in this block is then divided by the total bit count of the block. If this ratio, or failure density, in greater than 25%, then the die is considered an “ugly die” failure.

More than four failing bit cells within a row or column element should no longer be considered bit type defects. Also, there are rules for failing bit cell proximity to a defect classification. Normally, bit cells within two to five of the one in question are considered associated with the failure. This forms a mandatory passing buffer of bit cells around the failing bit cell pattern.

The sequence of the classification search is chosen to be gross failures first, then ugly failures, followed by contiguous blocks, periphery items, cross elements, column elements, row elements, and finally bit cells.

The objective of the failure mode distribution graph is to quickly see how a data set differs from the baseline. (See Figure 2). This unsorted snapshot has the advantage over the sorted pareto graph in that the expected distribution is easily depicted. The threshold for calling defects systemic in nature is generally three to five defects of the same classification type per memory instance. Choosing a threshold any higher than this distorts the defect classification pareto.

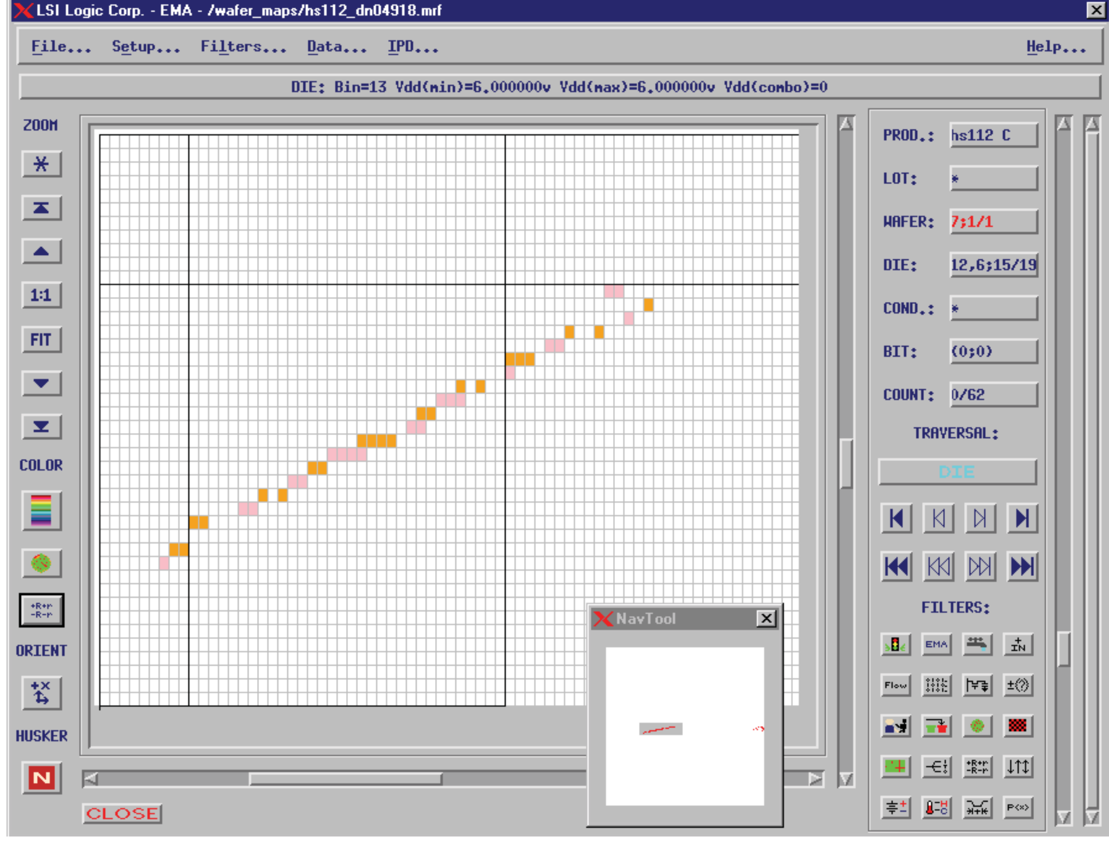


Fig 1. Example bitmap application

Variations in response

Variations in the behavior of each failure mode are classified as sensitive, dependent, or alterable. If the behavior does not change with an indepentent variable, such as voltage or temperature, it is considered a hard response. Sensitive is defined as the failing bit pattern changes in some way, but never completely passes under a specified change in test conditions. An example is a partial row that may pass more bits near the row driver under high Vdd conditions and pass fewer bits as Vdd is lowered. Dependent means the failure pattern will disappear completely under some condition. For example, a single bit may pass at a high Vdd corner. Alterable means the failing pattern permanently changes, either for the better or for the worse, because of a test condition applied. An example is a single column that is cured by a high Vdd stress test.

Root Cause Theorization

Overview

The root cause theorization process is so important because large amounts of physical failure analysis (PFA) have become impractical to complete. PFA is prohibitive because of the small feature sizes needing to be resolved, the high cost of both physical and human resources, and the short time-to-root-cause that is imposed.

Defect modeling

One method of anticipating failure modes is to do extensive defect modeling within the circuit layout. This can be done in several ways. It can be done through software tools, or it can be done with silicon. Examples include using extensive empirical data using knowledge-based systems [4], inserting defects onto the reticles used to maufacture the devices [5], and doing Focused Ion Beam (FIB) operations during the fabrication process. A pragmatic approach is often needed to augment the software circuit modeling because of the complexities encountered. Being able to model and predict silicon behavior accurately can be difficult at best, especially when sensitive signal margin issues are encountered.

Failure classification attributes

There is another level of failure classification analysis that is very important to study. Many additional attributes of the failure mode can be uncovered, and these hidden attributes are subtle affects that can be observed in the failure characteristics that add to the total signature of the failure. These affects can often lead to stronger theories for uncovering the root cause defects. A number of these attributes are outlined below.

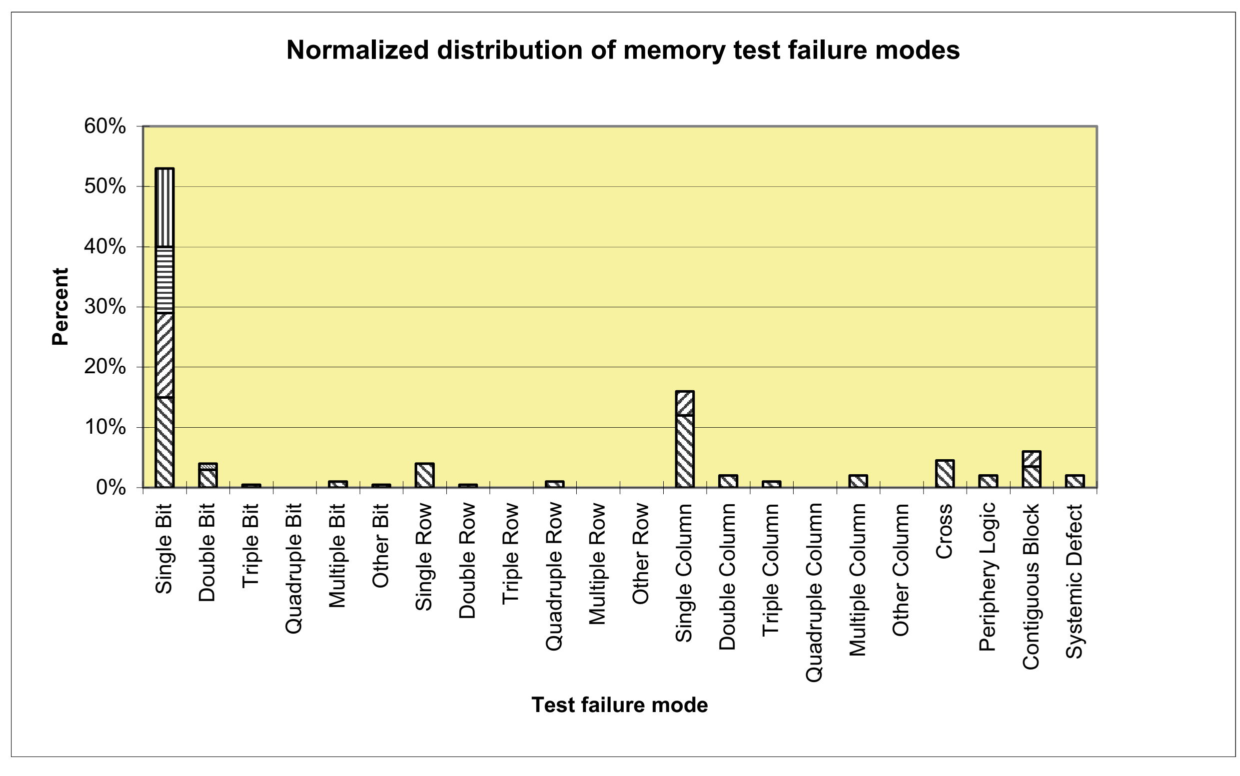


Fig. 2. Example failure mode distribution snapshot

Basic failure mode classifications for row and column elements includes a width aspect like single column or double row. An attribute that describes the length of the failing element is a simple way to add important information to the total signature. Partial elements are very common especially in memories that are partitioned. For example a half single row failure can point to an obvious problem related to a center row driver or row loading affect.

The distribution of even and odd physical row and column locations should always be tabulated. These four possible combinations can often reveal misalignment issues in the silicon processing, especially at active, polysilicon, and contact layers.
Another powerful way to tally physical locations is by discrete modulo values for the elements. For example, columns may be laid out with an eight-column repeat pattern. Tabulating the physical column modulo eight results of zero through seven may reveal a predisposition to bit cell failures at repetitive layout positions. Boundary elements like noise barriers, timing circuits, redundant elements, and spacer cells may have an affect on a failure mode because of very subtle process variations.

Layout variations including rotation, mirroring, and stepping of bit cells are very common, and are actually desirable for signature analysis purposes. If the failure rate is tallied by these possible orientations, it can reveal misalignment issues introduced in the wafer processing.

Classification rules should be used to determine the boundaries of a failure. Careful inspection of the bit cells that make up the failure should be studied. Attributes of fail logic states, fault types, and access port dependencies should be analyzed on a bit by bit basis to look for any irregularities.

There are, of course, many attributes that can be revealed in test pattern variations. The test pattern attributes of algorithms, background data patterns, and address sequences should be orchestrated through the test flow to provide important differentiations in the failure mode. This should include algorithms to identify which aggressor operations activated which victim bit cells, and write vs. read cycle analysis. It should also include test variables such as voltage, temperature, and timing. These are powerful ways to look for signal margin and leakage issues.

Correlation studies
Various correlation studies should be done in order to look for further signature information and improve the root cause prediction process. In-line process defect data can identify killer and nuisance defects when overlaid to the bitmap. Also, silicon process equipment tool commonality studies can be revealing. Careful analysis of test data such as prime die vs. repaired die are important as well. Yield loss patterns on the wafer may reveal edge affects or other defect clustering issues. Trend charts of classification failure rates by process technology and memory type can be used to set up Statistical Process Control (SPC) charts for quick reaction to process excursions. And finally, operational life failures may have a signature that is visible at wafer test. Careful study of the failure signature may reveal a link to characteristics and yield loss mechanisms that are easier to pursue FA on.

A first pass assumption for a maverick wafer is that there is only one root cause that makes this material deviate from baseline material. However, this usually manifests itself as more than one failure mode classification that is atypical. The importance of studying the distribution of failure modes can be understood in this fact, so one must look for multiple failure modes that can have the same root cause. If this model can not be justified, then one must speculate that multiple root cause defects must exist for the material.

Physical Failure Analysis Strategies

Overview
A systematic physical FA strategy must be closely managed. There are many choices of FA techniques available so optimizing the path to root cause is critical in controlling costs and cycle times. And because there are often red herrings in the data, team brainstorming is an important element throughout the analysis process.

Physical FA Flow
A template flow is presented that can be tailored depending on the failure analysis tool set available. Separate flows can be set up for front side and backside sample approaches.

1. In-line process defect monitoring tools [6] can visually overlay defect images with the layout. This should be the first line of defense in the yield enhancement and FA efforts.

2. Physical package analysis may include X-ray or Acoustic Micro Imaging (AMI) [7].

3. Decapsulation or backside sample preparation to expose the chip can be done next. If possible, this can be followed by a course optical microscope inspection.

4. Liquid crystal and emission microscopy [8] are used to look for thermal hot spots or photon emission under dynamic test conditions.

5. Operating the device under AC bias conditions and probing using mechanical probes, E-beam, Laser Scanning Microscope (LSM) techniques such as Light Induced Voltage Alteration (LIVA) or Thermal Induced Voltage Alteration (TIVA) [9] can be very effective.

6. After this point the sample will no longer be fully testable so a gate should be placed here for reviewing the FA progress and conducting more team brainstorming sessions.

7. Deprocessing using mechanical polishing, wet chemical etch, or Reactive Ion Etch (RIE) techniques is done next.

8. Voltage Contrast (VC) with it's various techniques such as passive, static, etc. can be used [10].

9. DC characterization, for example of an individual bit cell [11], is useful. This often requires making routing cuts and depositing probe pads using a FIB.

10. Various imaging techniques such as Scanning Electron Microscopy (SEM), Transmission Electron Microscopy (TEM) can be employed [12]. Cross-section or plan views are chosen.

11. Scanning Probe Microscopy (SPM) such as Atomic Force Microscopy (AFM), and Scanning Tunneling Microscopy (STM) reveal physical or chemical characteristics of the sample.

12. Elemental analysis using Energy Dispersive X-Ray Spectroscopy (EDS), or Auger Electron Spectroscopy (AES).

Conclusion

Integrating and streamlining these five disciplines can be a monumental challenge, but the rewards are usually beyond one's imagination. As feature sizes continue to shrink, the physical failure analysis process can become tedious, time consuming, and financially burdening. The key to making significant productivity improvements lies in the data mining and signature analysis process. Making intelligent and efficient use of this data can drive failure analysis cycle times down and hit rates up dramatically.

References

1. B. Prince, "Semiconductor Memories, A Handbook of Design, Manufacture, and Application", John Wiley & Sons. Ltd. Chichester, U.K., pg. 15 (1983, 1991)

2. A.J. van de Goor, "Testing Semiconductor Memories, Theory and Practice", John Wiley & Sons. Ltd. Chichester, U.K., pgs. 25-60 (1991)

3. M. Faucher, "Pattern Recognition of Bit Fail Maps". International Test Conference Proceedings, pgs. 460-463 (1983)

4. T. Viacroze, G. Fourquet, M. Lequex, "An Expert System for Help to Fault Diagnosis on VLSI Memories". 14th International Symposium on Testing and Failure Analysis, pgs 153-159 (1988)

5. M. Santana, Jr., A.V. Herrera, "Methodology to Correlate Defect Reduction Systems to Electrical Test data via Artificially Manufactured Defects", 28th International Symposium for Testing and Failure Analysis, pgs 587-589 (2002)

6. M. Rehani, B. Whitefield, J. Knoch, "An Automated Recipe-Based Defect Analysis System for ASICs", http://www.kla-tencor.com/company/magazine/spring00/anautomated.pdf, (2000)

7. T. Adams, "Removing Hidden Internal Packaging Defects From Production", http://www.us-tech.com/dec99/spftr/spftr001.htm, (1999)

8. J.S. Seo, S.S. Lee, C.S. Choe, S. Daniel, K.D. Hong, C.K. Yoon, "Photoemission Spectrum Analysis – A Powerful Tool for Increased Root Cause Success", 21st International Symposium for Testing and Failure Analysis, pgs 73-98 (1995)

9. R.A. Faulk, "Advanced LIVA/TIVA Techniques", 27th International Symposium for Testing and Failure Analysis, (2001)

10. K.J. Bertsche, H.K. Charles, Jr., "The Practical Implementation of Voltage Contrast as a Diagnostic Tool", International Reliability Physics Symposium, pgs 167-178 (1982)

11. V.K. Wong, C.H. Lock, K.H. Siek, P.J. Tan, "Electrical Analysis to Fault Isolate Defects in 6T Memory Cells", Proceedings of 9th IPFA 2002, Singapore, pgs 101-104 (2002)

12. R.B. Marcus, T.T. Sheng, "Electron Microspy and Failure Analysis", International Reliability Physics Symposium, pgs 269-275 (1981)

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Beam-Based Defect Localization Techniques

Edward I. Cole Jr.
Sandia National Laboratories, Albuquerque, New Mexico, USA

Abstract

SEM and SOM techniques for IC analysis that take advantage of "active injection" are reviewed. Active injection refers to techniques that alter the electrical characteristics of the device analyzed. All of these techniques can be performed on a standard SEM or with a SOM using the proper laser wavelengths.

Introduction

The scanning electron microscope (SEM) is as standard a tool for IC failure analysis as the optical microscope. The SEM's advantages over light microscopy include greatly increased depth of field, much higher magnification, increased working distance, and improved imaging of surface topography. In addition, the interaction of the electron beam with the IC enables unique imaging and analytical capabilities. The strengths, limitations, and effects upon the device being analyzed must be understood to use the SEM effectively. SEM settings that are appropriate for one analytical technique may be unsuccessful, may give misleading results, or may unexpectedly damage the device being analyzed if used for a different technique.

Since the early 1990's, the scanning optical microscope (SOM) has been a valuable tool for photon beam based analysis of ICs. Like SEM, the SOM can produce reflected light images of improved spatial resolution and depth of field compared to conventional optical microscopy. Additionally, by taking advantage of silicon's relative transparency to infrared wavelength's, the SOM has become one of the main vehicles for backside failure analysis tool development.

This presentation reviews SEM techniques for utilizing charge injection for IC analysis, the analogous SOM analysis methods, and relatively new SOM analysis tool developments. All of these techniques can be performed on a standard SEM or with a SOM using the proper laser wavelengths. The review is designed to provide beneficial information to both novice and experienced failure analysts. Topics to be covered are (1) SEM techniques: Electron Beam Induced Current, Resistive Contrast Imaging, and Charge-Induced Voltage Alteration (both high and low energy versions), recent improvements in low energy SEM resolution and (2) SOM techniques: Optical Beam Induced Current, Light-Induced Voltage Alteration, Optical Beam Induced Resistance Change/Thermally-Induced Voltage Alteration, Seebeck Effect Imaging, Soft Defect Localization/Laser-Assisted Device Alteration, and Solid-Immersion Lens applications. Each technique will be described in terms of the information yielded, the physics behind technique use, any special equipment and/or instrumentation required to implement the technique, the expertise required to implement the technique, possible damage to the IC as a result of using the technique, and examples of using the technique for failure analysis.

Electron Beam Techniques

Electron Beam Induced Current (EBIC) Imaging

Information Technique Yields: EBIC [1, 2] imaging localizes regions of Fermi level transition. EBIC is primarily used to identify buried diffusions and Si defects.

Physics Behind Technique Use: When the primary electron beam is scanned across a sample many interaction products are generated (Figure 1). Collisions between the primary electrons and the target material form electron-hole pairs within the bulk of the sample. The relatively low ionization energies (less than 10 eV) of materials used in integrated circuit manufacturing allow a single 10 keV primary electron to produce an many as 500 to 1000 free electron-hole pairs. These pairs usually recombine randomly in the material; however, if production occurs in a space-charge (depletion) region, the charge carriers will be separated by the junction potential before recombination. The large number of pairs per primary electron generates an EBIC signal much larger than the incident beam current. The magnitude and direction of the induced current is used to generate an image localizing where junction potentials occur. In contrast to secondary electron and backscattered electron scanning electron modes, the EBIC signal detector is the device itself. By controlling the primary electron beam energy, the depth of the diffusions examined can be differentiated. The range of the primary electron injection is given in Table 1 and is calculated from the Kanaya-Okayama formula . [3] (Calculations courtesy of W. Vanderlinde, Laboratory for Physics Sciencies). Various device pins may be sampled to observe different EBIC signals. While normally performed as a two contact technique, single contact EBIC or SCEBIC which uses momentary displacement currents for imaging can be a useful FA approach.

Difficulty to Implement: The equipment necessary to perform EBIC imaging is a scanning electron microscope, current to voltage amplifier, and an electrical vacuum feed-through. The only sample preparation needed is lid/plastic package removal. The passivation may or may not be removed for EBIC analysis. No electrical driving equipment is necessary since the integrated circuit is normally driven only by the electron beam. Biased EBIC, while uncommon, has been reported. Because of the small signal generated, a digital image acquisition system would be advantageous to image acquisition/manipulation, but not necessary. Electrical testing to determine proper node selection is also desirable. An example of an EBIC image is shown in Figure 2. Diffusions across the entire IC are visible in Figure 2.

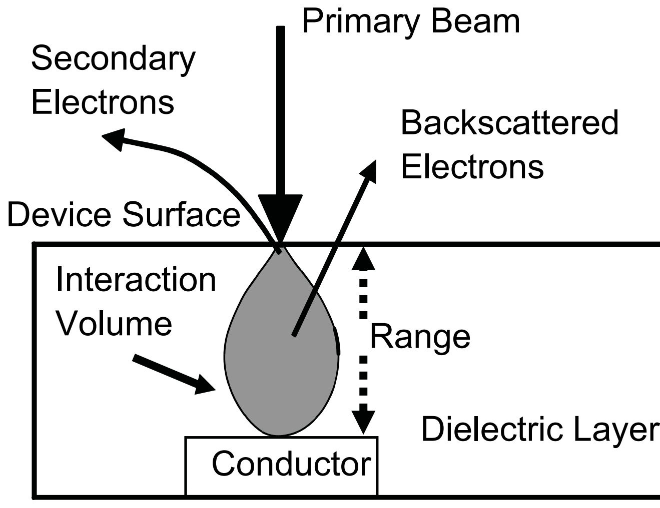


Figure 1. Primary electron beam interaction products.

Possible IC Damage: No direct "physical" damage occurs with electron beam testing. Alteration of the threshold voltage on MOS transistors is possible. This change in threshold voltage results from irradiation damage to the gate oxides of MOS transistors. The damage is generated by direct primary electron collisions and by x-rays generated through interactions in the surface layer(s). The primary electrons quickly alter the interface trap density and occupancy as well as the fixed charge levels in the gate oxide. If desired, low temperature annealing (100-150 ^{o}C) may be used to restore the threshold voltage to near its original value.

If surface layer removal is performed, there is the possibility of device damage during the removal process. Conductive coating to eliminate charging will prevent/hinder microelectronic operation and should be removed to restore device operation. Note that removal of any coatings may damage the IC under examination.

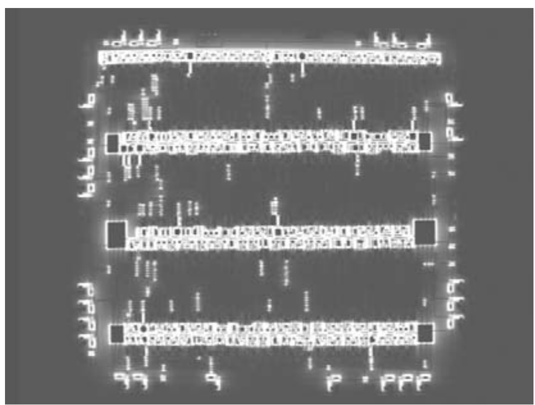


Figure 2. Example of EBIC imaging of an entire die.

Table 1. Beam range in Al and SiO_2. [3]

Beam energy (KeV)	Range in Al (μm)	Range* in SiO_2 (μm)
1	0.028	0.036
3.5	0.22	0.29
5	0.41	0.52
10	1.32	1.66
20	4.19	5.23
30	8.24	10.40

*for a ρ of 2.0 g cm^3

Resistive Contrast Imaging (RCI)

Information Technique Yields: RCI [4, 5, 6] generates a relative resistance map between two test nodes of a passivated integrated circuit. The map generated will display buried conductors on an integrated circuit and may be used to localize open conductors.

Physics Behind Technique Use: RCI obtains resistance information by using the integrated circuit as a complex current divider. Figure 1 displays the electron beam interaction products between a passivated integrated circuit and a 10-keV primary electron beam. To obtain RCI information the primary electron beam energy is increased until the tip of the interaction volume intersects the buried conductor of interest. A portion of the primary electron beam current will be injected into the conductor. Using an amplification configuration as shown in Figure 3, the currents induced by electron beam exposure will have a path out of the integrated circuit. The relative resistance between the electron beam position on the circuit and the test nodes determines the direction and amplitude of current flow. The current, on the order of nanoamps, is amplified and used to make a resistance map of the conductors. Usually the power and ground inputs are used as test nodes because of their global nature across the integrated circuit. However, other node combinations may be used if desirable/indicated. If a resistance change occurs along a conductor relative to the test node combination selected, such as an open conductor, the RCI image will display an abrupt contrast change at the open site. The RCI image in Figure 4a localizes an electromigration open along a clock line. Figure 4b is a Backscattered Electron (BSE) image of a similar field of view at 30 keV. While not an active SEM imaging technique, the elastically scattered primary electrons in BSE imaging show the evidence of electromigration cracks in the metallization.

Difficulty to Implement: The equipment items necessary to implement RCI are: a scanning electron microscope, current to voltage amplifier, and an electrical vacuum feed-through. The only sample preparation needed is lid/plastic package removal. The passivation is not removed for RCI. By increasing the primary electron beam energy, multilevel conductors under metal may be observed. No electrical driving equipment is necessary since the integrated circuit is driven only by the electron beam. Because of the small signal generated, a digital image acquisition system would be advantageous to image acquisition/manipulation, but not necessary. Electrical testing to determine proper node

selection is also desirable. Another induced current effect, Electron Beam Induced Current (EBIC) should be avoided. EBIC generates currents several orders of magnitude greater than RCI which can mask the RCI signal. EBIC signals are generated primarily from buried diffusions and can be avoided by using lower primary electron beam energies. With practice in selecting the proper primary electron beam energy and current RCI data may be acquired readily. Unfortunately, not all internal conductors with defects will be identified using RCI. Escape from detection occurs when the current paths from the defect-containing conductor to the IC pins are too convoluted and no difference in resistance occurs across the open site relative to the IC pins.

Possible IC Damage: The damage concerns described in the EBIC section are independent of applied technique and are applicable here if high beam energies are used. At the primary electron beam energies used for RCI, no direct primary electron/gate oxide interactions occur. However, some bremsstrahlung x-rays generated by interactions in the passivation layer will deposit energy in the gate oxide. This x-ray dose will alter the interface trap density and occupancy as well as the fixed charge levels in the gate oxide. Experiments on 3 micron, commercial grade, MOS transistors indicate that 40 images with 1 micron spatial resolution may be acquired before the threshold voltage shifts by 5%. As with EBIC, low temperature annealing (100-150 °C) may be used to restore the threshold voltage near or to its original value.

Charge-Induced Voltage Alteration (CIVA) Imaging

Information Technique Yields: CIVA [7] was developed to localize open conductors on both passivated and depassivated CMOS ICs. CIVA facilitates localization of all open interconnections on an entire IC in a single, unprocessed image. CIVA has been applied to an analog bipolar technology with similar results.

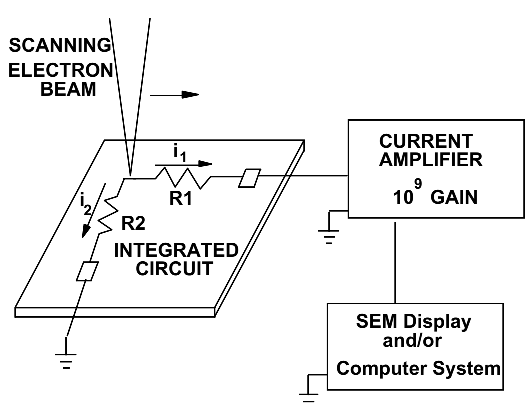


Figure 3. RCI imaging system.

a)

b)

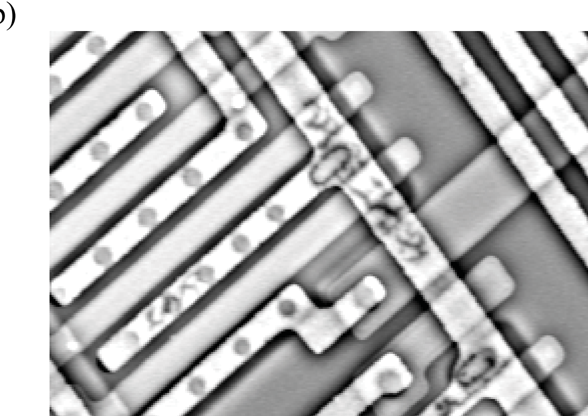


Figure 4. a) Example of RCI imaging locating an open conductor. b) BSE image of the same area showing metal voiding.

Physics Behind Technique Use: CMOS ICs with open conductor lines may function at low to moderate (< 50 kHz) frequencies. The reason for this functionality is that significant quantum

mechanical electron tunneling across the open can transport enough charge at low frequencies to maintain functionality. The maximum operating speed depends upon the nature of the open.

Even though the ICs may be functional with open conductors, charge injection into the floating portion of the conductor may cause significant loading that can overwhelm the open's tunneling capacity. CIVA takes advantage of this tunneling capacity to create an image of "loaded" areas. The CIVA image is generated by monitoring the voltage shifts in a constant current power supply as the electron beam is scanned over a biased integrated circuit. As electrons are injected into non-failing conductors the additional current, on the order of nanoamps, is readily absorbed and produces little change in the supply voltage. When charge is injected into an electrically floating conductor, the voltage of the negative conductor becomes more negative. This abrupt change in voltage on the floating conductor generates a temporary shift in the voltage demand of the constant current source supplying bias to the integrated circuit. The shift in power supply voltage can be either positive or negative depending on the proper state of the floating conductor. As the electron beam moves away from the floating conductor the previous equilibrium is quickly (~ 100 msec depending upon the bandwidth of the current source) reestablished. The shifts observed in the power supply voltage, even for opens that exhibit significant tunneling, are on the order of 100 mV with a 5 V supply voltage. These relatively large shifts produce images in which the contrast is dominated by the open conductors. Transistors with "weak" drive capacity have also been identified using CIVA.

Figure 5 displays the experimental setup used to generate CIVA images. Figure 6 shows an example of CIVA imaging on a passivated IC with an open conductor. The top left image in Figure 6 shows a CIVA image with no processing. The other three images show an overlay of the CIVA signal and secondary electron images at different magnifications. The highest magnification shows the open conductor at a polysilicon step.

Difficulty to Implement: The equipment necessary to perform CIVA imaging are a scanning electron microscope, a constant current source, and an electrical vacuum feed-through. The only sample preparation needed is lid/plastic package removal. A digital image acquisition would be advantageous to image acquisition/manipulation, but is not necessary. The passivation may or may not be removed for CIVA analysis. The IC must be biased into a non-contention state, but no complicated vector set is required. By increasing the primary electron beam energy, multilevel conductors under metal may be detected. Like RCI, the selection of proper primary electron beam energy includes using energies that just reach the buried conductors and avoid EBIC signal generation

Possible IC Damage: Same as with RCI imaging.

Charge-Induced Voltage Alteration (CIVA) Using Low Primary Beam Energies

Information Tool Yields: CIVA at low primary electron beam energies (< 1.0 keV) localizes open metal conductors beneath passivation layers as with conventional CIVA [8]. The major difference between the two approaches is that low beam energy CIVA (LECIVA) does not introduce any irradiation damage to the IC under test. Additionally, low energy CIVA can be implemented on electron beam test systems which operate only at low primary electron beam energies.

Physics Behind Tool Use: (NOTE: This description builds on the physics of CIVA and assumes the reader is familiar with conventional CIVA.)

Conventional CIVA localizes open conductors on an IC by powering the IC with a constant current source and monitoring the voltage fluctuations in the power supply with electron beam position. Passivated and multilevel interconnect ICs are

examined by increasing the primary electron beam energy until electrons from the beam are injected into the buried conductors. This capability to probe through buried layers is very powerful, but care must be exercised to avoid irradiation damage of the IC under examination. The requirement that the interaction volume of the primary electron beam interact directly with the buried conductor also limits the utilization of CIVA to SEMs with variable beam energy. Most electron beam testers have a fixed primary electron beam energy around 1.0 keV. CIVA at low primary beam energies (< 1.0 keV) circumvents these problems by producing CIVA images of buried IC conductors without direct interaction.

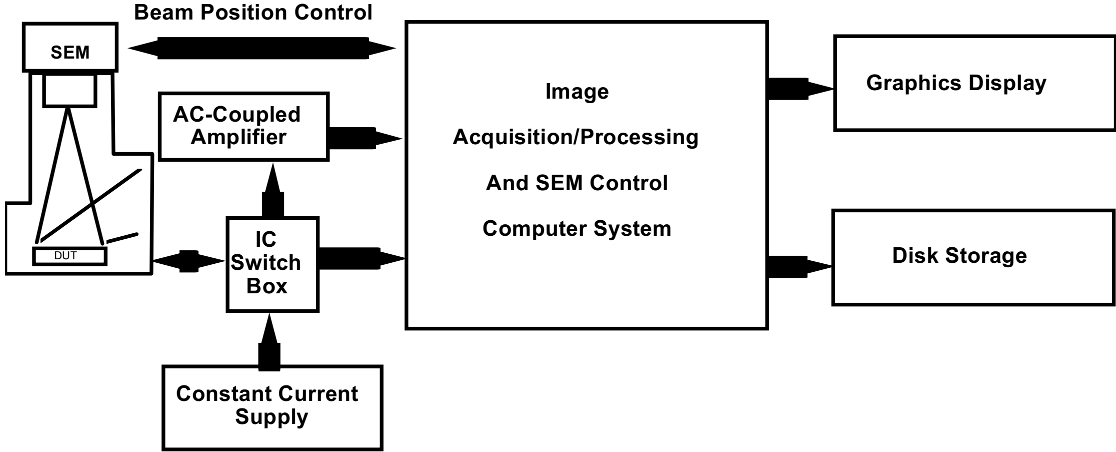


Figure 5. CIVA imaging system.

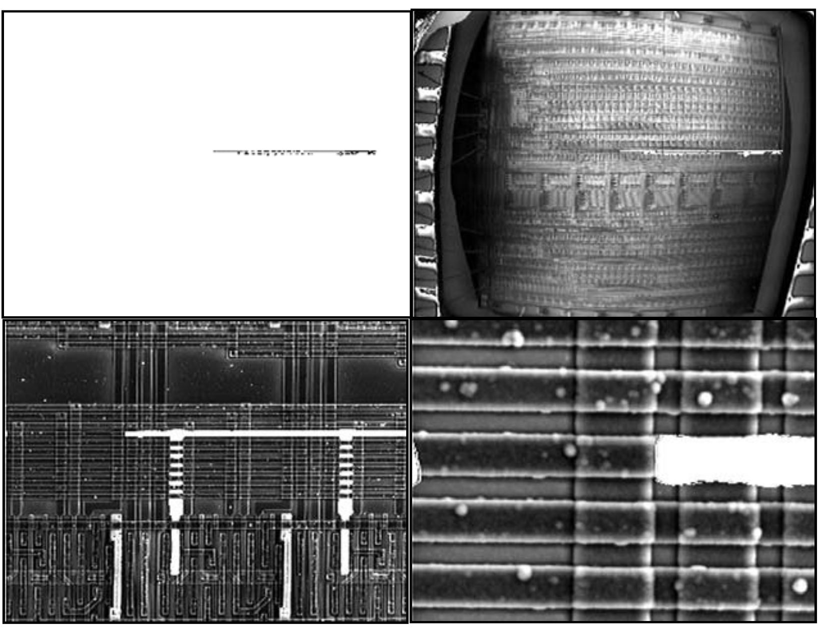


Figure 6. Image examples of CIVA at low magnification (upper left), overlaid with an SE image(upper right), and higher magnification CIVA/SE images localizing an open conductor (lower images).

At low primary electron beam energies (< 1.0 keV) the surface passivation of an IC will reach a positive equilibrium voltage as with Capacitive Coupling Voltage Contrast (CCVC). While the time to reach the positive equilibrium voltage is directly proportional to the incident electron beam flux, the value of the equilibrium voltage has previously been thought to be independent of beam flux. Recent experimental work has shown that at very high primary electron beam currents (> 20 nA) the equilibrium surface voltage will become negative at low primary electron beam energies. During a low energy beam exposure, a surface initially at 0V is thought to go positive and then negative at high electron flux. The negative voltage is believed to be due to a depletion of the available secondary electrons at the surface during a high electron flux, low energy exposure. (This effect has **not** been previously documented in literature describing electron beam surface interactions. In fact, it is generally believed that the surface equilibrium voltage is independent of primary electron beam current.) This rapidly changing surface equilibrium voltage will polarize the passivation and produce a changing bound charge at the metal conductor-passivation interface, similar to the bound charge used to make a CCVC image. The changing bound charge will introduce a small voltage pulse on the buried conductor. As with conventional CIVA, this small voltage change is readily absorbed by non-failing conductors, but the voltage pulse can change the voltage of floating conductors and alter the power supply demands of the IC under test. By imaging the changing voltage demands of a constant current supply powering the IC under test, low power CIVA images are produced.

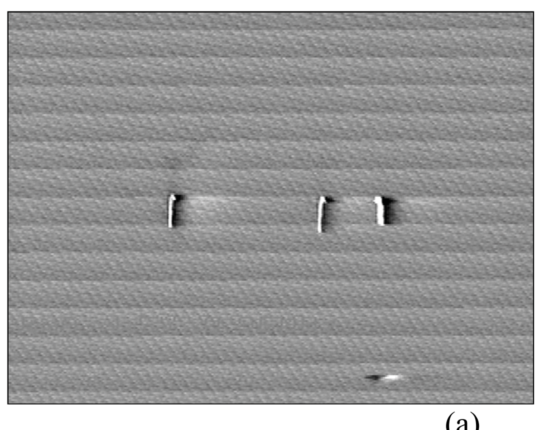
AC operation of the IC produces an improved low energy CIVA image by "depleting" the bound charge on the IC conductors. AC operation also increases the probability of localizing open conductors by increasing the chances that a given failing conductor will be placed in a logic state susceptible to CIVA observation. AC operation is facilitated by performing CIVA on commercial electron beam test systems which are normally configured for AC analysis of an IC.

Low energy CIVA images of open conductors are show in Figure 7. The images were acquired using a 300 eV primary electron beam energy.

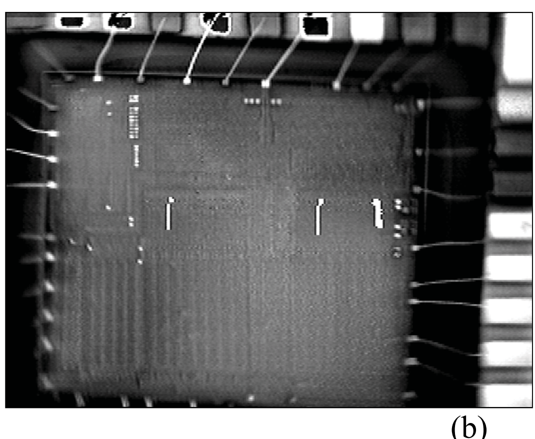
Implementation: The equipment necessary to perform low energy CIVA is identical to that for conventional CIVA: a scanning electron microscope, a constant current source, and an electrical vacuum feed-through. In addition, electron beam testers can be used. The sample preparation needed is lid/plastic package removal. The IC must be biased in a non-contention state(s), but no complicated vector set is required. High electron beam currents are required for low energy CIVA analysis, so procedures to maximize the currents such as larger apertures and lower condenser lens settings are advisable.

For standard IC technologies with 3 levels of metal interconnect low energy CIVA will be capable of imaging most structures. Higher interconnect layers between the passivation and the conductors of interest will absorb the bound charge of interest, preventing observation of the deeper conductor regions. If no metal layer obscures the bound charge path from the surface to a lower level conductor, the conductor of interest will be susceptible to the LECIVA effect. The increased thickness of dielectric material will reduce the low energy CIVA effect just as the CCVC signal is attenuated by thicker dielectric layers. Under these conditions conventional CIVA may be advisable, but low energy CIVA should be applied first because of its almost totally non-destructive nature.

Possible IC Damage: No physical or irradiation damage occurs with low energy CIVA examination. A possible "damage" scenario for ICs is the creation of snap-back or latchup conditions in a biased IC, which can occur rarely in limited situations. This can be avoided by prudent selection of the current compliance of the biasing power supply.



(a)



(b)

Figure 7. A low energy CIVA image (a) showing 3 open conductors. The CIVA image was acquired using a 300 eV primary electron beam. A combined secondary electron image (b) is shown for registration.

Recent Low Energy SEM Resolution Improvements

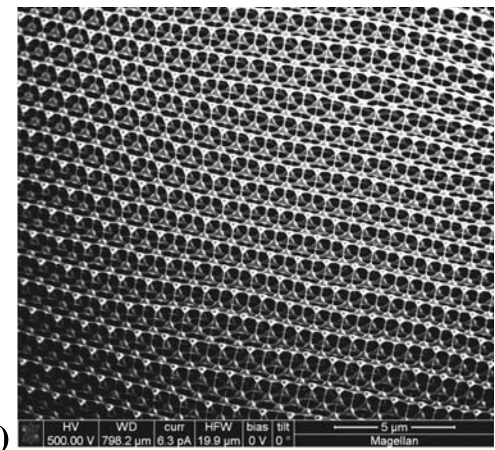
Resolution improvements are not an active SEM technique, but improvements yielding a smaller energy distribution in the selected primary electron beam energy and biased sample approaches have increase the spatial resolution of the SEM for low voltage applications. [9] These improvements are applicable to enhancing multiple SEM techniques.

Monochrometers in the upper SEM column yield a smaller energy spread and hence the ability to focus to a smaller spot at low beam energies. Samples that were previously “too fragile” for high resolution SEM imaging can now be observed at low energies with spatial resolutions of 1 nm or better reported with beam energies of 500 eV or greater. Figure 8 displays uncoated pyrolized photoresist with Au particles using monchromated 500 and 300 eV primary electron beams. Without monochromatization this resolution could only be achieved at higher electron beam energies that would damage the sample.

If the sample can be biased positively, the interaction or “landing energy” can be reduced further while still maintain improved spatial resolution.

Note that while the planar resolution is improved; the depth of focus is reduced compared to conventional non-monochromated SEM imaging.

a)



b)

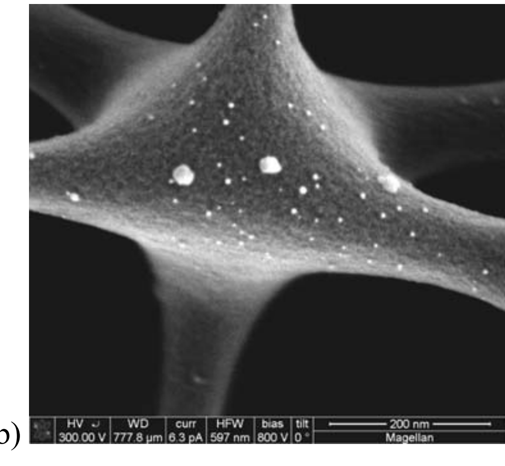


Figure 8. Uncoated pyrolized photoresist with Au particles using a) 500 eV and b) 300 eV primary electron beams on a monchromated SEM.

Optical Beam Based Techniques

Optical Beam Induced Current (OBIC)

Information Tool Yields: The optical beam induced current (OBIC) [10] mode of the scanning optical microscope (SOM) maps regions of Fermi level transition in a manner similar to the electron beam induced current (EBIC) imaging described in the electron beam analysis section.

Physics Behind Tool Use: The basic SOM consists of a focused light spot, normally from a laser source, that is scanned over the sample as some property of the sample is observed. This property is used to produce an image of the sample as the spot is scanned. The scan may be achieved by either scanning the sample or the light spot.

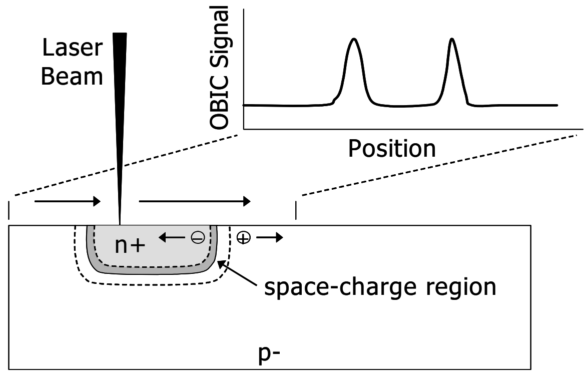


Figure 9. Physics of OBIC signal generation.

OBIC mode SOM is very similar to EBIC mode scanning electron microscopy. If the photon energy of the incident light is greater than the semiconductor bandgap electron-hole pairs are formed as light penetrates the semiconductor (Figure 9). This energy is 1.1 eV or 1.1 µm for silicon. The pairs will normally recombine randomly. However, if pair production occurs near a space charge (depletion) region, the carriers are separated by the junction potential before recombination and a net current is produced. The current generated on the integrated circuit or semiconductor sample may be monitored and imaged with the optical beam scan. Common failure analysis applications are localizing buried diffusion regions, defective/damaged junctions, gate oxide shorts (with deprocessing), and other regions with local Fermi level variations on the sample. Figure 10 shows the locations of diffusions on a CMOS device by observing the OBIC signal across V_{DD} and V_{SS}. While normally performed as a two contact technique, single contact OBIC or SCOBIC which uses momentary displacement currents for imaging, can be a useful FA approach. The use of an IR laser extends the usefulness of OBIC by permitting backside observation of devices through the substrate, taking advantage of silicon's transparency to IR light.

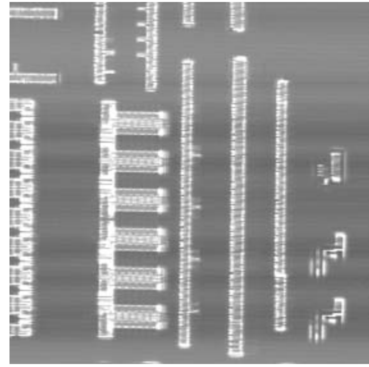


Figure 10. OBIC image of an unbiased CMOS device made by observing the current across V_{DD} and V_{SS}.

Possible IC Damage: No physical damage occurs with SOM examination. If biased OBIC is used it is possible to initiate a snapback or latch-up condition in the biased integrated circuit. Damage can be avoided or mitigated by prudent selection of the light source and the current compliance of the biasing power supply.

Light-Induced Voltage Alteration (LIVA)

Information Tool Yields: LIVA [11] is a Scanning Optical Microscopy (SOM) technique developed to localize diffusions associated with integrated circuit defects. The high selectivity of LIVA permits examination of the entire IC die in a single, unprocessed image. LIVA examination of an IC's surface from the front side is performed using a visible laser light source. By using an infrared light source LIVA has shown applicability

to backside examination of ICs. The LIVA approach has also been used to identify the logic states of transistors with much greater sensitivity than previous optical techniques.

Physics Behind Tool Use: The effects of photon injection on ICs have been well documented in the literature. As photons interact with semiconductor materials they produce electron-hole pairs which can effect IC operation. The most widely know SOM technique utilizing electron-hole pair generation is Optical Beam Induced Current (OBIC). In OBIC non-random recombination of electron-hole pairs near diffusions generates a net current that can be used to image the diffusions. OBIC can also be used to identify biased transistor logic states by observing the magnitude of the "photo-current" generated by optical beam exposure.

LIVA, like OBIC, takes advantage of photon generated electron-hole pairs to yield IC functional and defect information. LIVA is an optical beam corollary to the electron beam technique, charge-induced voltage alteration (CIVA). In both techniques the device under examination is biased using a constant current power supply. LIVA images are produced by monitoring the voltage shifts in the power supply as the optical beam from the SOM is scanned across the device. Voltage shifts occur when the recombination current increases or decreases the power demands of the IC.

The LIVA measurement and imaging of voltage shifts rather than directly observing the photo-currents has two advantages. First, the IC will act as its own amplifier producing a much larger LIVA voltage signal than a photo-current signal. This is in part due to the difference in "scale" for voltage and current on ICs. For example, a CMOS IC biased at 5V had a photo-current increase of 100 nA (from 90 nA to 190 nA) when a transistor gate was exposed to the SOM beam. When the same conditions were repeated for the IC biased with a constant current supply set to yield 5V with no illumination (90 nA supplied), the voltage decreased by **2.4 volts** (from 5 V to 2.6 V). A second advantage is that IC voltages are much simpler to measure than IC currents. Current measurement is in series, with the measurement system becoming "part" of the IC. Because of this series measurement, most current amplification systems will have maximum current limits (typically 250 mA) that limit the operational range without modifications. There is also the added complication of sometimes needing to measure a relatively small photo-current against a large dc background current. Voltage measurements are made in parallel and are therefore "separate" from the IC with none of the current measurement limitations. Small changes in voltage are easily measured using an ac coupled amplifier immune to background dc voltages. This relative signal difference and simpler equipment setup make LIVA far more attractive than conventional photo-current methods.

Defects on ICs, such as diffusions connected to open conductors and electrostatic discharge damage, produce relatively large LIVA signals when compared to the LIVA signal of biased diffusions under similar conditions (10 to 1000 or more times greater). This relative difference is used to produce highly selective LIVA images for defect localization.

When no defects are present, or when they are "deactivated" by changing the state of the IC such that the diffusions associated with defects have little or no electric field, the LIVA signal from non-defective diffusions can be used to determine transistor logic state.

Backside LIVA examination is performed using an infrared light source. The source wavelength must be large enough to take advantage of silicon's greater transparency to infra-red light, but small enough to generate electron-hole pairs in the diffusion regions of the IC. An 1152 nm, 5 mW, HeNe laser has been used to identify LIVA defects from the backside of an IC die. This laser does not produce enough electron-hole pairs for backside logic mapping using LIVA. A 1064, 1.2 W

Nd:YAG laser has been used to successfully identify CMOS transistor logic states from the backside of an IC die.

Figure 11 displays how a defect can be localized using LIVA from the backside of an IC. The IC is a radiation hardened version of the Intel 80C51 microcontroller. A 5 mW, 1152 nm laser was used to acquire the images. The LIVA image in Figure 11a displays a small signal in the lower right. Figure 11b is a reflected light image showing the same field of view. Note that the entire IC is being examined in this field of view. Figures 11c and 11d are higher magnification LIVA and reflected light images of the area identified in Figures 11a and 11b. Later analysis indicated that the LIVA signal is caused by an open metal-1 to silicon contact. Note that the open contact area is completely covered by a metal-2 power bus that would obscure any front side examination.

Figure 12 displays how logic state mapping can be performed using LIVA from the backside. A 1.2W, 1064 nm laser was used to acquire Figure 12. The field of view shows a portion of the SRAM embedded in the microcontroller examined in Figure 12. Figures 12a and 12b are same field of view LIVA and reflected light images respectively. The logically “off” p-channel transistors produce the dark contrast seen in Figure 12a. Logically “on” p-channel transistors produce a bright contrast. This difference in contrast permits reading of the SRAM’s contents.

Difficulty to Implement: The equipment necessary to perform LIVA imaging are a scanning optical microscope (SOM), a constant current source, voltage amplifier (preferably AC coupled), and electrical connections to the IC under examination. Almost all commercial SOMs will have an auxiliary or OBIC input port suitable for the LIVA input. The IC must be biased in a non-contention state, but no complicated vector set is required. For front side LIVA examination any visible wavelength will generate enough electron-hole pairs for LIVA imaging. Backside LIVA examination must consider the wavelength restrictions mentioned above.

Possible IC Damage: No physical damage occurs with LIVA examination. Two possible "damage" scenarios for ICs are: (1) erasure of nonvolatile memories which are susceptible to certain light energies and (2) creating snap-back or latchup conditions in a biased IC. This can be avoided by prudent selection of the light source and the current compliance of the biasing power supply. Die thinning is not necessary for LIVA, but polishing the backside of the IC die to reduce light scattering greatly improves LIVA and reflected light image resolution. All of the examples shown above have had polished backsides.

Optical Beam Induced Resistance Change (OBIRCH), Thermally-Induced Voltage Alteration (TIVA) and Seebeck Effect Imaging (SEI)

Information Tool Yields: OBIRCH [12], TIVA and SEI [13] are different variants of a Scanning Optical Microscopy (SOM) technique developed to localize opens (SEI) and short circuits (OBIRCH/TIVA) on integrated circuits with localized heating. The high selectivity of both variants permit examination of the entire IC die in a single, unprocessed image. At the wavelengths used localized heating can be produced from the front or backside of the IC making defect localization possible from either side.

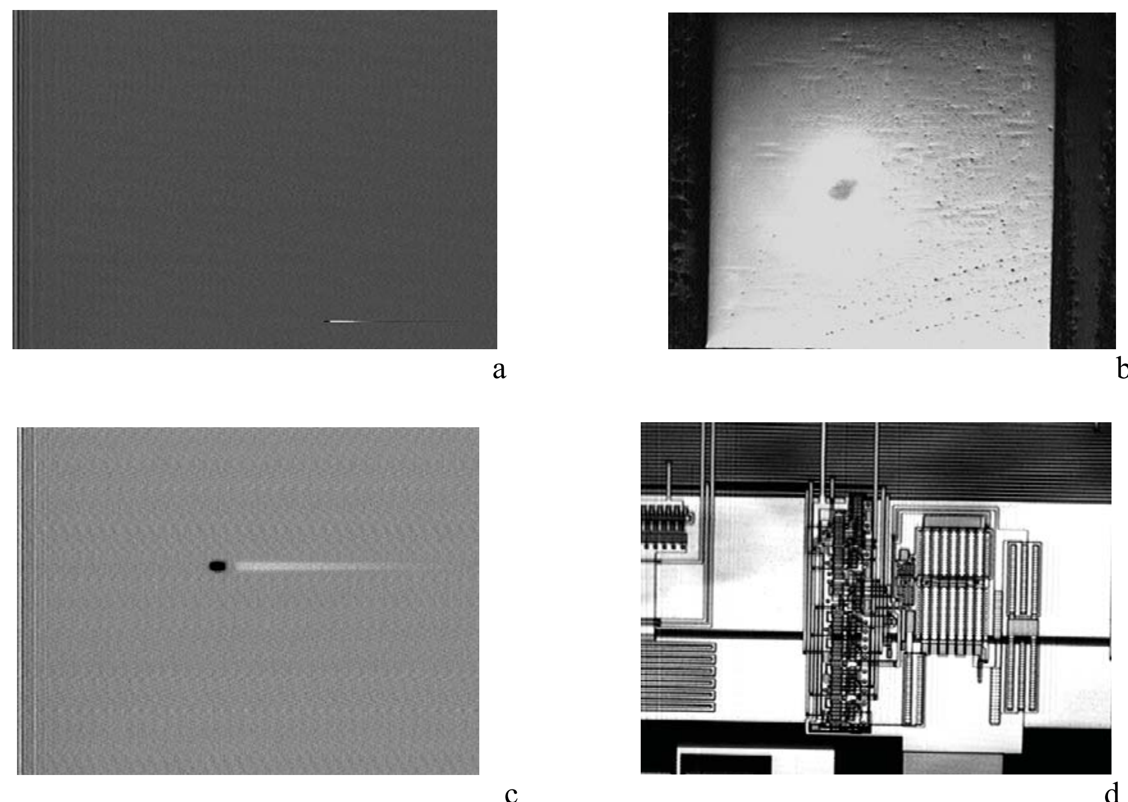


Figure 11. Backside LIVA examination of a microcontroller: LIVA image (a) indicating the area of an open metal-1 to silicon contact, backside reflected image (b) showing the same field of view as Figure 11a, higher magnification LIVA image (c) of the defect in Figure 11a, and a reflected light image (d) of the same field of view as Figure 11c. The defect site is completely covered by a metal-2 power bus from the IC's front side.

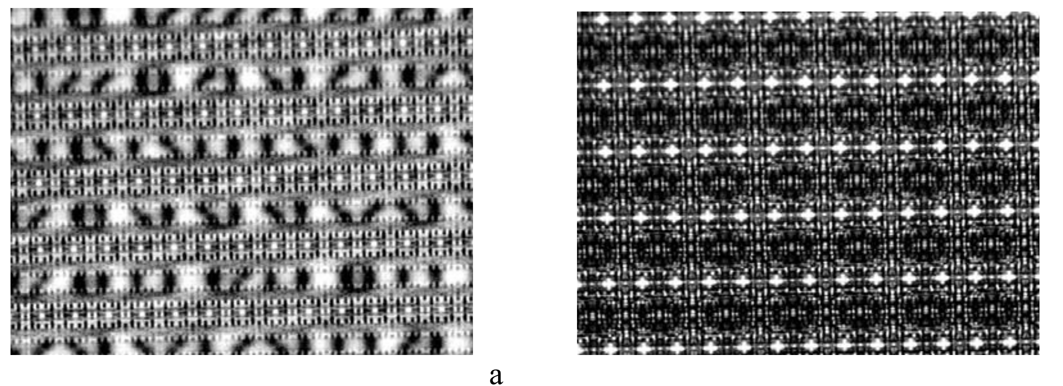


Figure 12. Backside LIVA image (a) showing the logic states of transistors in an SRAM embedded in a microcontroller and backside reflected light (b) image of the same field of view.

Physics Behind Tool Use: OBRICH, TIVA, and SEI take advantage of the interactions of a localized heat source (a focused infrared laser) and two common IC defects (an open or a shorted interconnection). All use an infrared laser wavelength longer than 1.1 µm (the indirect bandgap in silicon) to produce localized heating on an IC without electron-hole pair production. In OBIRCH and TIVA, the localized heating changes the resistance of a short site. If the short site is a conductor, the resistance will increase with temperature. If the short site is a semiconductor, the resistance will decrease with increasing temperature. In any event the resistance of the short site changes with temperature. As the resistance of the short site changes the power demand of the IC changes, assuming the short site has a voltage gradient across it. This effect of localized heating on IC short circuits and the subsequent IC power demand changes was first shown in the OBIRCH (optical beam induced resistance change) technique. TIVA displays an increase in detection sensitivity using the constant current biasing approach applied in CIVA and LIVA. (See the descriptions of CIVA and LIVA in this manuscript for a description of the defect sensitivity advantages of constant current biasing.) Figure 13a displays an entire 1MB SRAM with a short site as indicated by the TIVA image. Note that the short site has been localized in a single image. Figure 13b is a reflected light image of the same field of view for registration. Figure 14 is a backside TIVA/reflected light image pair of a short site. The particle producing the short cannot be seen in Figure 14b because it is on top of the shorted metal-1 interconnections. The example in Figure 14 demonstrates how heat can travel through a metal layer to detect a defect where photons cannot.

Localized heating can also be used to locate open conductor sites from the front and backside of an IC. Open conductor sites are identified through the use of the Seebeck effect. Thermal gradients in conductors generate electrical potential gradients with typical values on the order of mV/K. This is known as thermoelectric power or the Seebeck Effect and refers to the work of Thomas Johann Seebeck (1770-1831). The most common application of thermoelectric power is the thermocouple, which uses the difference in thermoelectric voltages of two different metals to measure temperature. If an IC conductor is electrically intact and has no opens, the potential gradient produced by localized heating is readily compensated for by the transistor or power bus electrically driving the conductor and essentially no signal is produced. However, if the conductor is electrically isolated from a driving transistor or power bus, the Seebeck Effect will change the potential of the conductor. This change in conductor potential will change the bias condition of transistors whose gates are connected to the electrically open conductor, changing the transistors' saturation condition and power dissipation. An image of the changing IC power demands (via constant current biasing) displays the location of electrically floating conductors. For the laser and SOMs used to date a maximum temperature gradient of about 30 ^{o}C has been achieved, with most work producing temperature changes of 10 ^{o}C or less. The resulting small changes in open conductor potential need the sensitive constant current biasing approach for defect detection. Figure 15a shows a backside SEI image of a FIB-opened conductor. Figure 15b is a reflected light image for registration. The open conductor can be seen as well as strong contrast at the metal-polysilicon interconnections. The enhanced contrast is believed to result from the thermopower difference in the two different materials (polysilicon and the metal conductor).

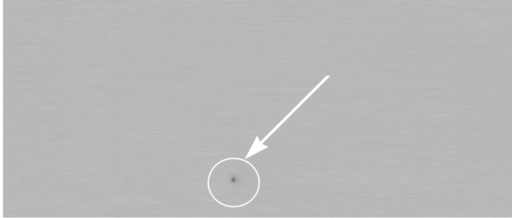
Difficulty to Implement: The equipment necessary to perform OBRICH, TIVA, and SEI imaging are a scanning optical microscope (SOM) with the proper wavelength laser, a constant voltage (OBIRCH) or current source (TIVA/SEI), current amplifier (OBIRCH) or ac coupled voltage amplifier (TIVA/SEI), and electrical connections to the IC under examination. Like LIVA, almost all commercial SOMs will have an auxiliary or OBIC input port suitable for the OBIRCH/TIVA/SEI input. The IC must be biased in a non-contention state, but no complicated vector set is required. The signals can be small depending on the details of the defect (especially for SEI), therefore increased laser power will produce stronger signals for all of the techniques.

Possible IC Damage: No physical damage occurs with OBIRCH/TIVA/SEI examination, but newer, more powerful SOMs have the potential of depositing enough power density to damage ICs. This source of damage can be mitigated by keeping the laser power density below any damage threshold. Die thinning is not necessary for backside analysis, but will increase the signal strength about 10X with each "halving" of the die thickness. Additionally, polishing the backside of the IC die to reduce light scattering and the application of an anti-reflective coating greatly improve reflected light image resolution. All of the backside examples shown above have been polished, but not purposely thinned.

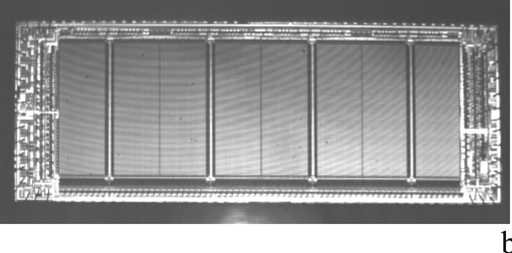
Externally Induced Voltage Alteration (XIVA)
XIVA is not an additional technique as much as it is an alternative biasing scheme that can be used with all in the "IVA" approaches. [14] In XIVA an ac choke circuit is used with a constant voltage supply. The XIVA signal measures changes in the choke circuit with momentary power demand fluctuations. Using XIVA can mitigate problems of large power changes with constant current biasing that can drop the supplied voltage below operational levels. An implementation of XIVA is shown in Figure 16.

Soft Defect Localization (SDL)
Information Tool Yields: SDL [15,16] is a SOM technique that localizes soft defects (resistive vias, leaky gate oxides, oxide defects, process variations, timing marginality, etc.) from the front and backside of an IC.

Physics Behind Tool Use: Soft defects are a failure source that can be extremely difficult to localize. In general, a soft defect will produce a failing signature under certain operating conditions and a passing signature by changing the conditions (temperature, voltage, and operating frequency are common variables). SDL uses a scanned 1.3 µm laser to locally heat the sample IC in a fashion similar to SEI/TIVA/OBIRCH. The 1.3 µm laser is used to avoid photocurrent effects which can mask defects or the property being examined. Heating will cause soft defects to change their electrical properties during dynamic testing, thus changing the pass/fail status of the test. Defects are detected by operating the device and monitoring (imaging) the pass/fail status as the laser is scanned. The resulting SDL image yields the location of the soft defect. Figure 17 is an example of using SDL to localize a resistive interconnection. Figure 18 displays a gate producing a race condition on an IC.

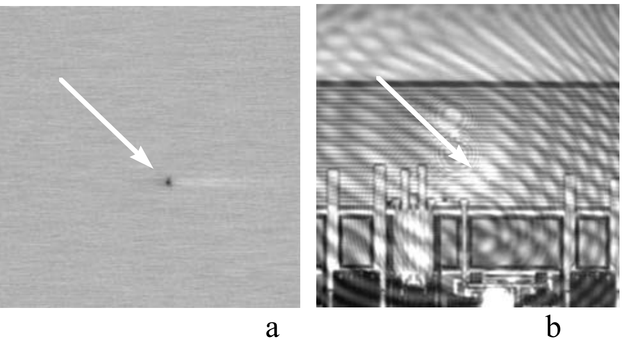


a



b

Figure 13. Entire die front side TIVA image of a 1MB SRAM showing (a) the site of a particle short. A reflected light image (b) is show of the same field of view for registration.



a b

Figure 14. Backside TIVA image (a) localizing a short circuit site on a 1MB SRAM and backside reflected light (b) image of the same field of view. Note that the shorting particle on top of metal-1 is not visible from the backside in (b).

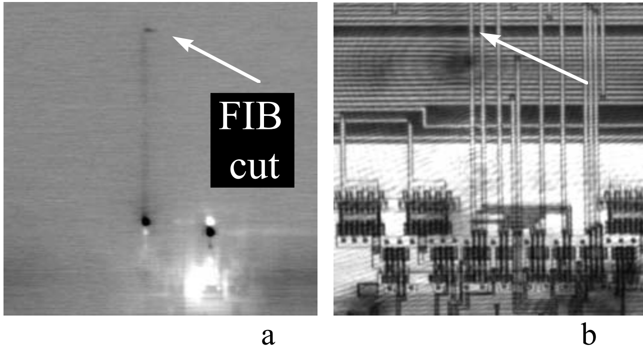


Figure 15. Backside SEI image (a) showing an open conductor resulting from an FIB cut and a reflected light image (b) for registration.

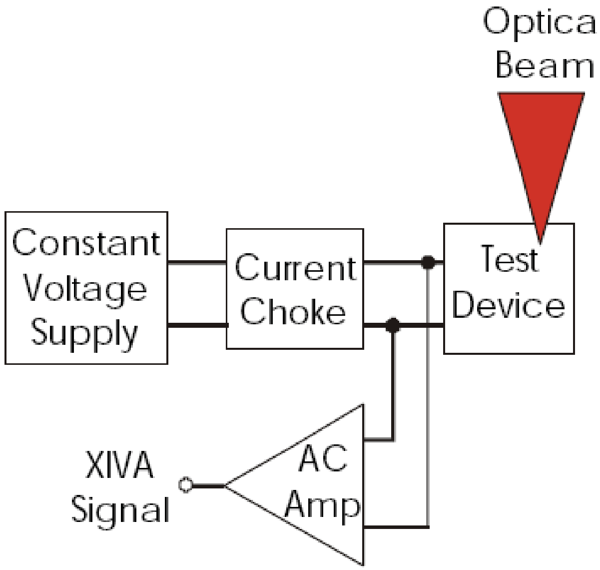
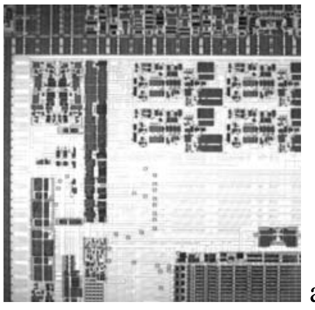


Figure 16. Schematic of XIVA system.

There have been techniques analogous to SDL published. Laser-Assisted Device Alteration (LADA) [17] uses laser induced local electron-hole pair generation for stimulus and has been used for timing marginality. Stimulus Induced Fault Testing (SIFT) [18] proposes a wide range of localized stimulus and detection criteria to localize various aspects of device operation.

Difficulty to Implement: The SOM system needed for SDL is identical to that for OBIRCH/TIVA/SEI. SDL requires the addition of dynamic stimulus and a means to determine the pass/fail condition of the target device.

Possible IC Damage: The damage considerations are similar to those of OBIRCH/TIVA/SEI. Since SDL requires dynamic stimulus, cooling of the sample may be required if self-heating is an issue.



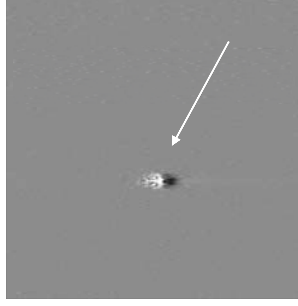


Figure 17. Backside images of (a) reflected light and (b) SDL of the same field of view. The SDL site was identified as a resistive interconnection.

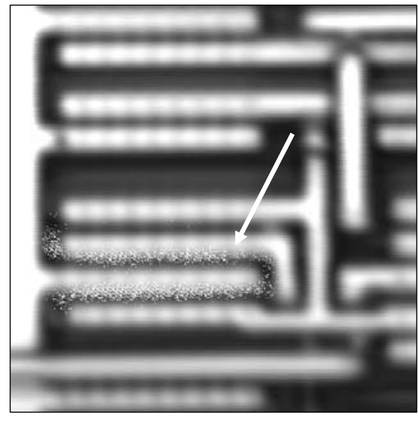


Figure 18. Front side SDL/reflected light combined image localizing a gate causing a parallel path race condition.

Solid Immersion Lenses (SILs)

As with electron beam imaging there have been efforts to improve the spatial resolution of optical beam techniques. This is perhaps even more important as the micron-sized wavelengths of infrared light used for many optical beam techniques has inherent spatial limitations for observing the small feature sizes of modern microelectronic devices. SIL technology has been successful in improving the spatial resolution limits.

SILs essentially increase the light collection efficiency of the optical system, increasing the numerical aperture (NA) which is inversely proportional to the spatial resolution.

The most popular SIL approach is the NAIL (NA increasing lens). [19] An example of a NAIL is shown in Figure 19. The curvature of the NAIL lens, usually Si or GaAs, increases the NA of the optical system. The NAIL/sample interface must be very flat and clean and the sample thickness tailored to the NAIL specifications, usually 100 µm, + 20 µm. The NAIL can be moved from location to location as needed. Figure 20 shows a thermal stimulus image on a 90 nm sample with better than 200 nm resolution using a SIL with a 1340 nm laser.

Two other SIL approaches have been described. One uses a machined hemisphere on the back of the silicon substrate to form a FOSSIL (forming substrate into SIL). [20] A second approach uses a machined diffraction or Fresnel lens. [21] Both the FOSSIL and Fresnel approaches have no interface problems because the SIL is part of the substrate, but neither are movable like the NAIL.

Recent reports show even improved spatial resolution is possible using an aperture to block the central axis of the SIL. The evanescent or "dark-field" approach may produce images with resolutions approaching 70 nm with 1 µm light. [22]

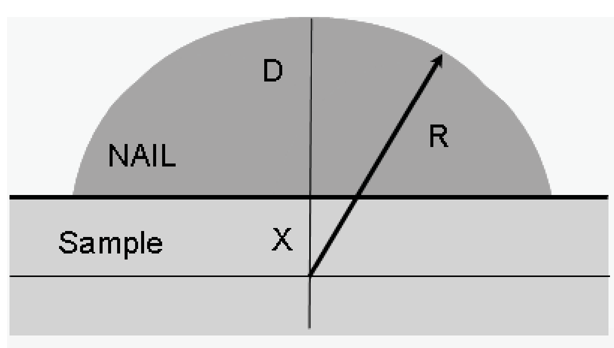


Figure 19. Schematic of a NAIL on a sample substrate.

Conclusion

The spatial resolution, depth of focus, ease of use, ready availability, and localizable interactions of a focused beam with microelectronic technologies have made beam-based analytical methods powerful techniques for IC failure analysis. The ability to perform backside analysis with infrared light sources has generated a number of useful techniques in recent years. Because of their great effectiveness and future potential, beam-based electron and photon probing will continue to be mainstays in IC failure analysis for the foreseeable future and topics for development and improvement.

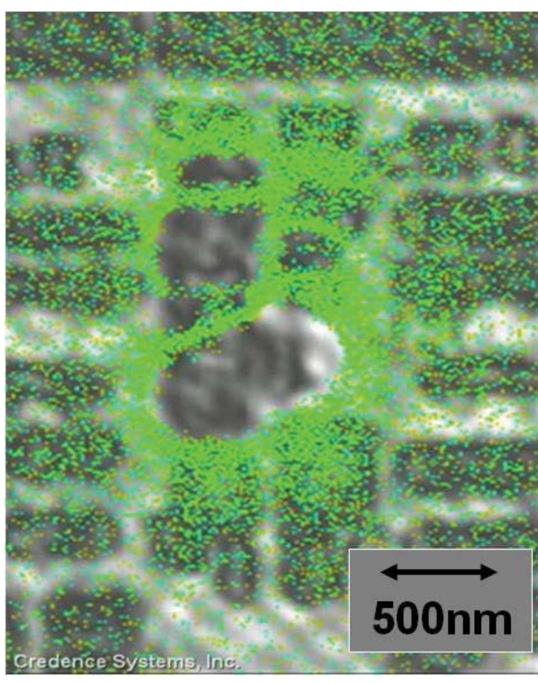


Figure 20. Example of SDL using a SIL showing better than 200 nm spatial resolution using a 1340 nm laser source. The green contrast indicates a passing condition. (Image courtesy of Steven Kasapi, Credence Systems, Inc.)

Acknowledgements

The author would like to thank Richard E. Anderson, Daniel L. Barton, Michael R. Bruce, Ann N. Campbell, James Colvin, Christopher L. Henderson, Paiboon Tangyunyong, Jerry M. Soden, and David J. Stein for their careful review of and valuable contributions.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security

Administration under contract DE-AC04-94AL85000.

References

1. J.R. Beall, *Voltage Contrast Techniques and Procedures" and "Electron Beam-Induced Current Application Techniques*, Microelectronic Failure Analysis Desk Reference, Supplement Two, November (1991)
2. E.I. Cole Jr. et al., *Advanced Scanning Electron Microscopy Methods and Applications to Integrated Circuit Failure Analysis*, Scanning Microscopy, vol. 2, no. 1, 133-150, (1988)
3. J.I. Goldstein et al., Scanning Electron Microscopy and X-Ray Microanalysis, A Textbook for Biologists, Materials Scientists, and Geologists, 2nd Edition, Plenum, New York, 89, (1992)
4. E. I. Cole Jr., *Resistive Contrasting Applied to Multilevel Interconnection Failure Analysis,* VLSI Multilevel Interconnection Conference, 176-182, (1989)
5. C.A. Smith et al., *Resistive Contrast Imaging: A New SEM Mode for Failure Analysis,* IEEE Transactions on Electron Devices, ED-33, No. 2, 282-285, (1986)
6. E. I. Cole Jr., *A New Technique for Imaging the Logic State of Passivated Conductors: Biased Resistive Contrast Imaging*, International Reliability Physics Symposium, 45-50, (1990)
7. E.I. Cole Jr. and R.E. Anderson, *Rapid Localization of IC Open Conductors Using Charge-Induced Voltage Alteration (CIVA),* International Reliability Physics Symposium, 288-298, (1992)
8. E.I. Cole Jr. et al., *Low Electron Beam Energy CIVA Analysis of Passivated ICs,* International Symposium for Testing and Failure Analysis, 23-32, (1994)
9. J.R. Michael, *What's New in Electrons, Ions, and X-Rays,* AMFA Workshop, (2009)
10. K.S. Wills, T. Lewis, G. Billus, and H. Hoang, *Optical Beam Induced Current Applications for Failure Analysis of VLSI Devices,* International Symposium for Testing & Failure Analysis, 21-26, (1990)
11. E.I. Cole Jr. et al., *Novel Failure Analysis Techniques Using Photon Probing With a Scanning Optical Microscope,* International Reliability Physics Symposium, 388-398, (1994)
12. K. Nikawa and S. Inoue, *New Capabilities of OBIRCH Method for Fault Localization and Defect Detection,* Sixth Asian Test Symposium, 219-219 (1997)
13. E.I. Cole Jr. et al, *Backside Localization of Open and Shorted IC Interconnections*, Proceedings of IEEE International Reliability Physics Symposium, 129-136, (1998)
14. R.A. Falk, *Advanced LIVA/TIVA Techniques*, International Symposium for Testing & Failure Analysis, 59, (2001)
15. E.I. Cole Jr. et al., *Resistive Interconnect Localization*, International Symposium for Testing & Failure Analysis, 43-50, (2001)
16. M.R. Bruce et al., *Soft Defect Localization (SDL)*, International Symposium for Testing & Failure Analysis, 21-27, (2002)
17. J.A. Rowlette and T.M. Eiles, *Critical Timing Analysis in Microprocessors Using Near-IR Laser Assisted Device Alteration*, International Test Conference, 264-273, (2003)
18. J. Colvin, *Functional Failure Analysis by Induced Stimulus*, International Symposium for Testing & Failure Analysis, 623-630, (2002)
19. S.B. Ippolito et al., *High Spatial Resolution Subsurface Microscopy*, Applied Physics Letters, 4071-4073, (2001)
20. T. Koyama et al., *High Resolution Backside Fault Isolation Technique Using Directly Forming Si Substrate into Solid Immersion Lens*, International Reliability Physics Symposium, 529-535, (2003)
21. F. Zachariasse and M. Goossens, *Diffractive Lenses for High Resolution Laser Based Failure Analysis*, International Symposium for Testing & Failure Analysis, 1-7, (2005)
22. P. Perdu, *Failure Analysis Year in Review*, International Reliability Physics Symposium Reliability Year in Review, (2010)

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Electron Beam Probing

John TL Thong
Centre for IC Failure Analysis & Reliability, Faculty of Engineering,
National University of Singapore, Singapore.

1. Introduction

Electron beam probing, a means to measure voltage waveforms on internal IC nodes using a focused electron beam, has its roots in the voltage contrast (VC) phenomenon which was first observed by Ken Smith from Cambridge University who reported "the variation of the mean signal intensity (in an SEM image) as the potential of the specimen is changed" in his 1956 Ph.D thesis. The next 3 decades saw many developments in understanding the phenomenon, instrumentation and detectors, and applications to voltage contrast imaging and waveform measurements on ICs, which culminated in the launch of the first commercial e-beam tester add-on to an SEM by Lintech in 1982, and the first full-fledged CAD-linked e-beam tester by Schlumberger Technologies in 1987.

The 1980's saw possibly the greatest interest in the technique due to a confluence of several additional factors – ICs were getting to the level of complexity that a technique to replace mechanical probing was needed, most of the fundamental issues and problems associated with e-beam waveform measurements had been understood and tackled, while the recognition that CAD navigation and automation are essential features prompted the development of user-friendly tools that did not require an SEM expert to sit alongside the IC designer or failure analyst. In a sense, e-beam testing tools and methods arrived at the right time. By the end of the 80s, challenges in the waiting started to emerge – ICs were getting functionally so complex that special test patterns were needed to make e-beam waveform measurements tractable, while the increase in the number of wiring levels made buried signal nodes less and less accessible. Concepts for "design for e-beam testability" had to be incorporated at the outset of IC design by manufacturers who had the intention of using e-beam probing to validate, characterize, and debug their first silicon. For the failure analyst, e-beam probing being a front-side technique, the problem of decreasing observability is somewhat alleviated by pervasiveness of FIB tools to access buried nodes. The lack of observability from the front-side also means that voltage contrast imaging techniques for fault localization find limited application except in the simplest of devices. For waveform measurements from the backside, the Laser Voltage Probe (LVP) provides an alternative [1].

In this article, we will first describe the underlying theory of e-beam voltage measurements before considering the practical aspects of applying the technique to present-day ICs.

2. E-beam waveform measurements

a. Open & closed loop measurements

In e-beam probing [2], a focused electron beam (using an SEM or a dedicated e-beam tester) with energy typically between 500eV and 2keV, is used to probe the node of interest on the IC, in a manner analogous to mechanical probing. However, the focused electron beam (the primary beam) does not convey the electrical signal *per se* – instead, the emitted secondary electrons (SEs) carry with them information about the voltage of the node being probed. The principle underlying e-beam voltage measurements is the direct correspondence between the energy spectrum of SEs emitted from a point on the specimen and the surface potential of that point. When irradiated by the primary beam, a material emits SEs, the bulk of which have energies E below 20eV that are distributed with an energy spectrum N(E) with a peak typically around a few eV (Figure 1). Now, if the specimen (assumed to be a conductor) were biased by V_s = +3V, the energy of the SEs would decrease by 3eV, as a result of which the entire SE spectrum shifts by -3eV. Likewise, a negative bias on the conductor would result in a corresponding positive shift of the spectrum.

Hence changes in the conductor voltage would translate directly to changes in the position of the SE spectrum. To make use of this phenomenon, a SE spectrometer is required. E-beam testers make use of high-pass retarding field spectrometers in one form or another. The principle of operation is

illustrated with reference to Figure 2, which shows an idealized hemispherical retarding-field analyzer. By biasing the retarding grid at an appropriate voltage, only those SEs with sufficient energy (the portion represented by the shaded part of the SE distribution) are transmitted by the grid and detected. The so-called “S-curve” represents the portion of the detected SE signal as a function of retarding grid bias V_{ret} (Figure 3).

If the retarding grid bias is fixed, typically around the steepest part of the S-curve, then the detected signal intensity increases with a negative bias on the specimen, and decreases with a positive bias (Figure 4), giving a qualitative measure of the specimen voltage that could be quite decent provided the voltage swings are small enough to avoid gross non-linearities in the S-curve. This is of course a voltage contrast signal, but in this idealized case only discriminates on the basis of the SE energy as opposed to other phenomena like trajectory deflection and local field effects. This so-called open-loop mode of measurement can quite often give a decent logic waveform measurement [3], bearing in mind that positive voltages (logic high) give rise to lower detected signal levels, and hence an inverted representation of the logic waveform.

Quantitative voltage measurements make use of a feedback loop that keeps the detected signal level constant by changing the retarding grid bias in tandem with the changes of the specimen voltage. With this closed-loop strategy, the measured waveform appears superimposed on the retarding-grid bias set point (Figure 5).

In practice, while the above basic principles of waveform measurements remain the same, a practical implementation for e-beam voltage measurements on ICs has to address many other issues. One fundamental problem is to extract as much of the SE spectrum as possible from the IC surface. Unlike the large specimen discussed earlier, the conductor tracks or points to be probed are much smaller in dimension and are often surrounded by other wiring tracks with dynamic voltages on them. The proximity between tracks means that a potential barrier can appear above a point being probed that prevents the lower energy SEs from leaving the IC surface. To minimize this problem, an extraction field is applied at the specimen surface to pull the SEs into the analyzer. There are other issues associated with SE trajectory effects caused by strong electric fields arising from neighboring tracks that can give rise to measurement crosstalk, but analyzer developments have largely tackled this problem and have reduced crosstalk to an acceptable level in modern e-beam testers.

One of the key advantages of e-beam probing as opposed to mechanical probing is its contactless, non-loading feature. While active electrical probes can have input capacitances as low as 20fF, such loading can nonetheless alter the circuit timing on sensitive nodes. In e-beam probing, non-loading conditions can be achieved by setting the primary (probing) electron beam energy such that the total number of electrons emitted (comprising SEs and backscattered electrons) balances the total number of primary electron impinging on the probing point. Depending on the material, the beam energy at which this occurs (the “second crossover”) ranges typically between several hundred eV to around 2keV. Even if the beam energy were not set at this energy exactly, the low probing beam current and beam-on duty cycle mean that few electrons are injected into or extracted from the probed point. On the other hand, on a charge-storage node such as a DRAM capacitor cell, the non-loading feature of e-beam probing is an important feature if managed properly.

b. Sampling and signal-to-noise issues

The Achilles’ heel of e-beam voltage measurements is the poor signal-to-noise ratio of the technique. Noise in the primary electron beam (limited beam current) ultimately limits the voltage measurement accuracy. Using a spectrometer, the minimum measurable voltage ΔV is related to the primary beam current I_{PE} and the measurement bandwidth, Δf by

$$\Delta V = \frac{C}{\sqrt{I_{PE} / \Delta f}}$$

where C , the spectrometer constant, characterizes spectrometer performance at its operating point. Typically, with 5nA current focused into a 0.1µm diameter spot, the bandwidth associated with 1mV sensitivity is only around 100 Hz. As $\Delta V \propto \Delta f^{-1/2}$, the corresponding sensitivity at 1 MHz bandwidth is 0.1V which would permit rough real-time measurements of low-speed logic waveforms.

Thus high-speed e-beam measurements are invariably based on sampling techniques and many of the sampling techniques used in digital oscilloscopes can be adopted. Waveform sampling

is carried out by pulsing the electron beam, which is easily achieved by deflecting the electron beam over an aperture in the electron-optical column to "blank" the beam (hence the beam pulsing system is sometimes called a "beam blanker"). On commercial systems, pulse widths as short as 30-40 ps can be obtained corresponding to a measurement bandwidth of around 10-12 GHz. An equivalent-time sampled reconstruction of the waveform is then obtained by delaying the sampling pulse relative to the trigger input.

However, the acquisition of a noise-free measurement requires that the measurement be averaged over many periods of a repetitive waveform until the desired signal-to-noise ratio is achieved, i.e., it is not possible to obtain a single-shot "real-time" sampled measurement that is noise-free at the same time. The increasingly long logic sequences would mean that waveform acquisition could be a very slow process. Lower logic voltage swings further exacerbate the problem by requiring more signal averaging to reduce measurement noise to an acceptable level.

On a positive note, in contrast to mechanical probing, e-beam probing can be carried out at many test nodes on the IC as repositioning of the e-beam probe simply involves deflecting the e-beam to the desired location and / or moving the device stage. With CAD support and image registration (an SEM image is readily available), many points can be probed automatically without user intervention.

c. Capacitive coupling voltage contrast

It is not only possible to measure voltages directly on conductors using e-beam probing, but also on conductors covered by insulators. While it is possible to penetrate through a dielectric to the underlying wiring track with a high-energy electron beam, thereby forming a conductive path through the insulator, this is not advisable due to potential damage to MOS devices. Moreover, the large conductive volume formed in the dielectric can short out neighboring tracks. The preferred method is to exploit capacitive coupling voltage contrast (CCVC) using the low energy electron beam energy range for measurements on covered tracks to maintain the minimally invasive nature of e-beam probing. A first level description of CCVC assumes that SEs are emitted from an equipotential region on the insulator surface that is made conductive by the irradiating primary beam. This region is seen to be capacitively coupled to the underlying (and neighboring) track(s) (Figure 6).

In reality, the situation is much more complex than presented by this simple model. A proper understanding of CCVC requires a rather detailed description of insulator charging and is beyond the scope of this article. Consider VC imaging, where the beam is being scanned over the specimen. At low magnifications, CCVC is readily observed in VC images with slow decay in the image contrast after a transient. At higher magnifications, the effective current density over the scanned area increases, and the VC decays more rapidly. At very high magnifications, we approach the situation of a stationary electron probe where the irradiated area is effectively that of the beam probe. Here the current density is very high, and any VC signal vanishes rapidly after a transient in the underlying conductor voltage. The time taken for the insulator surface voltage to decay to 1% of its original value after a transient is known as the storage time, and is inversely proportional to the irradiation current density. For e-beam voltage measurements, this is a real problem as the total electron irradiation dose required to obtain a single voltage measurement point on the waveform can easily exceed the storage time.

One way to reduce the effective current density is to avoid using a stationary probe, but to scan a small area above the conductor to be measured. An alternative is to deposit a small metal island (using FIB for example) to act as a capacitor plate that can be probed by the e-beam. An effective way of minimizing the problem of voltage droop during e-beam sampling is, as far as is possible, to randomize the sampling phase between consecutive sampling events. With random sampling, there is a high probability that consecutive samples are obtained from different logic states, causing the insulator surface to charge and discharge randomly, and hence averaging out negative and positive going voltage droops, albeit at the expense of introducing statistical noise into the measurement.

So far, the above issues only address CCVC measurements on a large conductor covered by an insulator. For measurements on real IC wiring tracks, we have to further take into account geometrical effects of the track geometry. Where the insulator thickness is comparable to the width of the track being measured, then the potential at the insulator surface being probed has contributions not only from the underlying track, but from neighbouring tracks (or the substrate if there are no other tracks in its vicinity) as well – the probing point is effectively capacitively

coupled to all surrounding conductors through the intervening dielectric. This gives rise to measurement crosstalk from dynamic signals, and electrostatic screening from static potentials on neighbouring tracks. This problem is particularly severe on modern ICs where not only are the geometries of wiring tracks small, but the tracks are also deeply buried, often under several layers of more superficial lines. Indeed, attempts to obtain meaningful CCVC measurements on level 1 wiring on 2-layer metal IC technology in the 1980's had already been fraught with difficulties.

Hence measurements via CCVC on anything but the most superficial tracks are not advisable. To perform measurements on buried nodes, it is necessary to bring the node up to the surface (if this had not already been done for validation of first silicon) using an FIB tool. Such an approach assumes that the node is physically accessible from the IC surface, but often overlayers of metallization obscure the underlying nodes, limiting the utility of e-beam or mechanical probing under such circumstances.

3. Preparation for E-beam Probing

The utility of e-beam probing depends greatly on the level of information and knowledge about the device that is available to the user. For silicon characterization and debug, the entire design and simulation database is available, and indeed test nodes for e-beam probing and e-beam DFT techniques may have been incorporated into the first silicon. Here the silicon debug engineer can make use of the CAD linkages and navigation tools available on a modern e-beam tester (Figure 7) to navigate to the probing point, acquire the waveforms, and compare them against simulations. Measurements can be performed at the wafer level using wafer probe cards, or on packaged devices.

At the opposite extreme of knowledge availability, failure analysis on vendor components faces the challenges of limited information on the device provided by the vendor, the usual problems of packaged and passivated devices, and scarcity of failed samples from the field. For such FA tasks, e-beam techniques are primarily used for fault localization, and can be demarcated into image-based techniques, and waveform-based techniques.

On simple devices with few wiring levels where CCVC can show up voltages on subsurface wiring, voltage contrast image techniques such as stroboscopic imaging and logic state mapping can be used to assist in fault localization. Although image-based approaches can be used without CAD layout data, they can be extremely time-consuming as large amounts of image data are required. Image comparison techniques are particularly useful when comparing images acquired from the same device under different device conditions to identify marginal failures. Alternatively, images can be acquired from the failing device and a golden device and compared within the same phase range of the test program. As the fault takes time to propagate to the device pins and be detected by the ATE, image acquisition should start several clock cycles earlier. E-beam probing on the offending node can then provide more detailed on timing and voltage levels. Faults that originate and propagate on buried wiring that are unobservable will require several test patterns that will activate the defect to eliminate faults not immediately related to the defect. Waveform-based approaches use comparisons between measured and reference waveforms, but require full CAD layout and netlist or schematic CAD data. The reference waveform can either be from simulation data, or a golden device. The software driven guided-probe approaches include the European funded ESPRIT project, ADVICE [2,4] and the IDA software package [5].

a. DUT preparation

The obvious prerequisite for the use of e-beam probing (or indeed mechanical probing for that matter) is that the measurement node must be accessible to the e-beam. At the package level, this means that the IC die frontside must be exposed to the electron optics in vacuum, often with a small gap between the IC and the spectrometer. This presents a problem for packaged devices using flip-chip BGA technology, and there are few options but to remove the die. Special test sockets are commercially available with access apertures and pogo-pins to contact to partially or fully-removed balls (Figure 8). Without the package and associated heatsinking, the die has to be cooled, for which a liquid-cooled heatsink in direct contact with the die backside would meet the size constraints of the device load module. Another type of problem may be encountered with center-bond ICs where the bond wires would obscure the underlying IC die. Such devices can be re-bonded with half the die uncovered, allowing for testing half the device at a time. Devices on wafers present few problems, and e-beam tester manufacturers provide wafer probes that are compatible with the working distance of the e-beam prober.

A more serious issue is that of node observability at the die level. If an IC is to be characterized or debugged using e-beam probing from the outset, then the IC designer needs to decide which are the critical nodes for probing bearing in mind that observability in a MPU design with 8 or more levels of interconnects is severely limited, especially if access to local wiring is desired. The use of upper metallization levels as power planes further complicates matters. DRAMs with 4 levels of metal are more manageable, and with their considerably simpler test sequences, are more amenable to e-beam probing.

For the failure analyst, a partial solution to accessing buried nodes is through the use of an FIB tool to mill a hole. The hole can then be refilled with metal to the surface using a masking / metallization step or the FIB tool itself in conjunction with ion beam induced deposition from an organometallic precursor, typically containing platinum or gold. Such a strategy is only viable on a relatively unoccupied area of the IC die or where the node is not too deep. The introduction of such an access via will have capacitive loading and crosstalk implications that can be significant on high-speed nodes. Indeed, on a marginal fail, such an invasive procedure can change the failure characteristics entirely.

Yet another alternative is to perform e-beam probing from the backside of the die. To carry this out, a tapered probe hole can be FIB milled to reach level 1 metal from the backside of a thinned die. The tapering is required for the SEs to escape the hole, otherwise only a small fraction can leave from a high aspect-ratio hole due to obfuscation by the sidewalls [6]. It is found that measurements acquired from the backside can be cleaner than those from the frontside due to electrostatic screening by the grounded substrate that eliminates crosstalk from adjacent wiring and field effects from high switching currents on the device surface. As with all backside probing techniques, the issue of heat removal is a serious one, especially in the case of e-beam probing where the backside of the die is now in vacuum.

On a relatively superficial node or a passivation-covered conductor, a metal pad deposited by FIB on the insulator surface can be used to improve waveform measurements via CCVC. Having a real metal pad in place of an equivalent scanned area avoids some of the problems associated with insulator charging dynamics and improves quantitative waveform measurements.

b. ATE Interface

The ideal solution for driving the device under test is the use of a standard ATE directly docked to the e-beam prober via a load-module (Figure 9 a, b). The load module should ideally preserve the bandwidth and delay characteristics at the ATE test head interface, and moreover provide the means to cool the device, there being no other heat dissipation mechanism in the vacuum chamber.

On less sophisticated DIY setups, the DUT can be connected to an ATE via coaxial cables and mass-terminated coaxial cable assemblies provide a solution where the I/O count is large. Homemade load boards should take into account that heat dissipation will be a problem unless some means is provided to remove the heat from the DUT, such as a cold metal strap. With cabling, the DUT is no longer at the tester driver, and significant propagation delays will mean that the ATE is unlikely to operate correctly, with only the e-beam measurement providing an indication of the correct signal feed to the DUT. Without immediate feedback from the regular test scheme, it is difficult to determine whether the ATE, connections, or the device is at fault. The device must then be returned to the regular ATE to verify the device operation.

c. Test Patterns for E-beam probing

Measurements using e-beam probing requires averaging the acquired signal over many repetitions of the waveform to achieve a relatively noise-free measurement. The e-beam prober (sampling system) must then be triggered by an external signal that is synchronized to the signal being measured. Due to the increasingly long logic sequences, regular ATE test vectors may not be suitable for e-beam measurements, and special e-beam test vectors with short repeat cycles should be developed where possible to maximize waveform measurement throughput. Fortunately, the requirement for manufacturing testability means that the same design for testability features such as scan paths can be exploited to support e-beam debug and FA. Nonetheless, the long waveform acquisition times call for a reasonable guess to be made as to where and when a particular failure is occurring in the device through the pin signals and a collection of FA tools before using e-beam probing.

4. Conclusions

From its inception, e-beam probing and its roots in voltage contrast have spanned nearly 50 years of history. The tool and its limitations are well-

understood, and the current state-of-the-art in commercial e-beam testing instrumentation can meet the requirements for design debug and failure analysis of at least the next generation of devices. The biggest challenge to the application of e-beam probing comes from the limited observability of device nodes from the front-side of the die, which increasingly restricts its use to validation and characterization with selected test nodes incorporated, or FA on designs of lower complexity with the aid of FIB tools.

Acknowledgments

The author would like to thank Ted Lundquist of Credence Systems Corp. for his comments and for providing photographs for this article.

References

1. M. Paniccia, T. Eiles, V.R.M. Rao, W.M. Yee, "Novel optical probing technique for flip chip packaged microprocessors", Proc. Intl. Test Conf. 1998, pp. 740-747 (1998)

2. J.T.L. Thong (ed.), Electron Beam Testing Technology (Plenum, New York), (1993).

3. H. Wang, H. Koike, M. Ikeda, K. Kanai, "Open loop method for waveform acquisition on electron-beam probe systems", in Advances in Microelectronic Device Technology (Qin-Yi Tong, Ulrich M. Goesele; Eds.), Proc. SPIE 4600, pp 154-159, (2001)

4. M. Melgara, M. Battu, P. Garino, J. Dowe, Y.J. Vernay, M. Marzouki, F. Boland, "Automatic location of IC design errors using an e-beam system", Proc. Intl. Test Conf. 1988, pp. 898-907 (1988).

5. A.C. Noble, "IDA: A tool for computer-aided failure analysis", Proc. Intl. Test Conf. 1992, pp. 848-853 (1992).

6. R. H. Livengood, P. Winer, J. A. Giacobbe, J. Stinson, J. D. Finnegan, "Advanced micro-surgery techniques and material parasitics for debug of flip-chip microprocessor" Proc. 25th ISTFA (1999).

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Failure Localization with Active and Passive Voltage Contrast in FIB and SEM

Ruediger Rosenkranz
Fraunhofer Institute for Nondestructive Testing, Dresden, Germany

Introduction

The common Passive Voltage Contrast (VC) localization method is based on brightness differences of conducting structures in SEM and FIB images [1-10] and can be used for failure localization issues. The Active Voltage Contrast method (AVC) as it was described as Biased Voltage Contrast by Campbell and Soden [11] and by the author in [12] offers even more localization possibilities. Particularly the PVC methods became widely accepted in the semiconductor FA community. Nearly all labs make use of it. There are several Voltage Contrast mechanisms in SEM and FIB and not all of them are always completely understood. This paper will give a comprehensive overview over all phenomena related to this subject. The multiple advantages, possibilities and limits of VC failure localization are systemized and discussed.

Passive Voltage Contrast in FIB

Basic principles of PVC localization in FIB

PVC means brightness differences in FIB (or SEM) images of structures having different electrical properties. The die isn't connected to any power supply or signal sources (passive) and is in most cases partially deprocessed.

Failure localization with PVC is based on the fact, that floating structures are charging up under the influence of the primary beam in FIB and SEM (Fig. 1). They appear dark. In contradiction to that grounded structure appear bright. In example 1 one part of a contact chain appears dark because it is floating due to a contact open. The bright part is grounded by design. In this context the term "grounded structure" means, that it has any contact to any substrate region or to a large network in contradiction to the conventional meaning where "grounded" indicates a resistance to mass potential of almost zero.

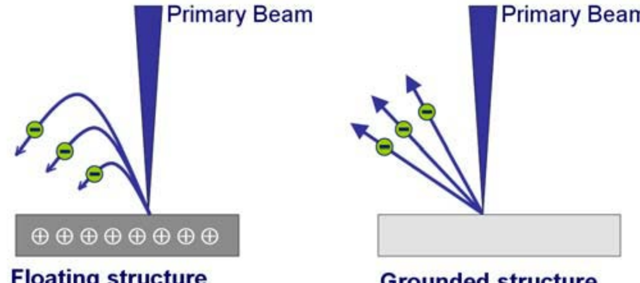


Fig. 1: Insulated structures are charging up positively under the influence of leaving secondary electrons. Immediately after that the majority of produced secondary electrons are prevented from leaving the sample by the electric field. These structures appear dark in the image. Grounded structures do not charge and appear bright because of the high secondary electron yield.

If a structure that usually supposed to be floating appears bright, then it is in urgent suspicion to be faulty. On the other hand, if a structure that supposed to be grounded appears dark, it is also a strong indication for a fault. The first case is not just limited to structures that are floating a priori because of the design. Structures can be made floating by deprocessing as it is shown for a gate and its contact in Fig. 2.

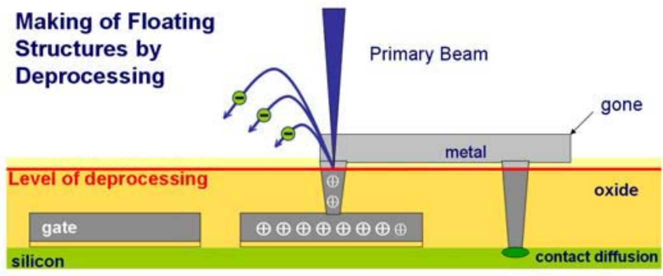


Fig. 2: Structures can be made floating by deprocessing

Another possibility to make structures floating is cutting conducting lines with a FIB cut as it is shown in Fig. 3. If such a line is bright instead of dark, the unwanted substrate short can be found by cutting the line into pieces (see Example 2). The remaining bright part is bearing the fault.

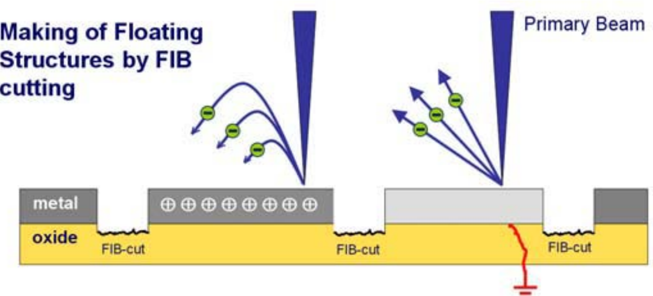


Fig. 3: Structures can be made floating by cutting

For other failure localization issues it can be helpful to ground a structure if it isn't a priori grounded because of the design. Fig 4 shows, how a short between two adjacent floating structures can be detected by grounding one of them by FIB.

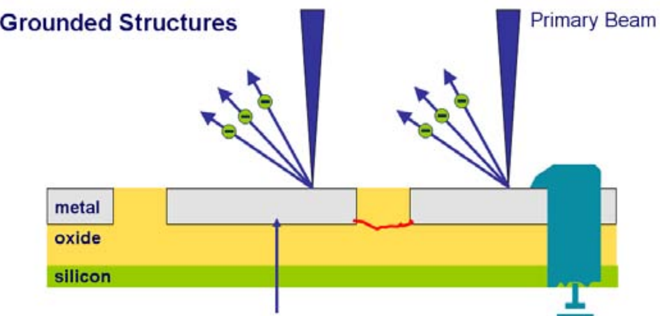


Fig 4: Localization Example: The right structure was grounded by milling a FIB hole through it followed by a metal deposition. The left structure is supposed to be dark because it is floating. If it isn't, there must be a short (red) to the adjacent one.

floating = dark structures	Application examples
by deprocessing	to find single contact opens
by FIB cutting	to find shorts (binary search strategy)

grounded = bright structures	Application examples
by design	to find open well contacts
by FIB connecting	to find opens in contact chains

Table 1: Contrast Generation Summary

Table 1 gives an overview for all four cases discussed including application examples.

Advanced mechanisms of PVC generation

If diodes, capacitors and transistors are involved, PVC generation and interpretation is much more complicated. In the case of substrate contacts the doping has to be taken into account. When large structures are investigated, their capacitance plays an important role. Even transistors can affect the PVC generation when they are broken. In Fig. 5 the PVC generation at S/D contacts is shown. The diodes, formed by n-contacts in p-wells are reverse biased. The charging generated by the leaving secondary electrons can't flow down to the substrate. These contacts appear dark. The p-diffusions in n-wells are forming forward biased diodes. Here the positive charges can flow easily to ground, such contacts are bright.

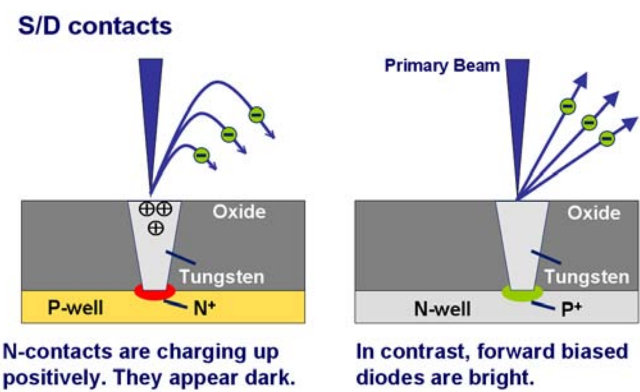


Fig 5: PVC generation at S/D contacts

Contacts connecting the wells (p-diffusion in p-well and n-diffusion in n-well) always appear bright because there is no diode. In example 3 the different contrasts of different contact types are to be seen.

Fig 6 shows how the capacitance of a structure can influence the generation of PVC. If a structure is very large, e.g. a very long conducting line, then there are not enough charge carriers produced in order to raise the potential of this structure significantly. It stays bright like a grounded structure or becomes gradually dark because it is charging slowly during the image scan like it is seen in example 4.

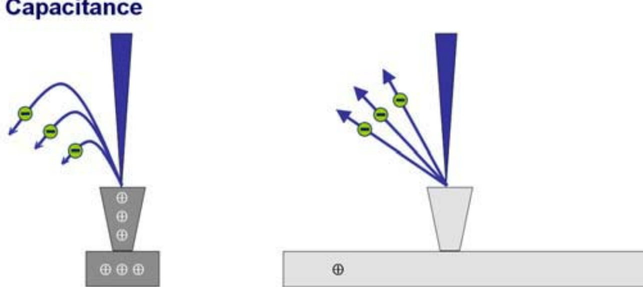


Fig. 6: Large structures are like big capacitors which can't be charged sufficiently during the short time the ion beam scans over them

Fig. 7 shows an equivalent circuit diagram for a structure having a certain capacitance C and a certain resistance R to ground. The current source is the sum of positive ions hitting the structure and of the secondary electrons leaving it. The solution of the corresponding differential equation delivers the voltage U_c which is building up at the structure.

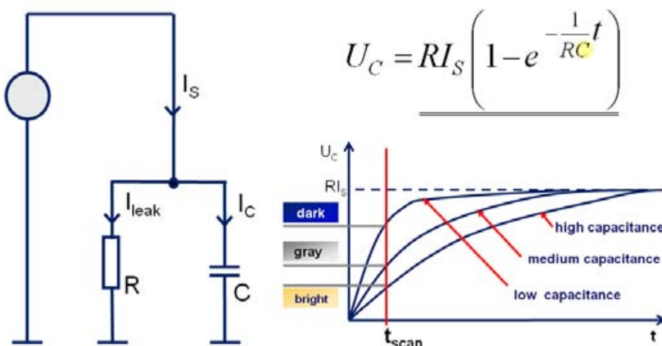


Fig. 7: Equivalent circuit diagram, solution of the corresponding differential equation and graphical representation for different structure capacitances

The time t_{scan} is the total time the ion beam scans over the structure. The larger the capacitance of the structure the brighter it appears.

The formula for the voltage U_c is also well suited for studying the influence of the leakage current from a structure to ground. These leakage currents can occur due to surface conductivity after deprocessing or due to the circuitry this structure is connected to.

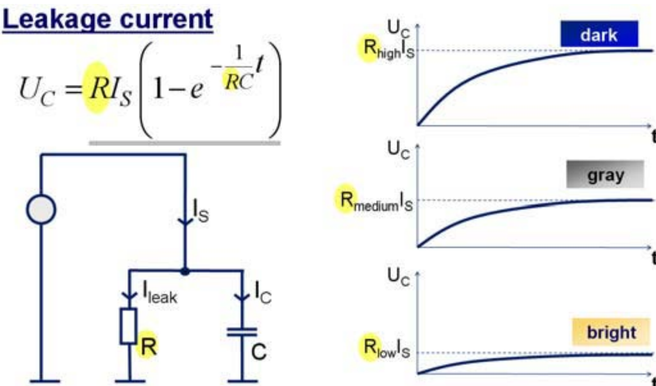


Fig. 8: Influence of leakage resistance to the voltage building up at the structure during ion beam scan

In Fig. 8 graphical representation of $U_c(t)$ is to be seen for a fixed capacitance C but different values of the leakage resistance R. For low values of R the Voltage U_c can't build up to an amount which is necessary to prevent the secondary electrons from leaving. They can reach the detector resulting in a bright appearing structure. In our experience R_{medium} is approximately 1 GOhm. Below this order of magnitude no dark structures can be observed.

At transistors we observed another interesting PVC phenomenon. The contrast of s/d contacts changed depending whether the corresponding gate was grounded or not.

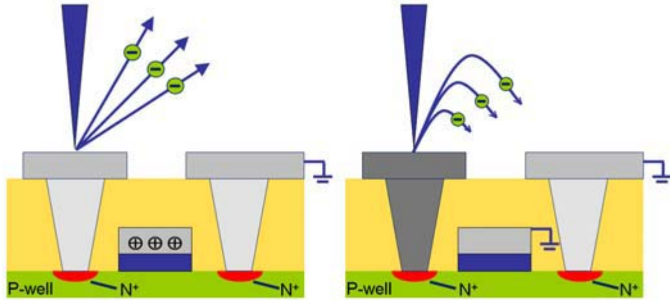


Fig. 9: The s/d contacts of n-MOS transistors are showing different PVC depending on whether the gate is opened or closed

The left source contact in the transistor of Fig. 9 is normally expected to be dark because of the reverse biased diode under it. In reality it is bright because the charges can escape through the open transistor to the drain which is grounded. The gate was charged positively through the oxide during the ion beam scan and thus opened. When this gate conductor is grounded by a FIB cut nearby (no metal fill is

necessary) the source contact suddenly turns dark as it is demonstrated in example 5.

The following table gives a summary of the PVC phenomena discussed in this chapter.

Implants	p-well	n-well
p-diffusion	bright	bright
n-diffusion	gray	bright
n-diffusion next to an open gate	bright	bright

Capacitance	high	medium	low
VC:	bright	gray	dark

Leakage Current	high	medium	low
VC:	bright	gray	dark

Table 2: Advanced mechanisms of PVC generation

Some experiences to share
Sometimes it is hard to achieve a noticeable PVC in the Focused Ion Beam. Knowing the mechanisms behind PVC generation, one can derive some basic rules to get or to improve the contrast between charging and floating structures. The goal to be achieved is either to reduce the leakage currents or to improve the charge carrier generation.

[1] The first thing one should try is to increase the primary beam current, because it improves the net charging of the structure. It should be noted, however, that there is always a tradeoff between damaging the structure by milling it away and the achievement of good PVC. In case of doubt the influence of a strong ion beam should be tested at an uncritical structure of the same kind.

[2] The next thing one should try is to change the sample orientation. If structures are to be charged that are longer than wider, it is more effective to align their long side horizontally so that the horizontally scanning ion beam is hitting a big portion of the structure abruptly which can overcome possible leakage currents and can increase the chance to charge it up sufficiently. The easiest way to do it is to change the scan direction by software buttons instead of rotating the sample stage.

[3] The scan speed also plays an important role. Slower scan speeds improve the chance of sufficient charging.

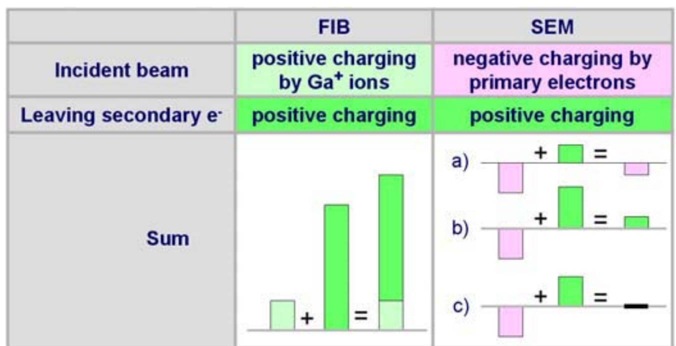
[4] A tilted surface is emitting more secondary electrons than a horizontal one. This improves the charge carrier generation. Therefore stage tilting may help.

[5] When lines are to be FIB cut, one should always use the fluorine GIS because it prevents the redeposition of conducting material on the surface and thus it reduces the leakage currents (see example 6)

[6] Typical beam currents for the most localization issues are 10-50 pA. For low magnifications a current up to 100 pA can be used, for special problems at high magnifications 1-10 pA may be best.

PVC in SEM versus FIB

In table 3 is explained why the charge generation in FIB is much stronger then in SEM.



	FIB	SEM
Incident beam	positive charging by Ga^{+} ions	negative charging by primary electrons
Leaving secondary e^{-}	positive charging	positive charging
Sum	+ =	a) + = b) + = c) + =

Table 3: Comparison of Charging in FIB and SEM. In SEM both negative and positive charging is gettable depending on primary beam energy. In FIB just positive charging is possible but this is much stronger.

In SEM, the incoming primary electrons and the leaving secondary electrons are partially compensating each other resulting in less charging then in FIB. Because this depends on primary beam current, sample material, sample orientation, magnification and other factors the charging in SEM can be both positive and negative, which can be an advantage in comparison to the FIB.

The accelerating voltage of the electron beam has the largest influence on the type of charging as it is shown in Fig. 10.

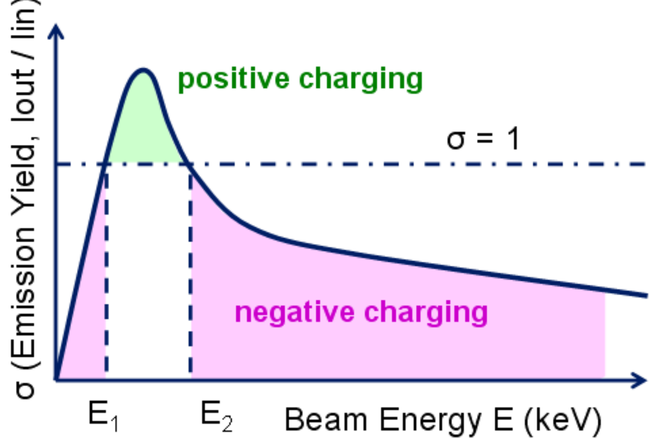


Fig. 10: Influence of the beam energy to the emission yield in SEM. The threshold voltages E_1 and E_2 depend on sample, sample material and SEM configuration like beam current, chuck to ground resistance, vacuum, etc.

The threshold voltage E_1 is in our experience approximately 0.5 kV and E_2 is around 2 kV. In Example 7 SEM images with two different beam energies show different PVC. Depending on the problem to be solved either FIB PVC or SEM PVC is the method of choice. In table 4 advantages and disadvantages of both methods are set against each other.

	FIB	SEM
Visibility of PVC	good	poor
Charging	just positive	both
Cutting	very helpful	not possible
Sample degradation	noticeable	negligible
Marking of lokalized failures	yes	no

Table 4: advantages and disadvantages of FIB and SEM PVC

Inline e-beam inspection

An important application of Passive Voltage Contrast in SEM is the inline e-beam inspection technique. In an inline SEM a retarding field at the chuck provides low landing energies of electrons which are necessary for VC localization. The primary energy of the electrons right after having left the cathode is not affected by the retarding field and is high as it is necessary for a high SEM resolution. With a charge control plate mounted above the wafer both positive and negative surface charging can be detected (Fig. 11).

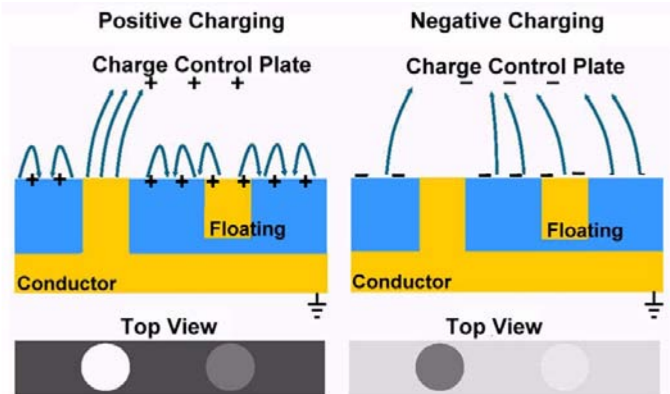


Fig. 11: Both positive and negative charges can be detected in inline e-beam inspection tools

In examples 8 and 9 is shown that both types of charging can be helpful for detection of faults at wafer level.

Active Voltage Contrast (AVC)

Sometimes contact chains are not completely open but have a high resistance far beyond the specification because of a partially open contact. The leaking current through this faulty contact prevents charging and its localization with PVC. This is shown in Fig. 12.

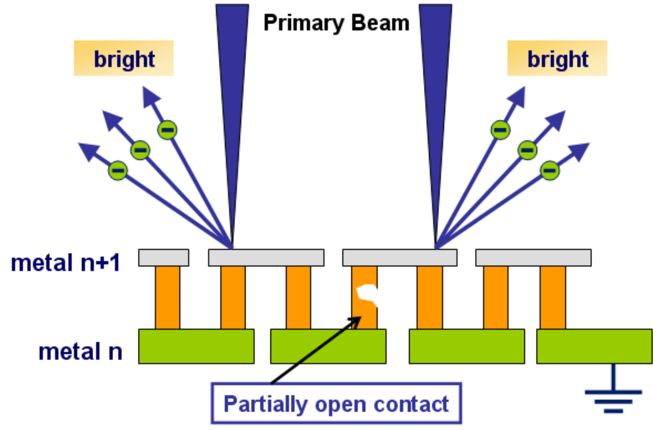


Fig 12: The contact chain part left from the faulty contact can't be charged because charges can escape through it faster as they were generated by the primary beam. PVC localization just works on totally open contacts.

With a nanoprober in SEM/FIB as shown in Fig. 13 definite voltages can be applied to definite structures.

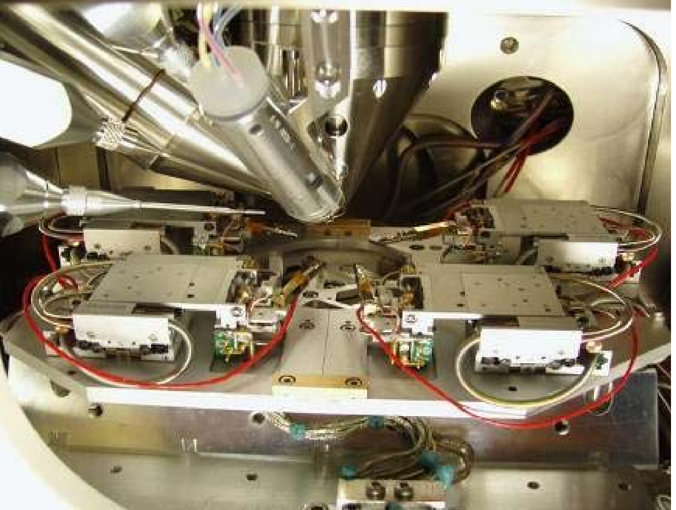


Fig. 13: 4-probe nanoprober, mounted on a stage of a Dualbeam FIB.

With an external voltage applied to such a chain as shown in Fig. 14 there are sufficient charges in the left part of the chain and at the faulty contact is a leap in potential according to Ohm's law.

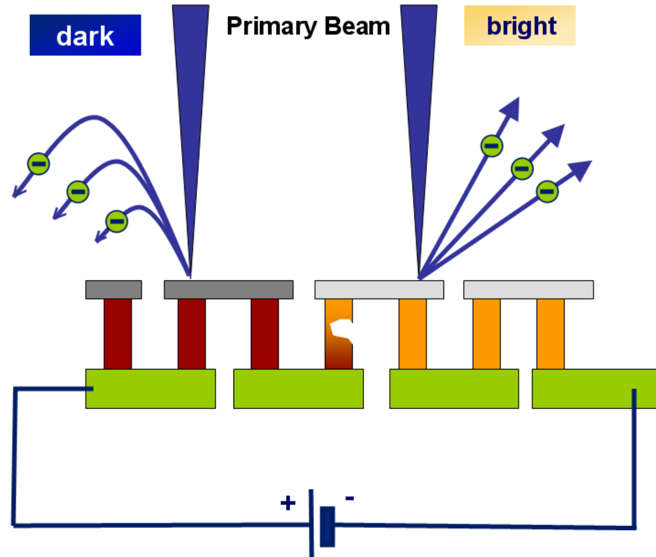


Fig. 14: Principle of Active Voltage Contrast. With an external voltage applied to a high-resistive contact chain different potentials emerge on the two parts of the chain resulting in a clearly distinguishable Voltage Contrast

Currently there are two basic types of in-chamber nanoprobing systems available – stage mounted ones and chamber mounted ones. Both have advantages and disadvantages depending on the purpose they are to be used for. Chamber mounted system do not allow any stage movement after the probes have been lowered. It doesn't matter for device characterization purposes where 4, 6 or even 8 probes have to contact structures in very small region.

For AVC sometimes test structures are contacted at their probe pads because it is easier than to do so at tiny conducting lines. Probe pads often are several hundreds of micrometers away from the structure itself. Here a stage mounted system is very helpful, because the pad is placed under the beam first in order to lower the tips under visual control and after that the stage is moved together with the tips until the test structure is under the beam in order to search for AVC.

Stage mounted systems have some restrictions regarding stage moving range and sample size. Some prober models do not allow the use of the load lock due to geometrical restrictions. On the other hand, this prober type can easily be used at any other tools of the same type, just an electrical feedthrough flange is necessary at each tool. In Table 5 advantages and disadvantages of both prober types are listed.

	Stage mounted	Chamber mounted
Stage movement	Stage movement with contacted probe tips	No stage movement with contacted probe tips
AVC	• Any distance between probing area and failing area possible • In large failing areas it can be searched for AVC by stage moving • Tilting in Dualbeam tools possible	• Probing and failing area must be in the range of beam shift • No tilting
Device Characterization	Well suited	Well suited
Restrictions	Stage moving range and sample size can be restricted	Nearly no influence to stage functionality
Flexibility	Easy to switch to another tool of the same kind (just feedthroughs must be on both tools)	Less flexibility
Time	For sample exchange the use of loadlock is not always possible. Chamber venting and evacuating takes time.	No chamber venting unless no probe tip change is necessary.

Table 5: Advantages and disadvantages of the two basic nanoprober types.

The number of probe tips for AVC application is not critical. It can be performed even with a single

tip, two are sufficient whereas a device characterization requires 4, 6 or 8 tips.

Another interesting question is whether one should use a SEM or FIB for AVC localization issues. A SEM with 2 kV primary beam voltage can mostly be used. As an exception, when lines are to be cut, a FIB is needed. A tool with both SEM and FIB column would be ideal.

The last question to be discussed here is which voltages should be applied to the structures. Since most of the secondary electrons have energies of 2-4 eV and below, a probing voltage of +5V is o.k. A general rule for the probing voltage is “the higher the better” but at voltages above 7V we occasionally observed irreversible gate oxide damage and the AVC image changed dramatically preventing any further localization.

Summary

For failure localization Voltage Contrast both in SEM and in FIB can be used. Both methods have their advantages and disadvantages.

Knowledge of all facts influencing the VC generation (capacitance, leakage, doping and circuitry) is very helpful for successful failure localization.

Active Voltage Contrast is more laborious but offers completely new opportunities for failure localization.

References

[1] J. Colvin, “A New Technique to Rapidly Identify Low Level Gate Oxide Leakage in Field Effect Semiconductors Using a Scanning Electron Microscope”, Proc. 16th Int. Symp. Testing and Failure Analysis (1990), p 331

[2] O. D. Patterson, H. Wildman, A. Ache, K. T. Wu, “In-Line Voltage Contrast Inspection of Ungrounded Chain Test Structures for Timely and Detailed Characterization of Contact and Via Yield Loss”, ISTFA 2005, Proceedings, 401-406

[3] V. Liang, H. Sur, S. Bothra, “Passive Voltage Contrast Technique for Rapid In-line Characterization and Failure Isolation During Development of Deep-Submicron ASIC CMOS Technology”, ISTFA 1998, Proceedings, 221-225

[4] O. D. Patterson, J. L. Drown, B. D. Crevasse, A. Salah, K. Harris, “Real Time Fault Site Isolation of Front-end Defects in ULSI-ESRAM Utilizing In-Line Passive Voltage Contrast Analysis”, ISTFA 2005, Proceedings, 591-599

[5] Nishikawa, N. Kato, Y. Kohno, N. Miura, M. Shimizu, "An Application of Passive Voltage Contrast (PVC) to Failure Analysis of CMOS LSI Using Secondary Electron Collection", ISTFA 1999, Proceedings, 239-243

[6] Caiwen Yuan, Susan Li, Andy Gray, “Method of Failure Site Isolation for Flash Memory Device Using FIB Passive Voltage Contrast Techniques”, ISTFA 2005, Proceedings, 202-205

[7] Cha-Ming Shen, Shi-Chen Lin, Chen-May Huang, Huay-Xan Lin, Chi-Hong Wang “Couple Passive Voltage Contrast with Scanning Probe Microscope to Identify Invisible Implant Issue”, ISTFA 2005, Proceedings, 212-216

[8] Tctsuya Sakai, Noriaki Oda, Takashi Yokoyama, “Defect Isolation and Characterization in Contact Array-Chain Structures by using Voltage Contrast Effect,” IEEE International Symposium 1999

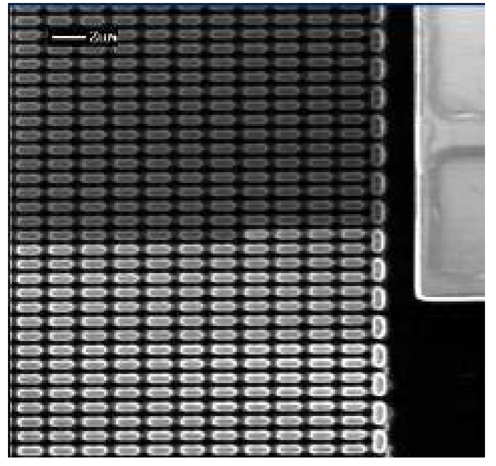
[9] Jone C Lee, C. H. Chen, David Su, J. H. Chuang “Investigation Of Sensitivity Improvement On Passive Voltage Contrast For Defect Isolation”, ESREF 2002

[10] R. Rosenkranz, “FIB Voltage Contrast for Failure Localization on CMOS Circuits – an Overview”, 8th European FIB User Group Meeting 2004

[11] A.N. Campbell and J.M. Soden, “Voltage Contrast in the FIB as a Failure Analysis Tool”, Microelectronic Failure Analysis Desk Reference 4th Edition, 1999, 161-167

[12] R. Rosenkranz, S. Döring, W. Werner, L. Bartholomäus "Active Voltage Contrast for Failure Localization in Test Structures", ISTFA 2006, Proceedings, 316-320

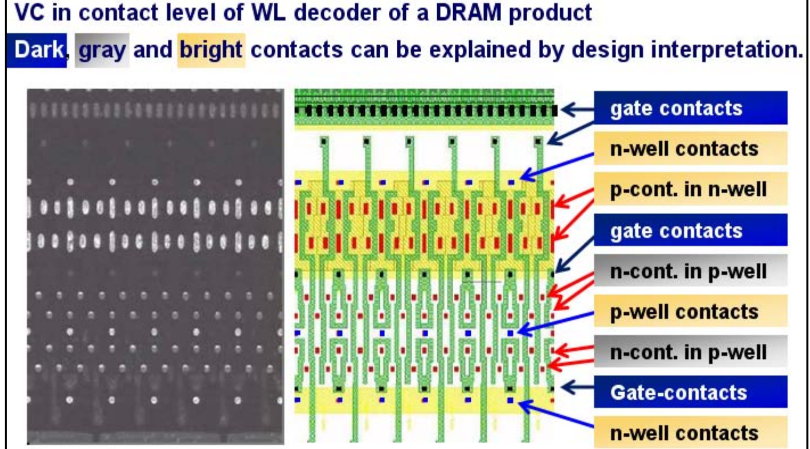
Appendix: PVC and AVC localization examples



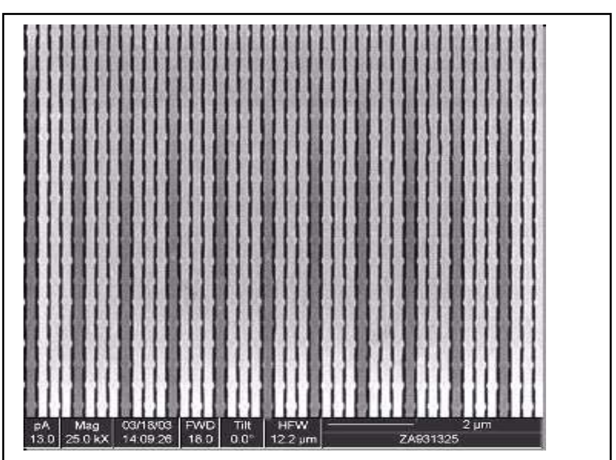
Leaking part of the metal line identified by successive cutting

Leaking contact found after FIB metal etch

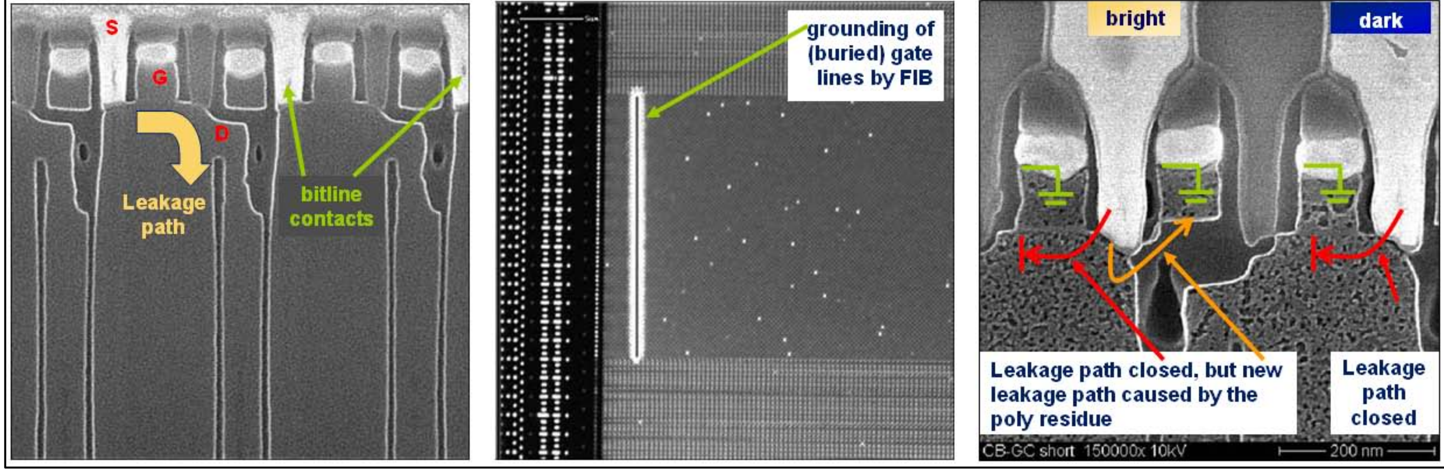
Example 1: Open in a contact chain. *Example 2: Finding a leaking contact by successive cutting.*



Example 3: Different contrasts at different contact types.

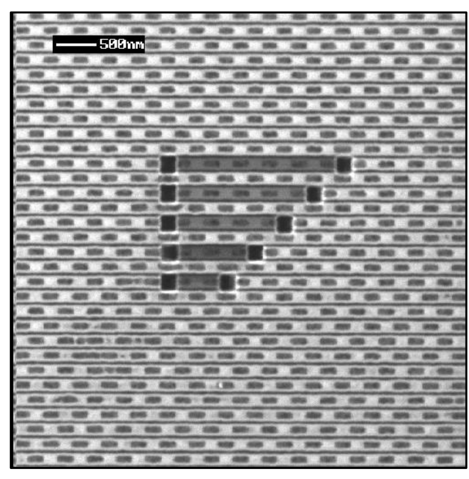


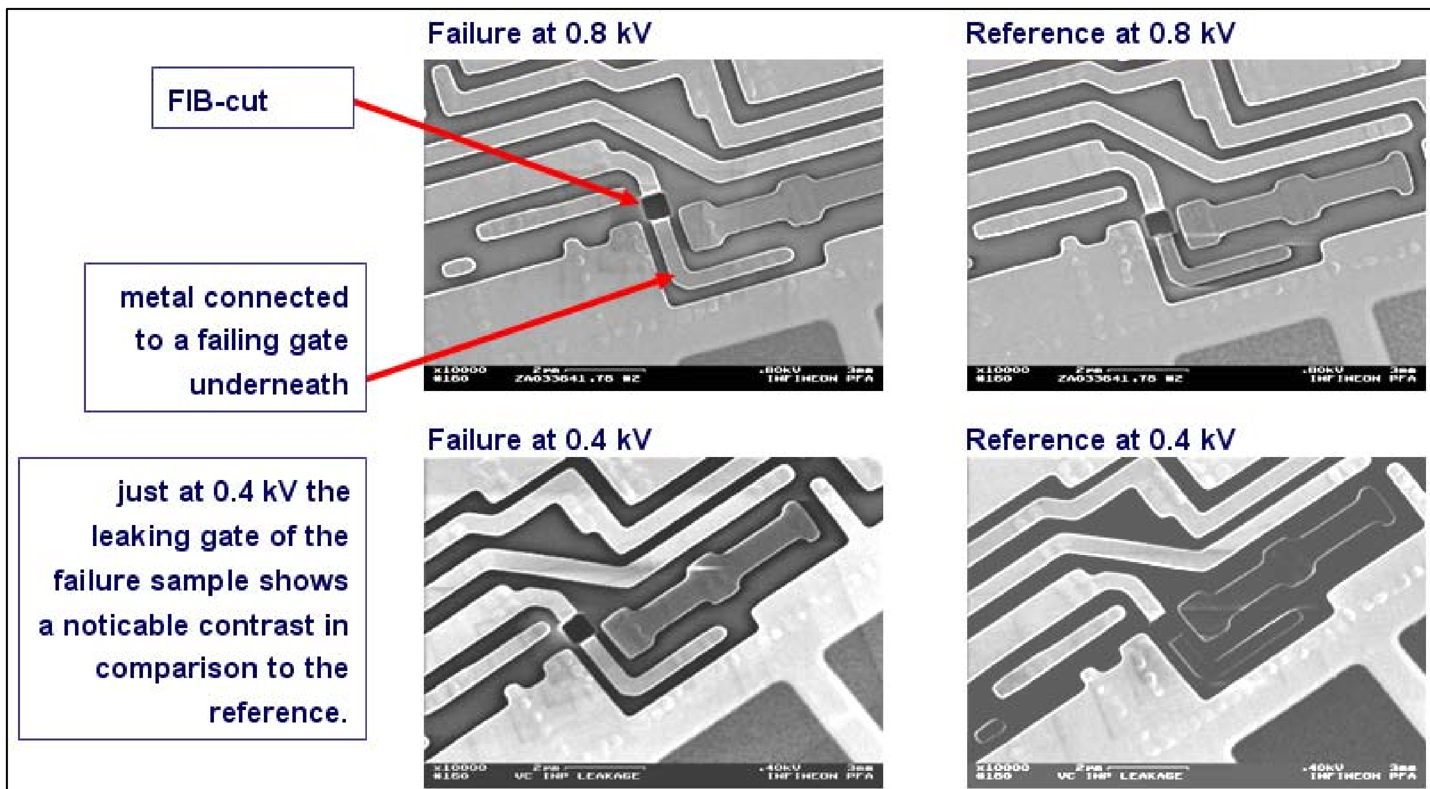
Example 4: One in four of these lines has a higher capacitance than its neighbors. It is charging up during the scan.



Example 5: Bitline contacts of a DRAM. When the gates of these npn-transistors are closed by grounding, the source (or bitline) contacts become dark. The only contacts remaining bright are these which are accidentally connected by poly residues to such grounded gates (white dots in the center image). This is a typical failure mechanism in Deep Trench DRAMs.

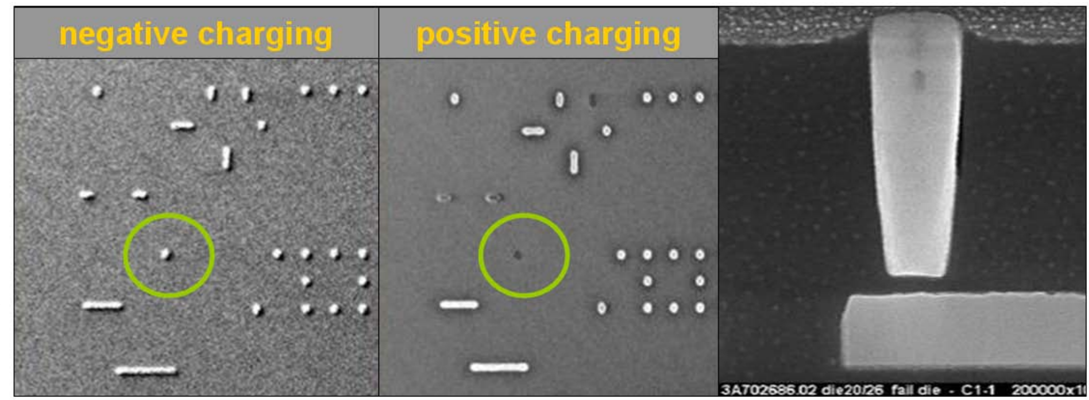
Example 6: *Isolated pieces of conducting lines are scanned without (left) and with (right) iodine valve opened. The latter prevents leakage paths on the surface by resputtering and allows the pieces to charge up positively.*



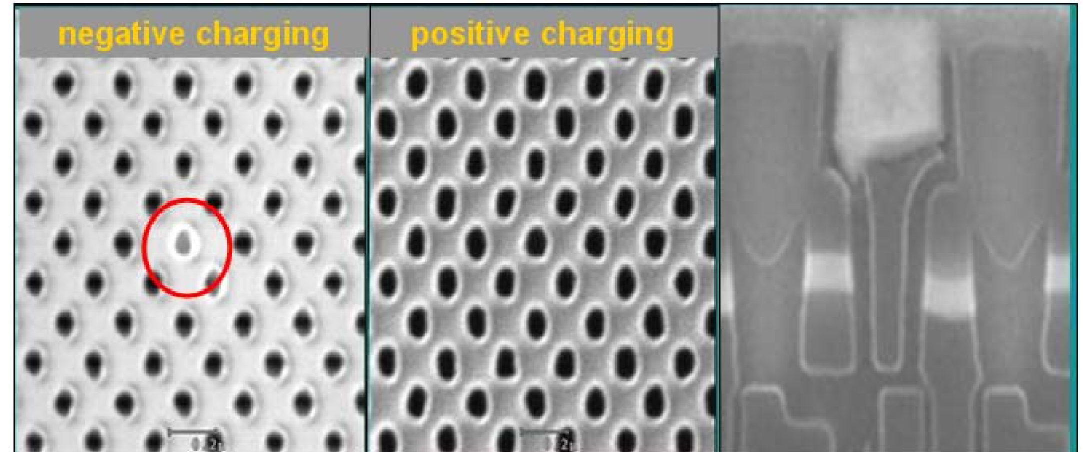


Example 7: PVC in SEM. Just at 0.4 kV noticeable PVC at the failing structure is to be seen.

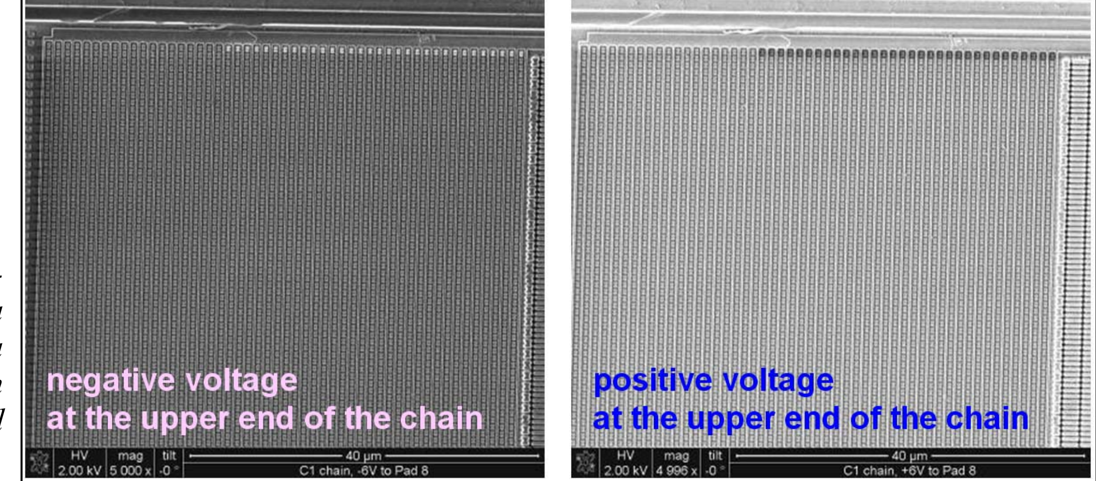
Example 8: Inline e-beam inspection. This contact open can be seen in positive charging only.



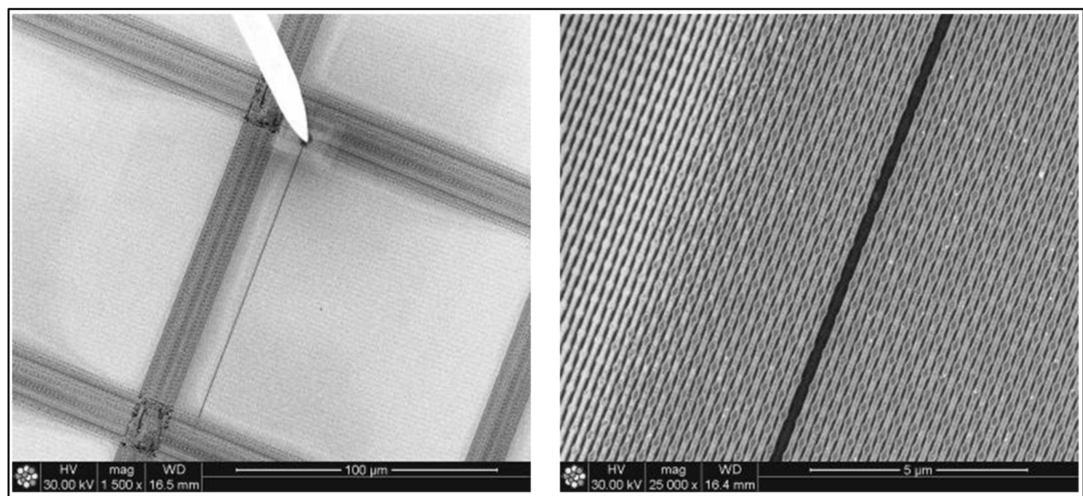
Example 9: Inline e-beam inspection. Contact holes that are not etched deep enough can be seen in negatve charging only.



Example 10: Active Voltage Contrast. With a probe tip a negative and a positive voltage has been applied to the upper end of an open contact chain.



Example 11: Active Voltage Contrast. A positive voltage is applied to a bitline of a DRAM product. If there was a bitline short, one of the adjacent bitlines would also appear dark.



Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Fundamentals of Photon Emission (PEM) in Silicon – Electroluminescence for Analysis of Electronic Circuit and Device Functionality

Christian Boit
TUB Berlin University of Technology, Berlin, Germany

1) INTRODUCTION

Photon Emission (PEM) has proven to be the most successful tool for localization of fails in electrical functionality of ICs also it is based on very faint light emission events. It has emerged almost 20 years ago as a very simple and not overly expensive technique that recorded emission positions as indicator or of failing circuitry quickly, much easier and inside silicon more sensitive than the alternative liquid crystal thermography.

Nowadays we have an optical path and detectors adapted to the challenges of the technique, which has increased applicability further even with the reduction of IC supply voltage. By understanding the physical mechanisms of electroluminescence behind the light emission, we can associate a much more exact fail hypothesis just from the PEM result. A very important breakthrough happened in the early nineties with the growing knowledge in which operation modes fully functional active devices produce light emission. This has finally led to the use of photoemission to track signal propagation in picosecond time resolving detectors.

Another challenge of functional circuit analysis techniques has increased the importance of PEM and all of it's derivatives: The necessity to access this information through chip backside, due to new package technologies like flip chip, and the increasing number of interconnect levels. PEM, being an optical technique, could quite easily be adjusted to this purpose, by just detecting the part of the emission spectra that silicon is transparent for. Also this transformation required knowledge of the emission and detector physics.

This article deals with the physical background of PEM. Here, the reader encounters the large variety of physical interactions in operating active devices and how to make distinctions between them. This explains why PEM is the premiere IC functional localization technique with the highest potential in design debug and failure analysis.

2) BASICS

Electroluminescence in semiconductors occurs in two basic mechanisms that can both be studied in the two operation modes of a pn-junction in silicon:

1) **Relaxation accompanied with light emission** of mobile charge carriers that have picked up a considerable amount of kinetic energy **in an electrical field (F-PE)**. This is an intraband process because it is all happening in the same energy band. This effect is most commonly used in PEM techniques for CMOS circuits.

2) **Radiative Band-Band Recombination** of electrons and holes **(R-PE)**. Recombination involves carriers from conductance and valence band and is called an interband process.

Fig 1 shows the I-V curve of a pn junction. F-PE requires some voltage so the carriers can pick up the necessary amount of kinetic energy. This mechanism is therefore correlated to reverse bias operation. Usually the reverse current is very small. In order to get a photon emission rate above detection limit, the current needs to be increased. So the device needs to either operate in avalanche condition (not ideal for PEM because it is an instable condition and the device may be destructed during measurement), or a leakage current increases the current to an observable level.

A forward biased junction injects many minority carriers and increases the recombination activity (R-PE) already at low voltages to observability level. This effect can be very efficient and is used in LEDs. But in silicon it is a faint effect. A look at the band structure

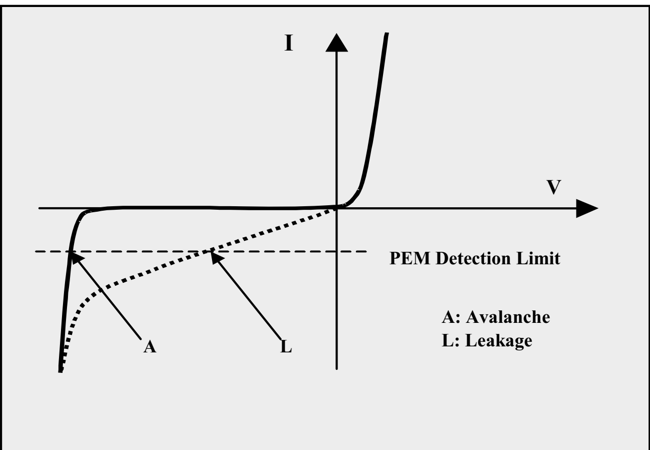


Fig 1: I-V Curve of pn-junction and PEM detection in reverse bias

(***Fig 2***) explains why. In the band diagram as function of a local coordinate (left side) the recombination processes 1 and 2,3 seem to be equivalent. Both fulfill the energy conservation law. A look at the momentum coordinate (right) shows: Silicon is an indirect semiconductor, the bottom of the conduction band and the top of the valence band have different momentum coordinates, a momentum gap for recombination. The momentum conservation law makes the radiant recombination process 1 far less probable than process 2,3.

A localized center in the bandgap has due to the uncertainty principle a blurred momentum coordinate. This raises the probability

for non-radiant recombination via centers several orders of magnitude above the one for radiant band-band recombination which requires a third partner with low energy and high momentum, typically a phonon, for the momentum exchange.

Nevertheless, a few recombination processes are radiant, and the lowered pn-energy barrier in forward bias condition to a level (***Fig1***) that allows the majority carriers to diffuse into the alternate part as minority carriers (minority carrier injection), which is accompanied by strong electron-hole recombination activity. The recombination activity via centers close to midgap level is enhanced proportional to the excess minority carrier concentration and the radiant recombination proportional to the excess minority carrier concentration, multiplied by the majority carrier concentration.

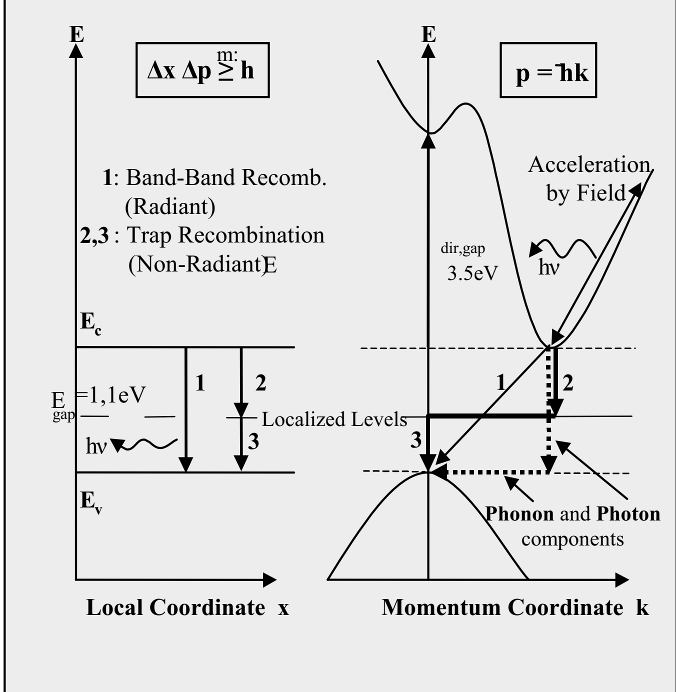


Fig 2: Silicon Band Structure

The radiation wavelength is about 1.1µm (1.1 eV band gap). The intensity I is very faint, even at high injection levels and the spectral distribution is a normal gauss profile ***(Fig 4)***. For small exponents -[$|(E- E_G)| > 3kT_e$] it can be approximated with an exponential function:

$$I = I_0 \cdot \exp -[\, |(E- E_G)| / kT_e \,] \qquad (1)$$

E_G: Bandgap energy
I_0: Intensity at E_G
T_e: electron temp.
E: electron energy

with T_e being the electron temperature, E_G the bandgap energy and I_0 the intensity at E_G. In the recombination case only heat contributes to T_e and the Intensity decreases by orders of magnitude within a wavelength range of 100nm

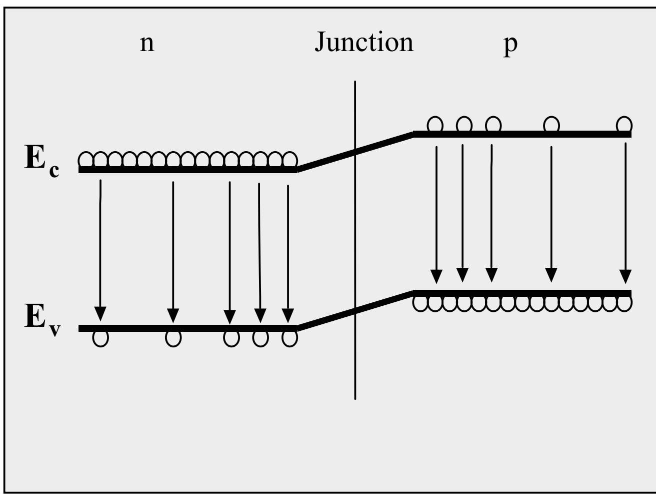


Fig 3: Forward Biased Junction (only radiative recombination shown)

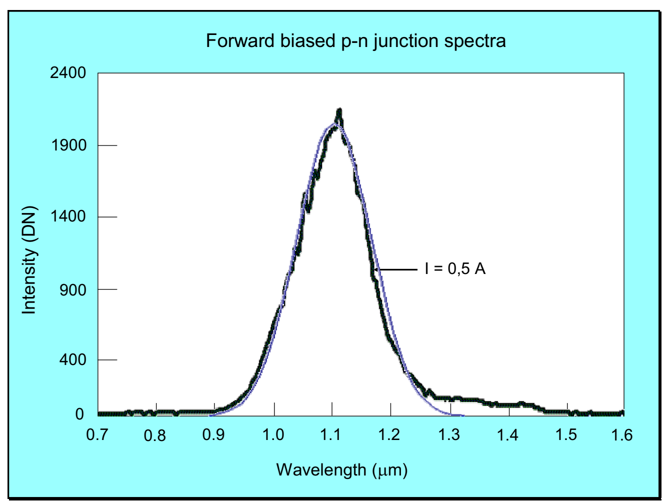


Fig 4: Forward Biased Diode Spectra: Curve obtained from PEM compared to Normal distribution

The recombination activity A decays with the distance x from the space charge layer of the junction by:

$$A = A_0 \cdot \exp(-x/L), \quad \text{with} \quad L = \sqrt{D \cdot \tau} \qquad (2)$$

τ: carrier lifetime
A: Recomb. Activity
A_0: A at junction
L: diffusion length
D: diffusion constant
x: distance from junc.

A_0 is the activity at the junction and L the diffusion length as the square root of the product of the diffusion constant D and the carrier lifetime τ.

A reverse biased junction raises the pn-energy barrier and creates an electrical field in a large space charge region (SCR). The passing charge carriers are accelerated and gain kinetic energy. They relax **scattering** at lattice disorders like phonons, crystal defects and charged coulomb centers, accompanied by light emission with high probability and in a few cases additionally by recombination (***Fig 5***). The population of energy levels follows a Boltzmann distribution with T_e defined by the energy the electrons gain in the electrical

field. The energy loss per scattering event varies and the probability decreases exponentially for increasing energy. Scattering results in a light emission with a broad spectral distribution from near IR into visible region (***Fig 6***). At small energies occurs a thermal kink due to long range coulomb scattering by plasmons in the drain (Fiscetti and Laux, ref. A/02/). The intensity decreases for $E>E_G$ again like **Eq 2.**

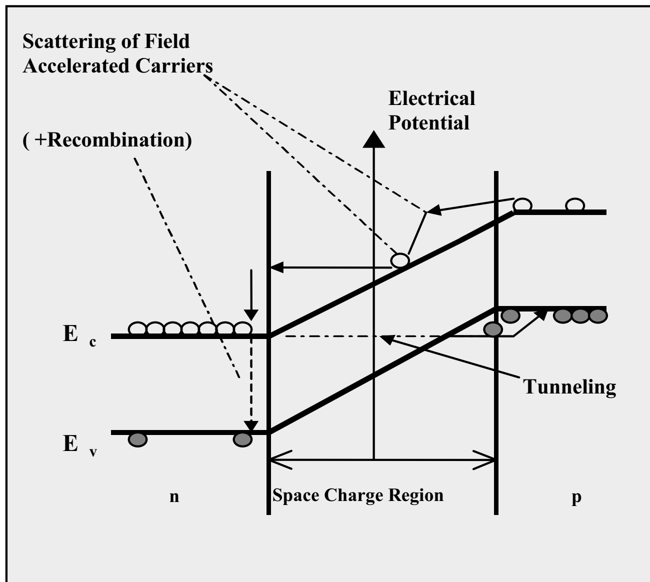


Fig 5: Reverse Biased Junction

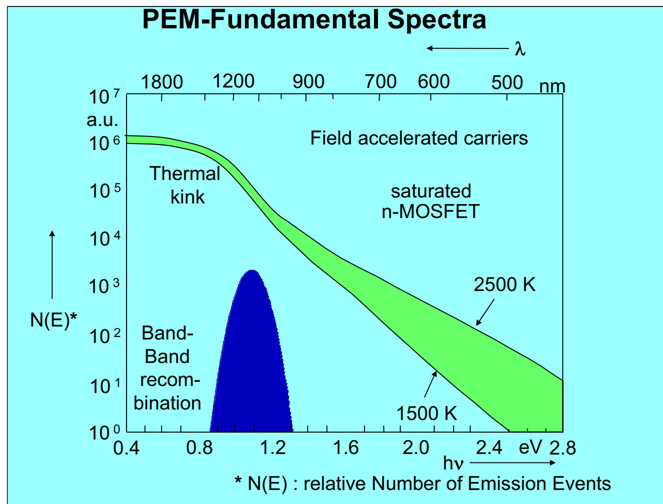


Fig 6: Spectra of Relaxation after Field Accelerated Carriers

These two effects which correlate to the active properties of semiconductor devices are the source for PEM. 'Ohmic' currents in homogeneously distributed density within their wires or designed current paths do not show in PEM unless they are accompanied by one of the photoemission processes. PEM is no thermography. Black Body Radiation in the detected spectral range is only significant at temperatures above 500K (Fig. 12).

3) EQUIPMENT AND APPLICATION

3.1) PEM Setup

A PEM tool (***Fig 7***) is usually based on a wafer prober with the product or structure under investigation being driven via electrical probes / probe cards, or a locally opened IC positioned on the stage. The microscope undergoes a few optimizations as is presented in chapter 4.2. The faint light must be detected very sensitively, usually with a cooled CCD or, especially for dynamic measurements, more sophisticated detectors. For localization, the resulting PEM image needs to be overlaid with a micrograph of the structure which requires some image processing.

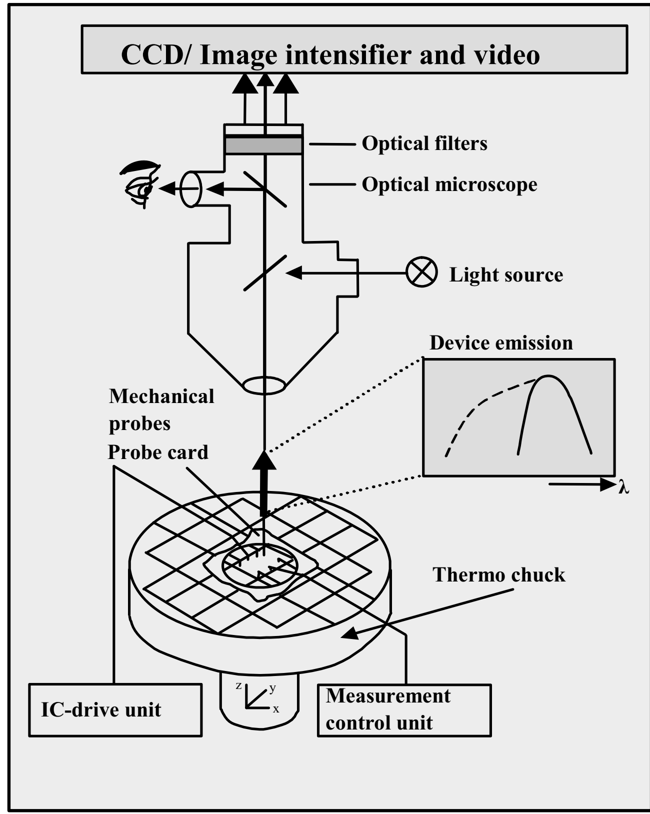


Fig 7: Setup of a Photoemission Microscope

3.2) Optical Path

The microscope, only an adjustment tool on a prober, has a crucial function in PEM. For an image of the light reflected from the IC surface, the lamp is bright enough. But, the faint photoemission signal needs the microscope as optical path to get transmitted to the detector. Any loss of intensity should be avoided. Most important is a high numerical aperture (NA) of the objectives with the high working distance (WD) necessary for manipulation still maintained. The NA is a measure for the light flux F transmitted with a magnification M from a light spot to a CCD pixel:

$$F = f(NA/M)^2 \quad \textbf{(3)}$$

M: magnification, NA: numerical aperture, F: light flux

Magnification	NA	rel. Flux
0.8	0.025	0,13
0.8 Macro	0.40 (eff. 0.14)	321 (39)
5	0.14	1
20	0.42	0,6
100	0.50	0,03

Table 1: Light flux of PEM microscope objectives

Table 1 shows numbers for typical objectives. The highest flux occurs at the higher magnifications. But an efficient application in failure analysis requires the highest sensitivity at lowest magnification: The first PEM image with the whole chip in field of view should show all the light emission spots in rough localization that are detectable. The FA engineer then can decide immediately if it makes sense to go to higher magnifications or try another technique. So, 0.8x magnification is essential for PEM because it allows to image the complete chip with one image shot. Here, optimum sensitivity is required. This can only be assured with a macro objective which is complicated to fit into a microscope optical path. The resulting compromise in handling is more than balanced by the gain in FA cycle time. If spots are detected at 0.8x, it is worth the time that may be required to make the signal visible at higher magnifications with lower light flux conditions.
The relative flux as in (***Table 1***) represents ideal microscope objectives. In reality, some intensity gets lost and the effective NA of macro objectives is significantly lower. The resulting relative signal intensity for each objective may differ from the calculated numbers, even by some extent.
High transmission means removal of all dispensable lenses and mirrors, i.e. zooms, during the recording of the emission image. Absolute darkness should be provided; any noise reduces the sensitivity significantly.

3.3) Detectors

The ideal detector is sensitive in the spectral range from 1500 down to 500 nm wavelength, images the lateral resolution of the microscope, has GHz frequency range with stroboscopic gating option and operates at TV image frequency. Here, only static photoemission is of interest. An example of a detector for dynamic photon emission is presented in the PICA article of this book. In the past Image Intensifiers (Photocathode, Micro Channel Plate, Vidicon tube) have been standard, but today's PEMs are usually equipped with Silicon CCDs, some have CCDs of compound semiconductors.

Cooled Si-CCDs (*Fig 8)* have a spectral range from 500 to ca. 1100nm. Both light emission mechanisms are detectable, but with a limited sensitivity. Typical pixel fields of 1056x1056 allow imaging with microscope resolution. Integration times for emission images are in the order of seconds and prohibit live imaging. Silicon as CCD material and operation under Peltier cooling give this detector excellent reliability and value for the money.

HgCdTe (MCT) CCDs (and some other compound semiconductors) extend their spectral sensitivity further into the IR regime and thus are sensitive in the spectral domain that the light emission is most intense. These detectors play an important role in PEM from chip backside (chapter 4.4) and low voltage supply of ICs (chapter 5.11.4). They have much smaller pixel numbers (typically 256x256), need to operate at liquid nitrogen temperature for noise reduction and are about an order of magnitude more expensive than Si-CCDs.

With the spectral sensitivity up to 2,5 µm wavelength, it is also sensitive to black body radiation. This offers the chance to use the same detector also for thermography. In order to avoid interference of electroluminescence and heat radiation, for PEM applications usually an optical band pass filter is used that cuts off sensitivity at ca. 1.4 µm wavelength.

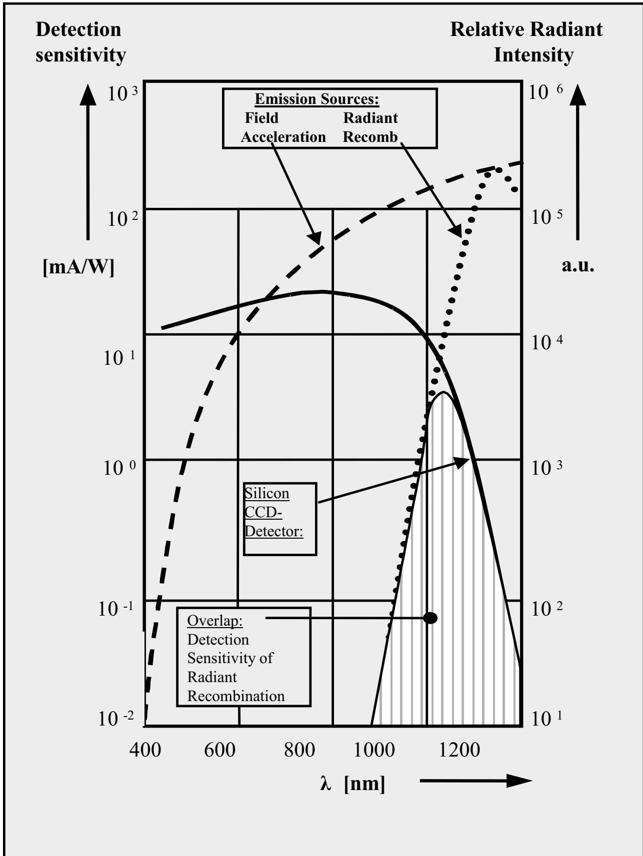


Fig 8: Spectral Distribution of PEM Signals and Sensitivity of Si-CCD detector

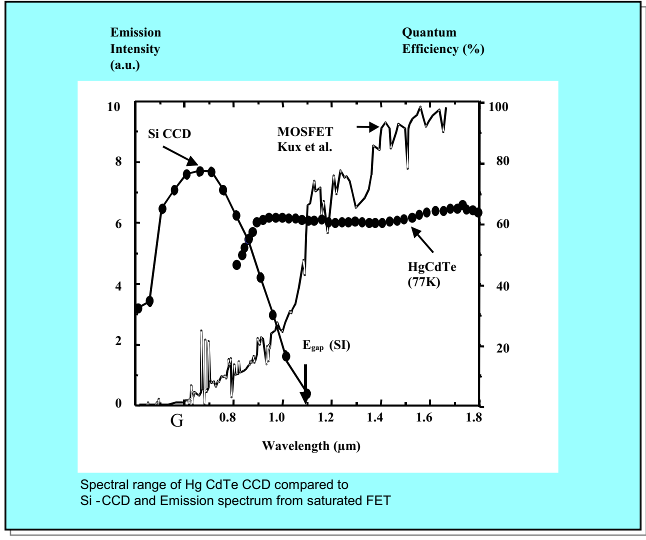


Fig 9: Spectral Range of HgCdTe CCD

3.4) Backside Inspection

Multi level metallization, packaging innovations or flip-chip technology are obstacles for PEM or other localization techniques, even direct probing from the top of the chip. Silicon can be transparent above 1100nm spectral range due to its bandgap. This spectral part of photoemission is accessible also from the back side of the chip which makes PEM one of the natural techniques in the backside approach. However, there are some challenges:

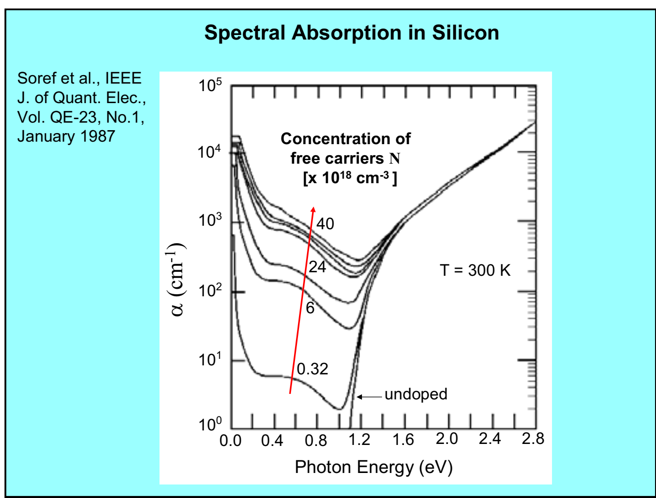


Fig 10: Absorption Coefficent and Free Carriers

1) The Si CCD detector does not fit backside detection, because its function is based on light absorption in Silicon which is the complimentary effect to transparency. Only by temperature the two curves overlap a bit as shown in ***Fig 13.*** This is good enough for the detection of many emission sites, but detectors with a spectral range further into the IR regime would be helpful. The alternative detector in most cases is the MCT that has exactly the spectral range that is required (see ***Fig 9*** in chapter 4.3 detectors).

2) Silicon has a diffraction index of 3.4. Dispersion is a problem then which blurs images because the optical path is in focus only for a small part of the spectrum. Corrected objectives are offered. They need to be adjusted to the right silicon thickness. Generally microscope images through chip backside are not as sharp as the frontside images. Laser Scan Microscopes can be an alternative but are not easy to integrate into PEM (at least with high WD and macro objective). Meticulous polishing is very crucial for the quality of the backside image

3) Heavy doping of the substrate reduces transmission even above 1100nm seriously due to bandgap-narrowing and absorption of free carriers (see ***Fig 10***). In this case, the small spectral effectivity of a Si-CCD for backside investigations is even further reduced (see ***Fig 11***), and the substrate needs to be thinned down to ca. 50µm to assure effective PEM. For every substrate doping level, the maximum bulk silicon thickness that still provides good backside PEM results, should be determined. In MCT cameras this problem is drastically reduced.

3.5) Heat Radiation

Black Body Radiation does not interfere much with PEM as long as the spectral range is not extended into the IR (chapter 4.4). Heat radiation emits in the regularly detected spectral range only at temperatures significantly higher than specified for ICs (***Fig 1***). Rarely, metallization shorts may reach this temperature range for a short time and produce an emission spot, but evidence should be verified in this case. If temperature is of interest, thermography techniques should be pursued.

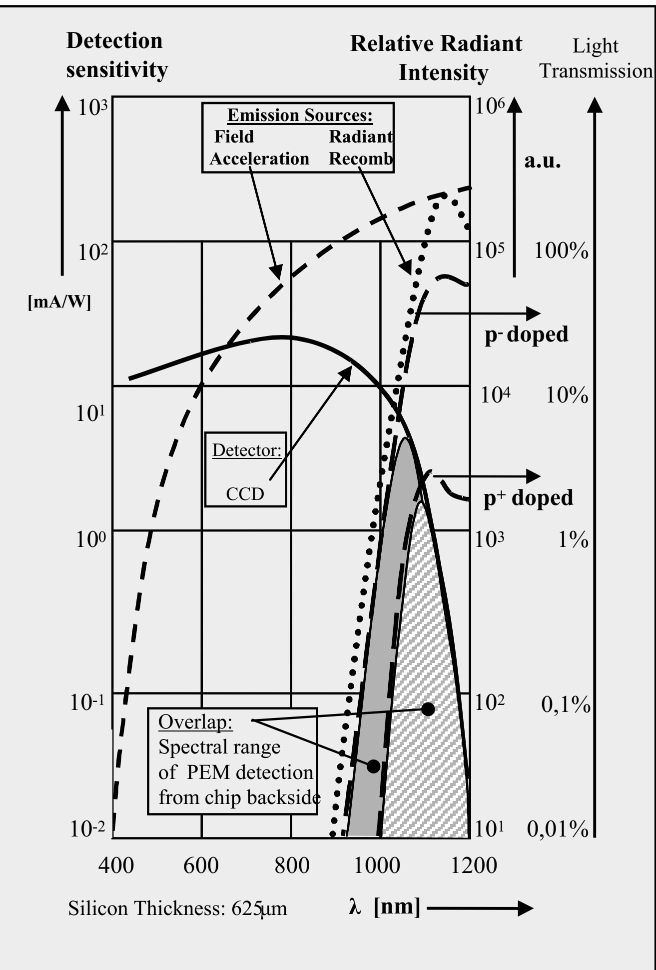


Fig 11: Spectral Response of Si-CCD for PEM through Chip Backside

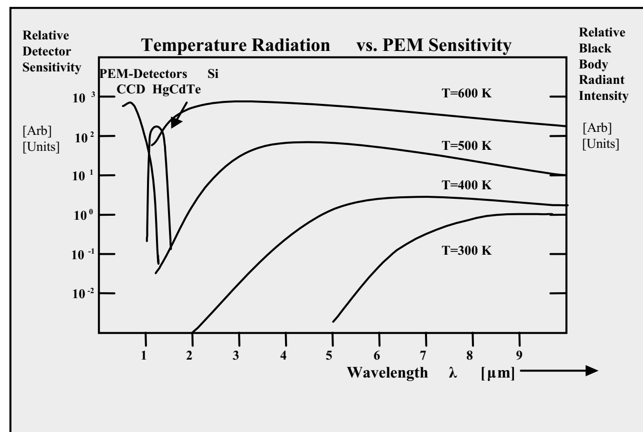


Fig 12: Qualitative Spectral Range Comparison of Black Body and PEM detectors

4) CLASSIFICATION OF EMISSION SOURCES

4.1) Light Emission Generated by Scattering of Field Accelerated Carriers (F-PE)

Most photon emission spots in microelectronic failure analysis are electrical field induced, as described in Table 2. This includes MOS transistors in saturation and most leakage currents in silicon. Even ohmic leakage currents as they occur for example at gate oxide (GOX) defects can be detected if the current density is locally high enough to create F-PE. GOX-defects usually need a local voltage drop of 2V or higher for PEM-detection. Fowler-Nordheim Currents create F-PE due to scattering in silicon or polysilicon, respectively. Effective field variations of the GOX area, resulting from lateral polysilicon resistance or local GOX thinning, are accompanied by FN-current variations that can be recorded by PEM.

4.2) Photoemission Generated by Radiant Electron-Hole Recombination (R-PE)

R-PE is subject to IC failure analysis mainly for latch up investigations. Forward bias effects occur here and there if wells are floating. It is very important for fast bipolar circuits and power devices in their switching properties.

F-PE	Space Charge Region (SCR)	Reverse Biased Junction
		Silicon Leakage Currents
		MOS Transistors / saturated mode
		ESD Protection Breakdown
		Bipolar Transistors / active mode
	Locally High Current	GOX-Defects / -Leakage
	Fowler-Nordheim Current	GOX Leakage Current
R-PE	E-H-Recombination	Forward Biased Junction
		Bipolar Transistors / saturated mode
		Latch up

Table 2: Classification of Photoemission Sources

4.3) Leakage Currents across a pn-Junction in Silicon

A good reverse biased pn junction is only carrying enough current to show up in F-PE when in avalanche mode. A leaky diode (***Fig 1***) is emitting light above detection limit at considerably lower voltages. (exception: well-substrate diodes. If a current is leaking all over their large area, current density may never reach detection limit).
Leakage currents in silicon are the classic PEM application for failure localization. They produce emission spots when passing through or created in an SCR (F-PE). Contact spiking is one more kind of junction leakage. Highly localized junctions or capacitors allow detection even in pA range.

4.4) Saturated MOSFETs/ Hot Electrons

The current through the inversion channel of a MOS transistor in ohmic state is low resistive and usually not accompanied by light emission. Only after pinch off, in saturation, it passes an SCR from pinch-off point to drain (***Fig 43***). The light emission is proportional to the back bias/substrate current (***Fig 54***). This means, the PEM signal is too low to be detected in the logical states “low” or “high” in digital circuits. In static operation mode, absolutely no photon emission should occur in good digital ICs (no forward biased junctions too). Each PEM signal is indicator of a fail: floating gate, leaky gate, leaky junction or avalanche. Only in dynamic operation, photons are emitted as indicators of correct function: whenever FETs switch, they pass the maximum substrate current condition and emit light. Dynamic measurement in picosecond resolution of photon emission can be used to track signal propagation and timing of the circuitry for design verification. The next article in this book about “PICA” by Dave Vallett deals with this approach in detail.

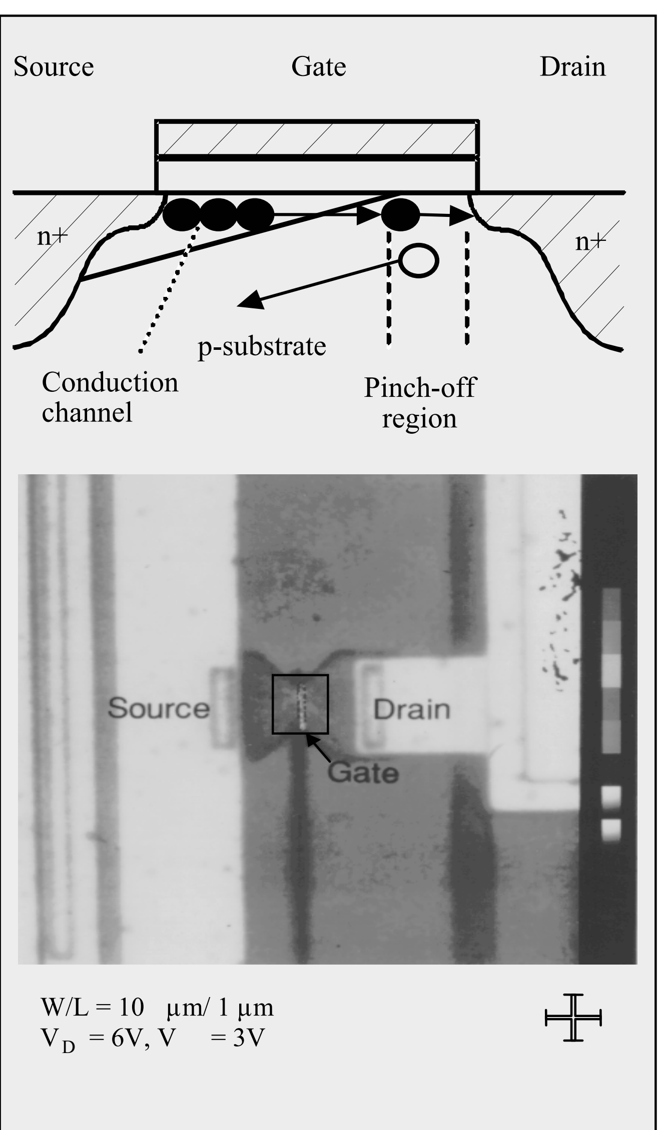


Fig 13: PEM at Saturated FET

All these effects make photon emission an essential effect in IC functionality verification and failure analysis. This may get lost in the future when supply voltage of ICs is dropped below 1V. The PEM signal level is mainly influenced by the strength of the electrical field and the width of the drain space charge region SCR. Scaling down the SCR has mostly been accompanied by increase of the field strength. Now, with reduction of IC supply voltage V_s, this is not the case anymore. The PEM signal drops with 1/ V_s significantly, as ***Fig 65*** illustrates. The different slopes in the graph depend on the spectral sensitivity of the detectors. The PEM signal is not only decreasing with V_s, it is also shifting into the low energy (IR) direction. Detectors that are cutting off the signal on the IR side see an even steeper decrease of the signal with V_s.

The PEM signal is proportional to the carrier multiplication in the SCR. Some of the carriers generated by multiplication in the pinch off region penetrate the gate oxide and become hot carriers. Light emission in an FET can be used as a measure of hot electron damage. Evaluation of the emission spectra allows to determine the electron temperature of hot electrons. (***Fig 22***) shows the spectral distribution to correlate with F-PE and a typical electron temperature of 1500 to 2500K.

4.5) Fowler-Nordheim (FN) Current

Leakage currents of intact gate oxides consist of a tunneling phase and a path in which the carriers get accelerated by the electrical field (***Fig 76***). The emission mechanism is again F- PE, but the spectral

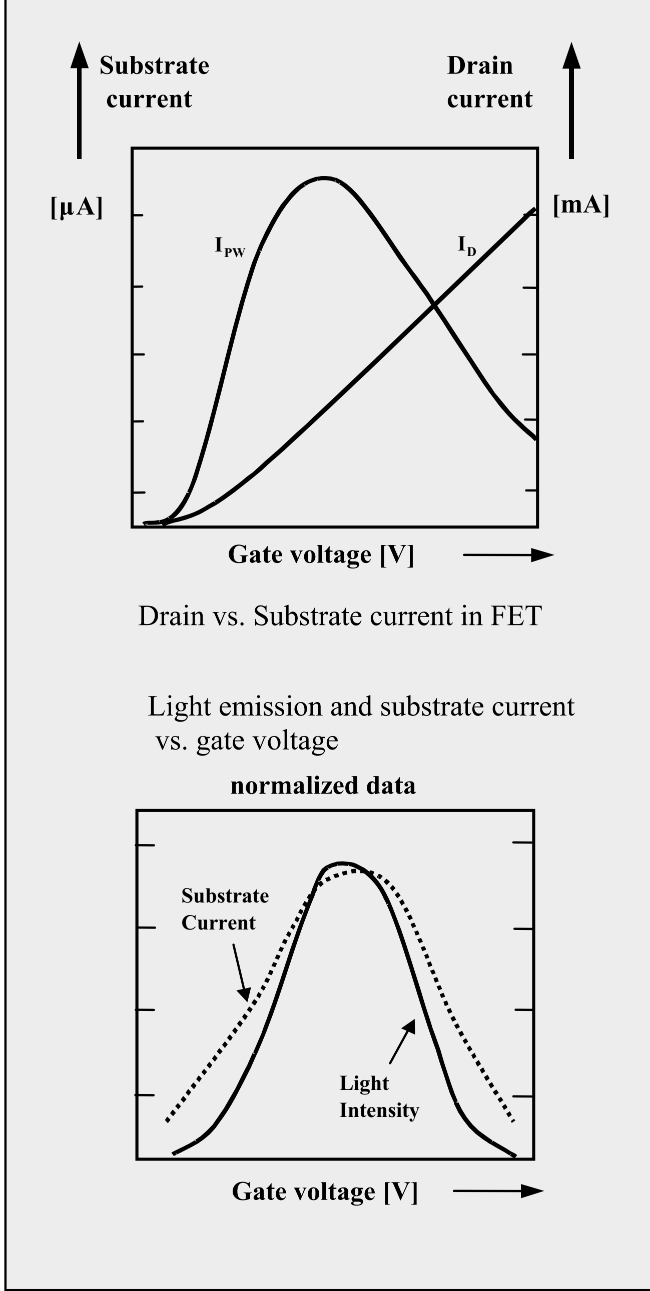


Fig 14: FET Substrate current vs. PEM Signal

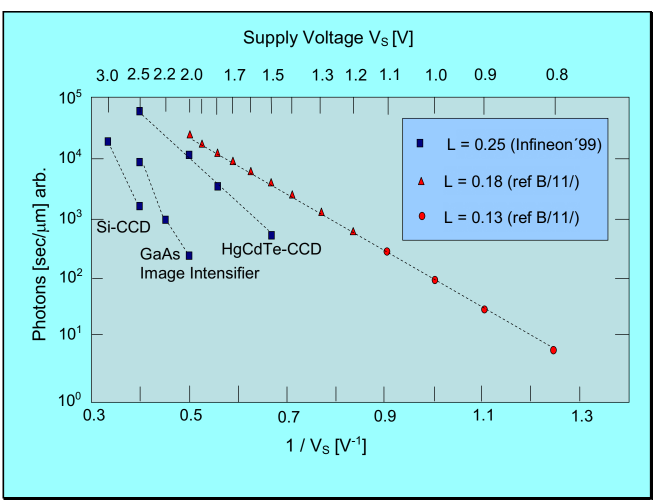


Fig 15: Supply voltage drop and PEM Signal Decrease

distribution (***Fig 22***) is very different from the other sources: The emission intensity is almost equally distributed over the recorded spectral region with a relative maximum at 1.8eV. As discussed before, the slope of the intensity over photon energy expresses the energy distribution of the carriers which undergo the light emitting impacts. In this case, only carriers contribute to the current that pass the Si-SiO_2 barrier, lowered by the tunneling condition. FN-current is in complex proportionality to E-field strength E~ $E^2 * \exp 1/E$.

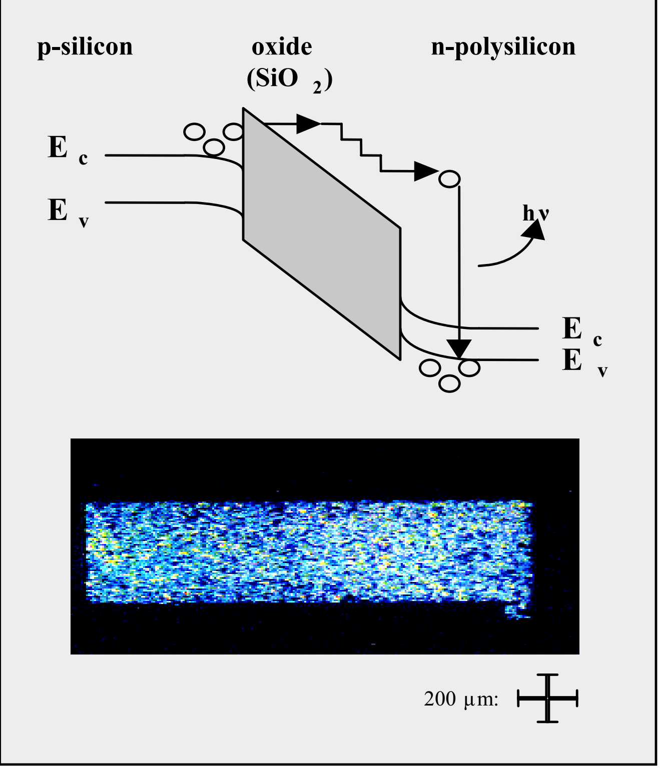


Fig16: PEM of FN-current

As the poly electrode is equipotential, oxide thickness variations create locally changing FN-currents that can be identified by PEM. The PEM signal in the FN-current case is the relative weak and noisy. As gate oxides are very thin, light emission in the blue or UV spectra is very likely which usual PEM is not sensitive for.

4.6) Locally High Currents / Thin Oxide Breakdown

GOX-defects belong to the most important failures which get localized by PEM. But the breakdown of thin oxides does not necessarily produce an SCR especially when poly and well are doped of the same type. The explanation is: the current density is high enough to create a voltage drop at the fail site which in turn implies F-PE in PEM spectral range (typically >2V). This makes GOX-breakdown usually observable in PEM (***Fig 17***). It also explains why not every breakdown can be found in PEM. Some emission spots are unstable or vanish after a while (local high current melts breakdown region, enlarges broken area and reduces current density) and some fails even heal electrically during operation (Joule heat blows off the poly piece and creates an open circuitry around breakdown area).

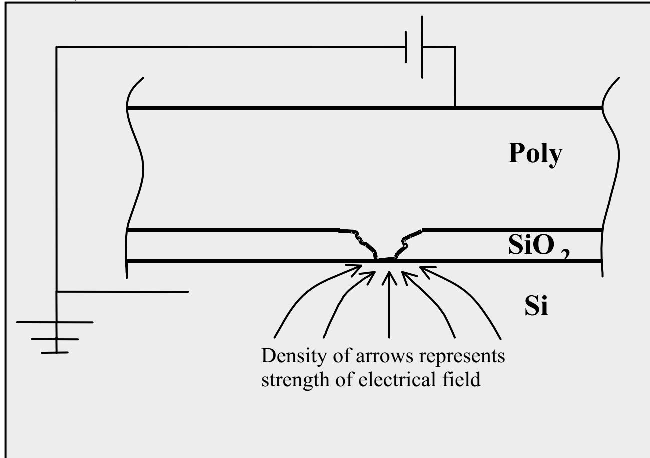


Fig17: Scheme of GOX-Defect

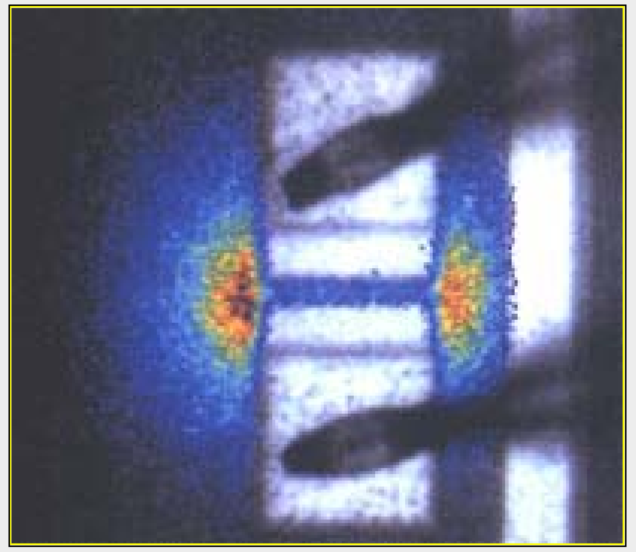
4.7) Bipolar Diode

The principal behavior of a diode in PEM has already been discussed in chapter 2. Here the focus is on the characteristic emission images from diodes in ICs or related test structures. ***Fig 18*** shows a diode test structure with two bond pads above and below and two metal 1 interconnect sheets covering most of the device itself which appears only as dark line in between.

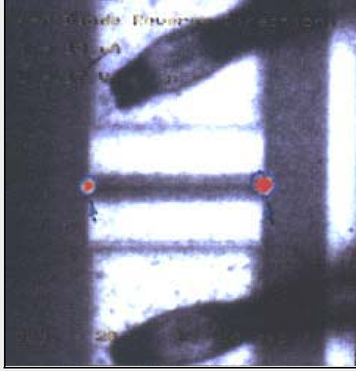
At forward bias, the interconnects are blocking most of the light emission underneath, only on the left and right edges the R-PE emission is reaching the detector. Furthermore the recombination of the carriers diffused away from the diode area is clearly visible. The diffusion length can be easily calculated from this emission image. A typical sign of R-PE is the wide spread of the signal due to carrier diffusion.

In the reverse biased case the image looks completely different: At the edges of the device sharply localized F-PE spots are visible in avalanche condition. This reflects the fact that the highest field strength occurs at corners and edges of he devices. If detectors allow real time mode, these light spots fluctuate strongly in intensity and location. Reverse current light emission is too faint for PEM unless avalanche multiplication raises the current density - and so the scattering probability - by orders of magnitude. Light spots appear at field maxima like corners of the SCR and are sharply localized. Fluctuations of avalanche multiplication are well known and multiplication maxima consist of so many carriers - they produce light flashes even visible with the bare eye.

The typical sign of F-PE is sharp localization at the peak of electrical field.



a) Forward Biased Diode, 100µA pad area 100x100µm^2



b) Reverse Biased Diode, 10µA, to location of image a)

Fig18: a) Forward Biased Diode; b) Reversed Biased Diode

4.8) Bipolar Transistor

For the bipolar transistor operation shown in ***Fig 19*** has been chosen for better observability a MOS transistor driven in its parasitic bipolar mode, with the well under the gate acting as base. This way, the structure is laterally orientated, whereas regular integrated bipolar transistors are vertically structured and the explained effect not that obvious. A bipolar transistor in active mode operates the emitter-basis junction forward biased and the basis-collector junction reverse biased. In saturated mode both junctions are forward biased. The emission images show a very sharply localized signal like F-PE in active, a broader signal like R-PE in saturated mode. The reason is, in active mode the reverse biased base-collector diode produces a much steeper diffusion slope in the base and concentrates the current into the collector direction.

On the other hand, the recombination probability in the thin base is very small. So, a typical recombination signal relies on the absence of SCRs because they separate carriers and keep them from recombination even if there is a high excess carrier concentration. Proof is given in the saturated mode which differs from active mode by the absence of the base-collector SCR. The lateral signal

distribution is characterized rather by carrier diffusion than by field acceleration. From PEM point of view, an advantage of a transistor over a diode is the chance to elevate blocking currents by raising the base current and make them detectable by PEM without all the drawbacks like fluctuations, degradation or destruction as under avalanche condition.

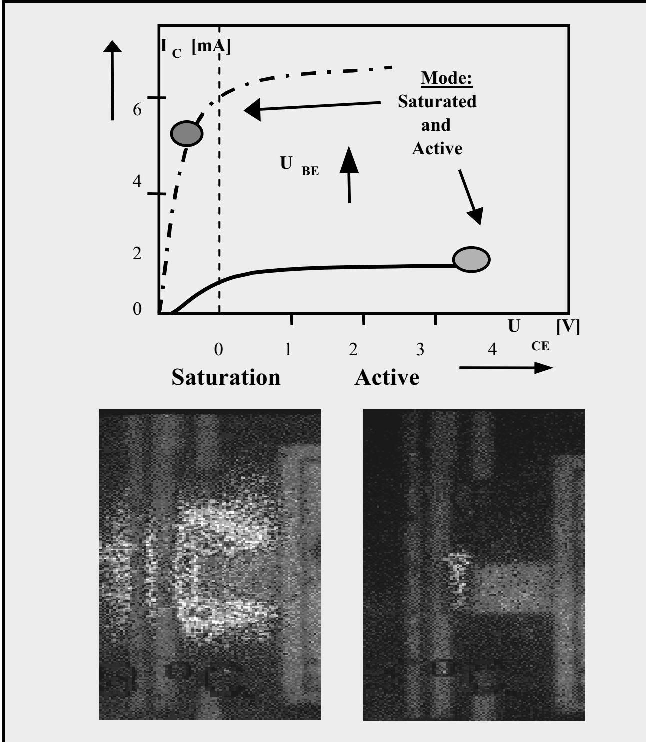


Fig 19: Parasitic bipolar Transistor in active and saturated mode

4.9) Latch up

Latch up is a parasitic bipolar effect of two wells and their diffusions which define a four-layer device like a thyristor. During latch up the thyristor turns on in forward direction (from the viewpoint of the two diffusions) by strong minority carrier injection of the diffusions which floods the wells and the junction of the two wells. Turn on condition is a combined amplification factor >1 of the two folded bipolar parasitic transistors (***Fig 20***). This may be reached by a parasitic base current as in the example of ***Fig 21***, where the minority carrier diffusion from the protection diode serves as additional injection or base current in the latching system and is turning it on. In other cases, a break over condition may be reached by field peaks combined with a tight design and avalanche produces latch up. The emission signal of a stable latch up is always R-PE because of the high excess carrier density of electrons and holes. At the turn on moment in the breakover case F-PE occurs due to avalanche multiplication.

4.10) Emission Spectra

All the emission spectra of the chapters 2 to 3.9 are concentrated in ***Fig22***. The curves of the separate effects have already been discussed previously. In this overview three basic types of spectra directly catches the eye:

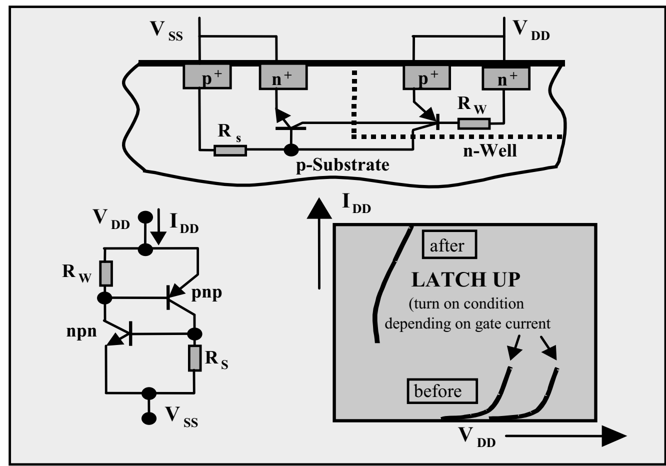
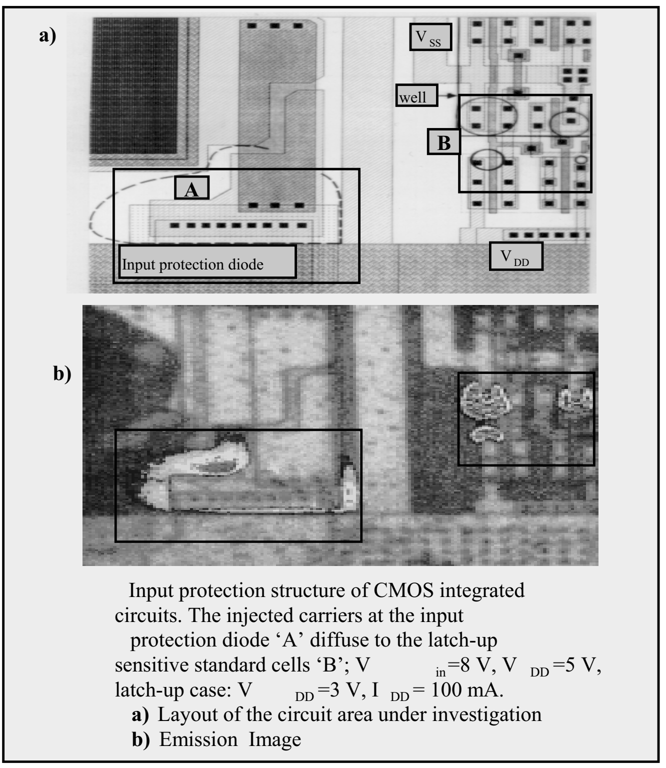


Fig 20: Latch up



Input protection structure of CMOS integrated circuits. The injected carriers at the input protection diode 'A' diffuse to the latch-up sensitive standard cells 'B'; V_{in}=8 V, V_{DD}=5 V, latch-up case: V_{DD}=3 V, I_{DD}= 100 mA.
a) Layout of the circuit area under investigation
b) Emission Image

Fig 21: Input protection structure of CMOS integrated circuits. The injected carriers at the input protection diode 'A' diffuse to the latch-up sensitive standard cells 'B'; V...

1) The recombination radiation (R-PE) with a steep slope due to the **low electron temperature** representing only heat (especially obvious in the latch up case: latch up heats the device significantly - even the slope shows).

2) Radiant scattering of field accelerated carriers (F-PE) with the carriers still approximated to obey Boltzmann statistics, so the exponential slopes represent the kinetic energy with electron temperatures of 1500K and more.

3) Radiant scattering of field accelerated carriers (F-PE) of FN-currents through intact thin oxides with and therefore not showing a

significant shape in the spectra (except for the relative peak at 1.8 eV).
All the quantitative evaluation of photoemission is based on the hypothesis that emission intensity is directly proportional to the carrier density supporting the emitting effect. There are certain limitations to quantitative evaluation of the spectrally resolved signal. The detected light intensity depends on optical transmittance of layers between the emission sources and the optical path. There have been many efforts to file fingerprints of the failure types for automatic recognition - none of these has become routine application because of the ever changing layers the light has to pass like filters - even within one technology.

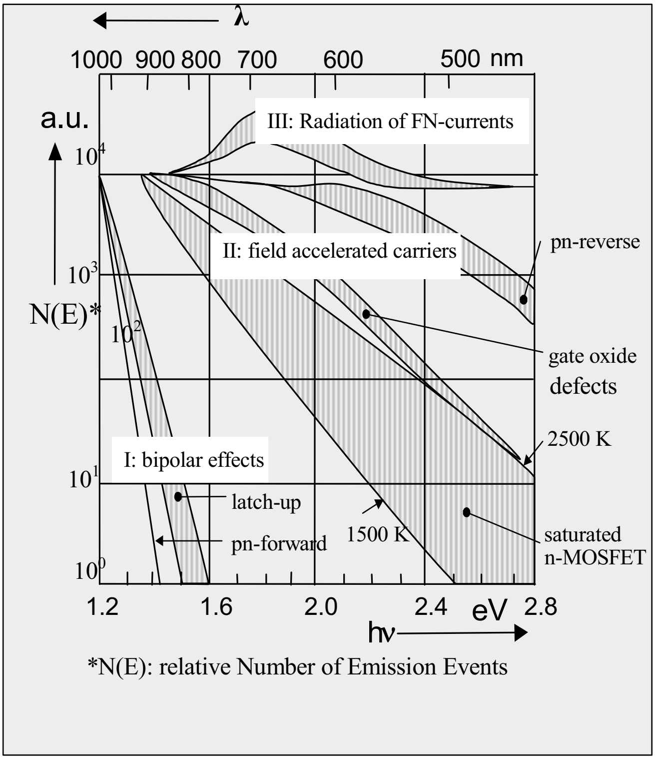


Fig22: Spectra of the Emission Sources

5) SUMMARY

Successful failure localization with PEM requires light emission spots generated in silicon. The key issue is a locally concentrated high current density that creates:

- a local electric field to accelerate a sufficient number of carriers for scattering accompanied by light emission in the PEM-sensitive range (GOX-Breakdown),
- enough carriers in a space charge region to certify a sufficient number of radiant scattering events
  or
- enough carriers to diffuse as minority carriers in the forward bias case to certify a sufficient number of radiant recombination events.

The failure analyst should always try and understand a possible and probable reason in the circuit that can cause the detected emission phenomenon before further preparation because these processes are final.
In general, the emission spot is not always identical to the failure location. A flip flop may not switch correctly, and the result can be a gate of another FET in intermediate state which may emit light, and so on.
The key is to always verify the reason for the light emission in terms of device and circuitry operation.
Often there is more than one reasonable explanation. The analysis job is not done before they all have been investigated.

For Starters:
PEM phenomena should be checked and verified on test structures first, they are more flexible in terms of electrical parameters. This is an opportunity to study thresholds of emission signals, forward bias / reverse bias effects, shapes of emission spots on the screen etc., vary frequencies, temperatures and comparison to alternate techniques.
This is helpful to get a feeling which parameter needs some more or some less stress to receive the desired localization result or how save the device operates when an emission signal can be recorded. With those experiments in mind, it is much easier on the failing chip to work with the limited range of electrical variability and get the highest possible success rate in localizing failures with PEM. There are not many failures in ICs that are inaccessible to PEM.
A powered-down digital CMOS IC (i.e. DC-condition) should not produce PEM-signals - FETs should not switch and have small substrate current, PN-junctions are all in reverse bias. Every emission spot represents a failure as mentioned above - leakage, spiking, GOX-defect, FET with floating gate, current density higher than designed, ESD protection damage, latch up etc. If certain areas show emission accumulations which may even alter each time the circuit is turned on again or even buses emit light, the chip is probably not in a defined state > Recheck if power-down is realized correctly.

6) REFERENCES

6.1) History

Electroluminescence from active silicon devices has been observed in the early 1950's. Back then, reverse bias light emission has been rather an indication of obstructive operation (avalanche mode) than something useful to get information from. As a diagnostic tool it has been utilized in the early 60's for power devices. Forward-bias recombination radiation has been recorded with S1-photomultipliers because it indicates the degree by which the middle layer of pin-diodes or thyristor is flooded by an electron-hole plasma. Laterally resolved measurements could detect the carrier diffusion length that the forward bias resistance of those devices is depending on. Time-resolved measurements allowed exact calculations how, in the switching period from forward to reverse bias, this plasma gets removed from the basis and the SCR starts to penetrate the device - a dangerous moment for a power device because there is still high current to remove the carriers from the base and already the high blocking voltage on the terminals.
These methods have been meticulously developed through the seventies and even the eighties. Power devices offered the amount of light and the geometry to comply with detector sensitivities and the resolution of the optical paths available by them. In the eighties, integration density and the victory of MOS technologies led to light emission observations of some IC failure analysts by looking through the microscopes of their probe stations due to hot electrons and leakage currents. At the same time, night vision technology

offered sensitive image intensifiers. Bringing together these two approaches, today's photoemission microscopy has been presented by N. Khurana and C.L. Chiang 1986. Until 1990, semiconductor industry checked out the range of application and the easiness of operation of the method with a still small number of those machines running. Most of the information on how to understand the phenomena has been collected in that period of time. From then on, with the news spread and the tools optimized, volume application started. Since then, the full potential and range of use has been explored, especially with all the techniques and processes for the approach through chip backside and design debug.

6.2) Acknowledgements

The author wants to acknowledge

- Edward I. Cole of Sandia National Laboratories, Albuquerque NM for contributing the figures on the correlation of FET substrate current and photoemission,
- Jeremy A. Rowlette and Travis M. Eiles of Intel, Santa Clara CA for some data of low voltage PEM,
- Karolin Bernhard-Hoefer, University of Regensburg Germany for her multiple contributions, especially on PEM inspection from chip backside,
- Mrs. Silke Neugebauer and Mr. Adrian Kuehn, Berlin University of Technology, Germany, for word processing and graphics.

6.3) Literature

This is only a selected list of publications about PEM. Many case histories and reports of PEM events are published in Proc. of the Int. Symp. for Test. and Fail. Analy. (ISTFA) from 1989 on.

Some principal aspects are discussed in more detail in: "Quantitative emission microscopy (J.Koelzer, C. Boit, A. Dallmann, G.Deboy, J. Otto, D. Weinmann) J. Appl. Phys. 71 (11), 1 June 1992, R23-R41.

A) Fundamentals

/01/ Deslandes, H.; Lundquist, T. R.;"Limitations to Photon-Emission Microscopy when Applied to "Hot" Devices"; Proc. ESREF 2003, Mic. Rel. 43 (2003); p. 1645-1650

/02/ Fischetti, M.V., and Laux, S.E.; "Long-Range Coulomb Interactions in Small Si Devices. Part 1: Performance and Reliability"; J.Appl. Phys 89, No.2 Jan 2001, p.1205-1231

/03/ Mayer, D.C.; Ferro, R.J.; Leung, D.L.; Dooley, M.A.; Scarpulla, J.R.; "Application of photoemission microscopy and focused ion beam microsurgery to an investigation of latch-up", Proc. ISTFA 1996, p. 47-51, 1996

/04/ Inuzuka, E.; Suzuki, H.; "Emission microscopy in semiconductor failure", Conf. Proc. 10th anniversary, IMTC/94, Advanced Technologies in I&M, 1994 IEEE Instrumentation and Measurement Technology Conference

/05/ Golijanin, D., "Photoemission microscopy – A novel technique for failure analysis of VLSI silicon integrated circuits", X-Ray Optics and Microanalysis 1992, Proc. of the 13th Int. Congress, UMIST, Inst. Of Physics, 1992, p. 465-468

/06/ J.T. May and G. Misakian, "Failure analysis of a turn-on degraded transistor using photoemission microscopy", Proc. ISTFA 1990, pp. 69-71, 1990

/07/ G.F. Shade, "Physical mechanisms for light emission microscopy" Proc. ISTFA 1990, pp. 121-128, 1990

/08/ A. Dallmann and G. Deboy, "Characterization of Trench-Trench Punch-Through Mechanisms by Emission Microscopy," Proc. ESREF 1990, p. 69-76, 1990

/09/ D. Weinmann, C. Boit and J. Koelzer, "Characterization of leakage currents by emission microscopy,", Proc. ESREF 1990, pp. 61-68, 1990

/10/ C. Boit, J. Koelzer, H. Benzinger, A. Dallmann, M. Herzog, and J. Quincke, "Discrimination of parasitic bipolar operating modes in ICs with emission microscopy," Proc. IEEE/IRPS 1990, pp. 81-85, 1990

/11/ Y. Uraoka et al., "Evaluation technique of gate oxide reliability with electrical and optical measurements," in IEEE ICMTS, Vol. 2, No. 1, p. 97, 1989

/12/ Brahme, U.; Li, D.; "Analysis of device gate oxide problems by photoemission microscopy," Materials Letters (1989) vol. 9, No. 1, p. 10-13

/13/ T. Aoki and A. Yoshii, "Analysis of Latch-up induced photoemission," Proc. IEEE/IEDM '89, 1989, pp. 281-284

/14/ S.C. Lim and E.G. Tan, "Detection of Junction Spiking and its Induced Latch-up by Emission Microscopy," IEEE Annual Proc. of the Reliability Physics Symp., pp. 119-125, 1988

/15/ M. Herzog and F. Koch, "Hot-carrier light emission from silicon metaloxide semiconductor devices," Appl. Phys. Lett., Vol. 53, 1988, pp. 2620-2622

/16/ N. Khurana and C. Chiang, "Dynamic imaging of current conduction in dielectric films by emission microscopy," Proc. IEEE IRPS, pp. 72-76, 1987

/17/ A. Toriumi, M. Yosimi, M. Iwase, Y. Akiyama and K. Taniguchi, "A study of photo emission from n-channel MOSFETs," IEEE Trans. on Electron Dev., Vol. ED-34, No. 7, pp. 1501-1508, July 1987

/18/ N. Khurana and C.L. Chiang, "Analysis of Product Hot Electron Problems by Gated Emission Microscopy," Proc. IEEE IRPS, pp. 189, 1986

/19/ C.L. Chiang and N. Khurana, "Imaging and Detection of Current Conduction in Dielectric Films by Emission Microscopy," IEEE International Electron Devices Meeting, pp. 672-675, 1986

/20/ T. Tsuchiya and S. Nakajima, "Emission Mechanism and Bias-Dependent Emission Efficiency of Photons Induced by Drain Avalanche in Si MOSFETs," IEEE Transactions on Electron Devices, Vol. ED-32, No. 2, pp. 405-412, 1985

B) IC-Functionality

/01/ Marks, H.L.; Ton, T.; Block, J.A. ; Kasapi, S.; Shaw, C.;"PC Card Based Optical Probing of Advanced Graphics Processor Using Time Resolved Emission", Proc. ISTFA 2003; pp. 36-39

/02/ Stellari, F.; Song, P.; Weger, A.J.; McManus, M.K.;"Time-Resolved Optical Measurements from 0.13µm CMOS Technology Microprocessor Using a Superconducting Single Photon Detector", Proc. ISTFA 2003, pp. 40-44

/03/ Deplats, R.; Eral, A.; Beaudoin, F.; Perdu, P.; Chion, A.; Shah, K.; Lundquist, T.R. ;"IC Diagnostic with Time Resolved Photon Emission and CAD Auto-Channeling"; Proc. ISTFA 2003, pp. 45-54

/04/ Vickers, J.; Pakdaman, N.; Kasapi, S.;"Prospects of Time Resolved Photon emission as a Debug Tool", Proc. ISTFA 2002, pp. 645-654

/05/ Bodoh, D.; Black, E.; Dickson, K.; Wang, R.; Cheng, T.; Pakdaman, N.; Cotton, D.; Lee, B.;"Defect Localization Using Time-Resolved Photon Emission on SOI Devices that Fail Scan Tests", Proc. ISTFA 2002, pp. 655-662

/06/ Weger, A.J.; McManus, M.K.; Song, P.;"Timing Characterization of Multiple Timing Domains"; Proc. ISTFA 2002, pp.663-666

/07/ Stellari, F.; Song, P.; Tsang, J.C.; McManus, M.K.; Ketchen, M.B.;"Circuit Voltage Probe Based on Time-Integrated Measurements of Optical Emission from Leakage Current",; Proc. ISTFA 2002, pp. 667-673

/08/ Weizman, Y.; Baruch, E.; Zimin, M.;"Pseudo-Dynamic Fault Isolation Technique Using Backside Emission Microscopy – Case Study, Where All Other Methods Fail"; Proc. ISTFA2002; pp. 539-542

/09/ Bockelman, D.R.; Chen, S.; Obradovic, B.;"Infrared Emission-based Static Logic State Imaging on Advanced Silicon Technologies"; Proc. ISTFA 2002; pp. 531-538

/10/ Varner, E.; Young, C.; Ng, H.; Maher, S.; Eiles, T.; Lee, B.; Single Element Time Resolved Emission Probing for Practival Microprocessor Diagnostic Applications."; Proc. ISTFA 2002, pp. 741-746

/11/ Rusu, S.; Seidel, S; Woods, G.; Grannes, D.; Muljono, H.; Rowlette, J.; Petrosky, K.; "Backside Infrared Probing for Static Voltage Drop and Dynamic Timing Measurements"; Proc. IEEE ISSCC 2001; p.276,277,454

/12/ J.A. Kash; J.C. Tsang; " Dynamic Internal Testing of CMOS Circuits Using Hot Luminescence"; IEEE Electron Device Letters, Vol. 18, N0. 7, 1997, pp. 330-335

/13/ Van Doorselaer, K.; Swerts, U.; Van Den Bempt, L.; "Broadening the use of emission microscopy", Proc. ISTFA '95, p. 57-65

/14/ Adams, T.; "Emission microscopes reveal IC defects", Test & Measurement World (1995) vol. 15, no. 2, p. 51-2

/15/ Finotello, A.; Gallesio, A.; Marchisio, L.; Riva, D.; "Emission microscopy: a powerful tool for IC's design validation and failure analysis", Proc. ISTFA 1992, p. 289-293, 1992

/16/ C. Khandekar, M. Hennis, P. Brownell and D. Bethke, "Photoemission detection of vendor integrated circuit failures during system level functional testing," Proc. ISTFA 1990, pp. 73-79, 1990

/17/ C. Hawkins, J. Soden, E. Cole and E. Snyder, "The Use of Light Emission in Failure Analysis of CMOS IC's," Proc. ISTFA 1990, pp. 55-67, 1990

/18/ R. Lemme, M. Gentsch, R. Kutzner, H. Wendt and H. Haudek, "Defect analysis of VLSI dynamic memories," Proc. ISTFA '89, 1989, pp. 9-14

C) ESD / Latch up

/01/ Stellari, F.; Song, P.; McManus, M.K.; Weger, A.J.; Gauthier, R.J.; Chatty, K.V.; Muhammad, M.; Sanda, P.;"Study of Critical Factors Determining Latch up Sensitivity of ICs Using Emission Microscopy" Proc. ISTFA 2003; pp. 19-24

/02/ Weger, A.J.; Voldman, S.H.; Stellari, F.; Song, P.; Sanda, P.; McManus, M.K.;"Transmission Line Pulse Picosecond Imaging Circuit Analysis Methodology for Evaluation of Electrostatic Discharge and Latchup" Proc. IEEE IRPS 2003; pp. 99-104

/03/ Liao, S.; Niou, C.; Chien, W.T.K.; Guo, A., Dong, W.; Huang, C.;"New Observance and Analysis of Various Guard-Ring Structures on Latch up Hardness by Backside Photon Emission Image" Proc. IEEE IRPS 2003, pp. 92-98

/04/ Song, P.; Stellari, F.; Tsang, M.K.; McManus, M.K.; Ketchen, M.B.;"Testing of Low Power CMOS Circuits Using Optical Emission from Leakge Current"; Proc. ESREF 2002, Micr. Rel. 2002, pp. 212-218

/05/ Salome, P.; Leroux, C. Chante, J.P.; Crevel, P.; Reimbold, G,; "Study of a 3D phenomenon during ESD stresses in deep submicron CMOS technologies using photon emission tool", IEEE IRPS Proceedings 1997, p. 325-3, 1997

/06/ C. Boit, J. Koelzer, C. Stein, J. Otto, H. Benzinger and M. Kreitmaier, "Characterization of Field and Diffusion Currents in ICs with Emission Microscopy" Proc. CERT '90 (Component Engineering, Reliability and Test Conference), p. 110-114, 1990

/07/ M. Hannelamm and A. Amerasekera, "Photon emission as a tool for ESD failure locaization and as a technique for studying ESD phenomena," Proc. ESREF 1990, pp. 77-84, 1990

/08/ K.S. Wills, S. Vaughan, Charvaka Duvvury, O. Adams and J. Bartlett, "Photoemission Testing for EOS/ESD Failures in VLSI Devices: Advantages and Limitations", Proc. ISTFA 1989, p. 183, 1989

/09/ J. Quincke, C. Boit, D. Fuehrer, "Electroluminescence measurements with temporal and spatial resolution for CMOS latch-up investigations", Proc. 2nd European Conference on EOBT, Oct. 1989, Duisburg, Germany, in Microelectronic Engineering, Vol. 12, 1-4, May 1990, p. 157-162

/10/ K.S. Wills, C. Duvvury and O. Adams, "Photoemission testing for ESD failures, advantage and limitations," Proc. of Electrical Overstress/Electrostatic Discharge Symp., pp. 53-61, September 1988

/11/ Lim, S.C.; Tan, E.G.; "Detection of junction spiking ant its induced latch-up by emission microscopy", Proc. IRPS 1988, Monterey, pp. 119-125

/12/ N. Khurana, T. Maloney and W. Yeh, "ESD on CHMOS devices, equivalent circuits, physical models and failure mechanisms," Proc. IEEE/IRPS '85, 1985, pp. 212-223

D) PEM Setup

/01/ Nothnagle, P.E.; Zinter, J.R; Ruben, P.L.; "Macro lens for emission microscopy", SPIE-Int. Soc. Opt. Eng: 1995, vol. 2537, p. 13-26

/02/ An Economical Approach to Correlation of Light Emission Microscopes, K.M. Baker, Proc. ISTFA 1991

/03/ N. Khurana, "Second generation emission microscopy and its application," Proc. ISTFA, 1989, pp. 277-283

E) Spectral Analysis

/01/ M. Rasras; I. De Wolf; G. Groeseneken; H.E. Maes; S. Vanhaeverbeke; P. De Pauw; "A Simple Cost Effective, and Very Sensitive Alternative for Photon Emission Spectroscopy"; 23rd ISTFA 1997, p. 153-157

/02/ Tao, J.M; Chim. W.K.; Chan, D.S.H.; Phang, J.C.H.; Liu, Y.Y.; "A high sensitivity photon emission microscope system with continuous wavelength spectroscopic capability [semiconductor device failure analysis]", Proc. IEEE IRPSiability Physics Proceedings,34th Annual, p. 360-5

/03/ Golijanin, D.; ed. Kenway, P.B.; Duke, P.J.; Lorimer, g.W.; Mulvey, T.; Drummond, I.W.; Love, G.; Michette, A.G.; Stedman, M.; "Photoemission microscopy -a novel technique for failure analysis of VLSI silicon integrated circuits", X-Ray Optics and Microanalysis 1992, p. 465-8

/04/ Bruce, V.J.; "Energy resolved emission microscopy" IEEE IRPS, 1993, p. 178-83

/05/ K.S. Wills, D. Depaolis and G. Billus, "Advanced photoemission technique for distinguishing latch-up from logic failures on CMOS devices," Proc. ISTFA, pp 335-341, 1991

/06/ T. Wallinger; "Characterization of Device Structure by Spectral Analysis of Photoemission" , Proc. ISTFA 1991

/07/ K.S. Wills, et. al. "Spectroscopic Photoemission Techniques to Understand Microprocessor Failures," Third International

Symposium of the Physical and Failure Analysis of Integrated Cicuits-Pan Pacific Singapore, 1991
/08/ N.C. Das and B.M Arora, ”Luminescence Spectra of an N-channel Metal-Oxide Semiconductor Field Effect Transistor and Breakdown,” Applied Physics Letters, Vol 56, No. 12, pp. 1152-1153, (19 March 1990)
/09/ K. de Kort and P. Damink, ”The spectroscopic signature of light emitted by integrated circuits” Proc. ESREF, 1990
/10/ N. Tsutsu et al., ”New detection method of hot-carrier degradation using photon spectrum analysis of weak luminescence on CMOS/VLSI,” IEEE ICMTS, Vol 3, No. 1, p. 143, 1990
/11/ H. Ishizuka, M. Tanaka, H. Konishi and H. Ishida, ”Advanced method of failure analysis using photon spectrum of emission microscopy. ” Proc. ISTFA ‘90, 1990, pp. 13-19
/12/ A. Toriumi, M. Yoshimi, M. Iwase and K. Taniguchi, ”Experimental Determination of Hot Carrier Energy Distribution and Minority Carrier Generation Mechanism Due to Hot-Carrier Effects,” IEEE International Electron Devices Meeting, pp. 56-59, 1985
/13/ S.D. Borson, D.J. Di Maria, M.V. Fischetti, F.L. Pesavento, P.M. Solomon and D.W. Dong, ”Direct measurement of the energy distribution of hot electrons in silicon dioxide,” J. Appl. Phys. Vol. 58, 1985
/14/ T.N. Theis, J.R. Kirtley, D.J. Di Maria and D.W. Dong, ”Spectroscopic Studies of Electronic Conduction of SiO_2” in Insulating Films on Semiconductors, J.F. Verweij and D.R. Wolters (ed.), North-Holland, pp 134-140, 1983

F) PEM Inspection from Chip Backside

/01/ Bailon, M.F.; Tarun, A.B.; Nery, M.; Munoz, J.;”Localization of Electrical Failures from the Backside of the Die: Failure Analysis Case Studies Using Infrared Emission Microscope”, Proc. IEEE IPFA 2003, pp. 193-198
/02/ Beaudoin, F.; Desplats, R.; Perdu, P.; Patrie, D.; Haller, G.; Poirier, P.; Jacob, P.J. ; Lewis, D.; “Emission Microscopy and Thermal Laser Stimulation for Backside Failure Localization”; Proc. ISTFA 2001, pp.227-236
/03/ Loh,T. H.; Yee, W. M. and Chew, Y.Y.;”Characterization and Application of Highly Sensitive Infra-Red Emission Microscopy for Microprocessor Backside Failure Analysis”; Proc. IEEE IPFA 1999, pp. 108-113
/04/ S.-S. Lee; J.-S. Seo; N.-S. Cho; S. Daniel; ”Application of Backside Photo and Thermal Emission Microscopy Techniques to Advanced Memory Devices”, Proc. ISTFA 1997, pp. 63-66
/05/ K. Naitoh; T. Ishii; J.I. Mitsuhashi; ”Investigation of Multi-Level Metallization ULSIs by Light Emission from the Back-Side and Frontside of the Chip”, Proc. ISTFA 97, pp.145-151
/06/ Wu, N.M.; Tang, K.; Ling, J.H.; ”Back side emission microscopy for failure analysis”, SPIE-Int. Soc. Opt. Eng: 1996, vol 2874, p. 238-47
/07/ Vallett, D.P.; ”An overview of CMOS VLSI, failure analysis and the importance of test and diagnostics”, Proceedings International Test Conference 1996, p. 930
/08/ Daniel L. Barton et al.; ”Infrared Light Emission from Semiconductor Devices”, Proc. ISTFA 96, pp. 9

G) History

/01/ K. Penner, ”Electroluminescence from silicon devices - a tool for device and material characterization,” Journal de Physique, Coll. c4, Suppl. 9, Tome 49, 1988, pp. 797-800
/02/ C. Boit, ”Quantitative evaluation of High Injection effects in forward biased pin-diodes by means of recombination radiation measurements” Ph.D. Thesis, Technical Univ. Berlin, Germany 1987
/03/ S. Tam and C. Hu, ”Hot Electron-Induced Photon and Photocarrier Generation in Silicon MOSFETs, ” IEEE Transactions on Electron Devices, Vol. ED-31, No. 9, Sept. 1984, pp. 1264-1273
/04/ S. Tam, F. Hsu, P. Ko, C. Hu and R. Mueller, ”Spatially resolved observation of visible-light emission from Si MOSFETs,” IEEE Electron Dev. Letters, Vol. EDL-4, No. 10, pp. 386-388, October 1983
/05/ F. Berz and J.A.G. Slatter, ”Solid State Electronics Vol. 25, No.8, p. 693 (1982)
/06/ H. Schlangenotto, H. Maeder, W. Gerlach, phys. Stat. Sol (a) 21, p. 357, (1974)
/07/ F. Dannhäuser, J. Krausse: Solid State Electronics Vol. 16, (1972), pp. 861-873
/08/ T. Figelski and A. Torun, ”On the Origin of Light Emitted from Reverse Biased p-n Junctions,” Proceedings of the International Conference in the Physics of Semiconductors, 1962, pp. 853
/09/ L.W. Davies and A.R. Storm, Jr., ”Recombination radiation from silicon under strong-field conditions,” Phys. Rev. Vol. 121, 1961, pp. 381-387
/10/ A.G. Chynoweth et al., ”Photon emission from avalanche breakdown in silicon,” Phys. Rev., Vol 102, No. 2, p. 369, 1956
/11/ R. Newmann, Dash, Hall and Burch, ”Visible Light From a Si p-n Junctions,” Phys. Rev., Vol. 100, p. 700, 1955
/12/ W. van Roosbroeck and W. Shockley, ”Photon-radiative Recombination of electrons and holes in germanium,” Phys. Rev., Vol. 94, 1954, pp. 1558-1560
/13/ R.N. Hall , Phys. Rev., Vol. 83, p. 228, 1951 and Vol. 87, p. 387, 1952

H) Compound Semiconductors

/01/ Roesch, W.J.; ”Light emission as an analysis tool for GaAs ICs”, III-Vs Review, 1997, vol. 10, no. 1, p. 24-7
/02/ E. Zanoni, S. Bigliardi, R. Cappelletti, P. Lugli, F. Magistrali, M. Manfredi, A. Paccagnellla, N. Testa and C. Canali,” Light emission in AlGaAs/GaAs HEMT’s and GaAs MESFET’s induced by hot carriers,” IEEE Electron Device Letters, Vol. 11, 1990, pp. 487-489

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Picosecond Imaging Circuit Analysis – PICA

D. Vallett
IBM Systems & Technology Group, Essex Junction, VT USA

Overview

PICA [i] is a photon counting and timing technique that uses naturally occurring emission from switching FETs to produce time-dependent optical waveforms. The ability to temporally resolve emission is what differentiates PICA from the more widespread static or time-integrated photon emission microscopy (PEM). [ii] Time-resolving the emission allows signal propagation through operating circuits to be directly measured. Such information is critical in the debug, failure analysis, and characterization of performance-related issues like clock skew and circuit delay, and manufacturing and reliability problems like AC defects. [iii]

The need for PICA arises because of the dense wiring and associated fill patterns that cover the front side of integrated circuits. These opaque layers prevent the use of traditional non-contact methods like electron beam or UV laser probing. Further, area-array I/O placement found on many complex devices requires a corresponding area-array probe head, which obviates contact probing and frontside non-contact acquisition altogether. This has led to the development of optical techniques for measuring circuit nodes through the backside of the silicon substrate. Since photon emission during transistor switching occurs predominantly at less than 1.1 eV, corresponding to about one micron in wavelength, it is visible through thin highly doped silicon substrates, making PICA an ideal non-invasive probe.

This chapter provides a short history of the technique; briefly discusses hot carrier emission in silicon FETs and pulsed luminescence from CMOS logic gates; describes basic PICA systems and suitable detectors; covers critical operating parameters and variables; shows example images and measurement data; and closes with a brief summary of limitations and future challenges.

Introduction

The first PICA systems were invented and built by Kash and Tsang of IBM's T.J. Watson Research Center in 1996, where additional characterization and modeling was done on the physical mechanisms behind the emission [1-3]. Their instruments used an imaging detector which spatially registered and temporally counted photons over a relatively wide area, and alternatively a silicon avalanche photodiode (APD) to observe single locations on the chip. Early experiments were carried out in collaboration with IBM's Systems & Technology Group (then known as the Microelectronics Division) which used data from simple ring oscillators to assess proof of concept, compare relative merits of imaging and single point detectors, compare PICA with electron beam or voltage contrast waveforms (the best-known method of on-chip timing analysis at the time), and validate PICA with circuit modeling [4].

Use of an imaging detector ('imaging' PICA) was advantageous in that it acquired many thousands of pixels of emissions in a single acquisition. It was used successfully in a number of cases at IBM [5-7] and elsewhere. However, the relatively poor quantum efficiency of the particular detector used, at wavelengths greater than about one micron (where silicon transmits), coupled with long test loops, ultimately drove signal acquisition times to many hours. The development of much more efficient photodiode and superconducting detectors led to the now predominant method of measuring emission from a very limited field of view, typically one transistor at a time, known as 'single point' PICA. The imaging approach will be reviewed briefly while the bulk of the discussion will center on single-point methods.

Hot Carriers in Silicon FETs

Photon emission is a well known effect in silicon devices and its use in characterization and failure analysis is widespread [8-10]. There are a handful of physical mechanisms that give rise to radiative transitions in the classical MOS structure. These can be broadly classified as either *inter*band (occurring *between* the valence or conduction band), or *intra*band (occurring *within* either the valence band or conduction band).

---

[i] Although the acronym PICA implies an imaging technique, it is first and foremost a time-resolved measurement, applicable equally to single-point methods. It is therefore synonymous with the terms 'time resolved photon emission' (TRPE) or 'time-resolved emission' (TRE).

[ii] Phot<u>on</u> emission is often erroneously described as phot<u>o</u> emission, the latter referring to the photoelectric effect in which incident photons create subsequent emission of electrons (as in *photoemission microscopy* or *photo emission electron microscopy* – PEEM, a materials surface analysis technique). While the photomultiplier detectors used in early phot<u>on</u> emission microscopes and PICA systems did happen to make use of the phot<u>o</u> emission principle, the primary purpose of the greater instrument however was measuring phot<u>on</u> emission. Further, the majority of present-day systems use photovoltaic (e.g., Si, HgCdTe, InGaAs, Ge, InP) or superconducting (e.g., NiN) detectors, and are thus more accurately described as phot<u>on</u> emission microscopes (or PEMs).

[iii] AC or delay defects are those that limit device operation at specified operating frequency, or occur only during targeted delay testing. They are typically signals that are slow to rise or fall.

The emission of interest in making PICA measurements is due to the distribution of hot carriers near the drain of a FET, though its exact cause is not agreed upon. [iv] Its existence is experimentally well established however, occurring predominantly in the near infrared (NIR) and, unfortunately, at a level that is quite weak [11-13].

Hot carrier emission arises when carriers (electrons or holes) are accelerated in an external electric field (e.g., voltage across the pinch-off region of a FET channel) that supplies the necessary kinetic energy. To get an idea of the acceleration required to generate emission of sufficient energy so as to be transmissible through silicon, consider an electric field (*E)* where the added kinetic energy is: *eEd;* with *e* being the intrinsic charge on an electron or hole, and *d* being the distance through which the carrier moves between scattering events in the channel. For a carrier starting at zero kinetic energy, the time (*t*) required to move over the distance (*d*) is: $\frac{1}{2} (eE / m^*) t^2$; where m^* is the effective mass of the carrier. In the continued presence of the electric field the carrier will accelerate and gain energy so that at *t* the kinetic energy will be: $\frac{1}{2} (eEt)^2 / m^*$. If however the carrier is scattered by the lattice or other carriers, it will lose most of this kinetic energy and will have to accelerate over again, with scattering times typically a few tenths of a picosecond. So to acquire about 1 eV of kinetic energy in 0.2 ps, an electron in silicon must be subject to a sizeable electric field, about 10^5 V/cm [14]. In such fields, energy is given to the free carriers at a greater rate than which those carriers can in turn transfer it to the surrounding silicon lattice. As such, while the lattice remains at ambient temperature, the accelerated carriers can reach 2000-3000K because of carrier to carrier scattering effects that help thermalize the distribution, hence the term 'hot' carriers.

These hot-carrier distributions and their associated light emission can respond essentially instantaneously to changes in electric field and current that occur on the picosecond time scale characteristic of modern CMOS integrated circuits. The calculated intensity of the emission depends linearly on those switching currents, and thus reflects the electrical state of a CMOS gate, and is an effective means of monitoring time-dependent electrical behavior.

The above discussion has not made any distinction between hot electrons in the conduction band and hot holes in the valence band. The magnitude of the electric charge is the same for both, so they experience the same force in an applied electric field. On the other hand, differences in effective mass and shorter scattering times for holes compared to electrons result in hole mobilities that are typically only half those of electrons. The result is that for the same electric field, holes will gain on average significantly less energy before scattering than will electrons. Therefore the temperature of a hot-hole distribution in a PFET is substantially lower than the electron temperature in a comparable NFET. This lower temperature means that the calculated optical emission from PFETs is typically more than an order of magnitude weaker than that from comparable NFETs [8, 15, 16], though as discussed below, the relatively high efficiency of single point detectors often makes it impossible to distinguish between the two.

Transient Luminescence in CMOS Circuits

It has been shown experimentally in many studies that when operating in saturation, both NFETs and PFETs emit detectable near-infrared radiation [8, 10, 17]. The intensity of this emission is a strong function of both the source-to-drain voltage V_{DS} and the gate-to-source voltage V_{GS}. For classical long-channel devices at fixed V_{DS}, the intensity of the emission shows a peak near $V_{GS} = V_{DS}/2$ (emission peaks from short channel devices tend to shift toward $V_{GS} \approx V_{DS}$). If V_{DS} is allowed to change while the density of the carriers in the channel is kept constant, the intensity of the emission follows exponentially with V_{DS}. The spectrum of the emission also demonstrates exponential (Maxwell-Boltzmann) behavior. These experimental results are consistent with the theoretical treatments of hot-carrier emission in silicon FETs discussed above.

Using a simple unloaded inverter as an example, when the gate or input voltage V_{in} is swept from ground to V_{DD}, there is a range of values where both NFET and PFET are conducting. A peak in short circuit current (I_{SC}) occurs from V_{DD} to ground as $V_{in} = V_{GS} = V_{DD}/2$ for a brief period, as both devices are briefly in saturation. This transition produces maximum optical emission, again, with NFETs brighter than PFETs. Figure 1 shows experimental data from such an inverter as V_{in} is varied from 0 to V_{DD} and V_{out} goes from V_{DD} to 0. For a classical long channel device, a finite I_{SC} exists in the circuit only when the NFET and the PFET are both conducting. Light emission is observed from the NFET during the first half of the transition as V_{out} begins to drop and V_{in} approaches $V_{DD}/2$. Light emission is observed from the PFET during the second half of the transition as V_{out} approaches 0 and the PFET is in saturation. Thus, observation of the emission from both devices can be combined to trace the start and the end of the switching transition.

CMOS circuits are characterized by gate-to-gate delays of 10-50 ps. The rise and fall times of these gates are on the order of twice the gate-to-gate delays. As indicated above, the NFET emission from a simple inverter for a $V_{out} = V_{DD}$ to 0 transition occurs during the initial rise of the input waveform. As a result, the switching-induced light pulse has a width of less than half the rise time of the gate. Therefore, the width of the emission pulse is comparable to the gate-to-gate delay.

[iv] The possible mechanisms are direct transitions between the higher and lower conduction bands; intraband processes in which the requirements for conservation of momentum are satisfied by the presence of defects; or intraband processes in which momentum conservation is satisfied by phonon emission or absorption, or interaction with impurities (i.e. Bremsstrahlung).

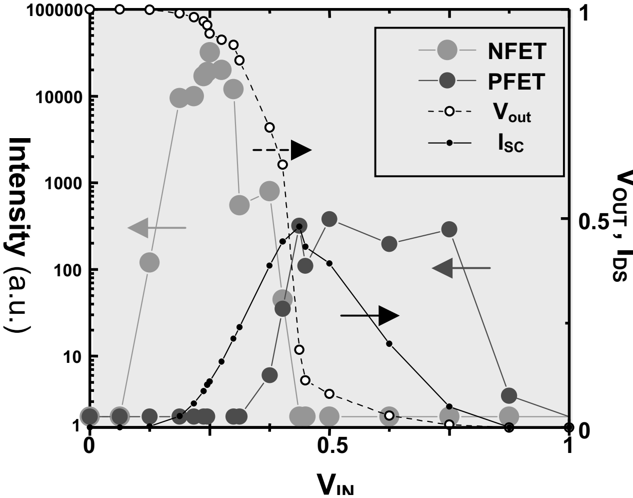


***Figure 1**: Emission intensity, and normalized V_{OUT} and I_{SC} vs. normalized V_{IN}, for an unloaded inverter.*

The above discussion was for an unloaded inverter – i.e., a circuit not driving any logic. A logic gate in a functional chip typically drives a fan-out of one or more circuits. In CMOS devices the load generated by these downstream circuits is due primarily to the interconnecting lines and gates of the driven FETs. Such loads are largely capacitive with a small resistive component. For an inverter driving a capacitive load, the output voltage cannot instantaneously follow changes in the input voltage. For example, V_{in} can reach $V_{DD}/2$ while V_{out} remains temporarily high. Under this situation, shown in Figure 2, the NFET emission can be significantly stronger than for the unloaded inverter because in addition to the short circuit current I_{SC}, current I_{SINK} from the downstream load is also discharged, even after the PFET is turned off. The presence of the load therefore increases the amount of emission from the NFET during a $V_{out} = V_{DD}$ to 0 transition, while the PFET emission remains solely due to I_{SC}. For a V_{out} = 0 to V_{DD} transition, however, the opposite situation occurs. The NFET can cut off while there is still a significant voltage drop across the PFET as the load is still being pulled up to V_{DD}, requiring additional current from the power supply through the PFET (the NFET current in this case being just I_{SC}).

In summary, a loaded CMOS gate will produce strong NFET emission for a $V_{out} = V_{DD}$ to 0 transition, while a V_{out} = 0 to V_{DD} transition will be accompanied by strong PFET emission. For either loaded or unloaded gates, switching transitions generate optical emission from hot carriers. [v] The dominant emission is associated with NFETs effecting $V_{out} = V_{DD}$ to 0 transitions. The presence of current in the NFETs and PFETs is necessary for observable light emission, but does not guarantee it, since the presence of hot carriers in a FET

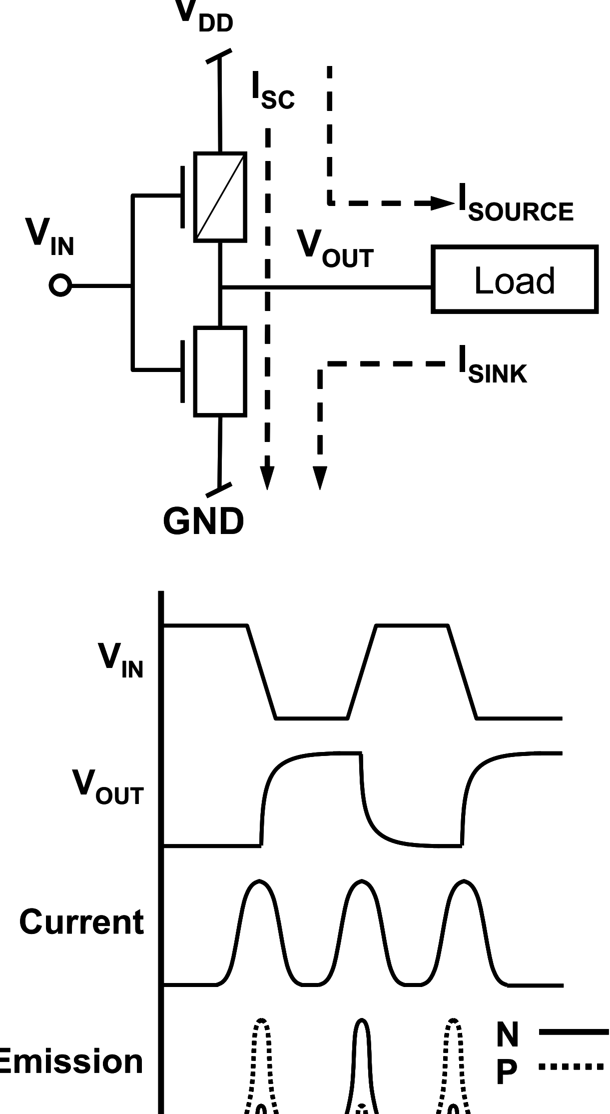


***Figure 2**: a.) Loaded CMOS inverter showing switching-induced current paths; b.) Transient voltage, current, and emission waveforms. (Note: N and P emission intensities are shown normalized - in practice, PFET emission is substantially weaker.)*

depends on the magnitude of the source-to-drain voltage and the relative value of the gate voltage. This latter point is most important as there exists a direct relationship between supply voltage, electron acceleration, energy, and subsequent emission wavelength and intensity, all of which are critical to the efficacy of PICA, as will be discussed below.

PICA Systems and Instrumentation

PICA measurements require six major subsystems – A tester for device stimulus; device under test (DUT) fixturing; optics and positioning for infrared imaging and navigation; DUT cooling; emission signal processing and timing electronics; and the photon emission detector, as shown in Figure 3. The tester supplies a repeating pattern and an event-start synchronization signal so that the system can observe

[v] There are transient emissions that occur in CMOS circuits besides those caused by hot carriers. Sufficiently short channel devices, for example, can produce time-varying FET off-state leakage current which is used to measure logic states or power supply noise, and is readily observed with PICA [18].

switching events and determine their location in time. DUT fixturing and optics are largely common to standard backside photon emission microscopes used for time-integrated imaging. DUT cooling is an increasingly critical requirement for high power dissipation parts as package intervention for backside access often removes heat sinking capability. Emission signal processing and timing electronics are proprietary and unique to particular instrument manufacturers. Hence the remaining discussion will focus on photon detectors, the key component of PICA systems that determines critical parameters of sensitivity, noise, and acquisition time. Figures of merit for photon detectors are: quantum efficiency; jitter; and dark count. [vi]

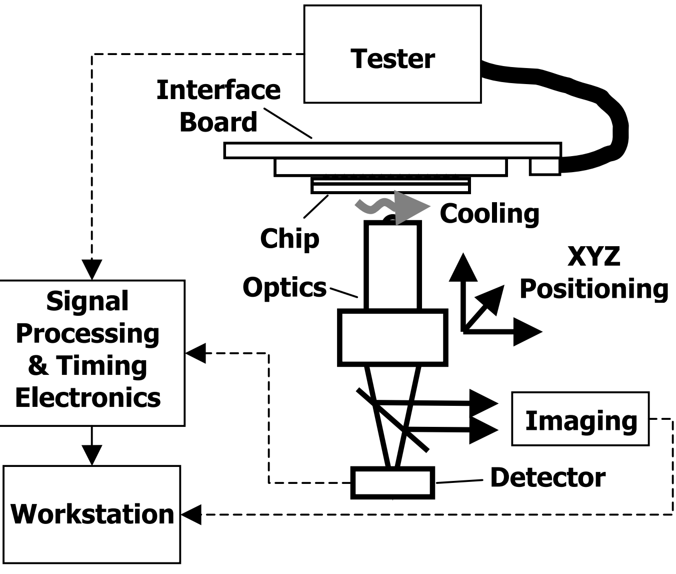


***Figure 3**: Block diagram of major sub-systems comprising a typical PICA measurement setup.*

Imaging Detectors

The original imaging PICA detector was a microchannel plate imaging photomultiplier described by McMullan, et al [19] and Charbonneau, et al [20], using a multi-alkali photocathode. The phosphor screen for visualizing emissions was replaced with a resistive anode back-plane [21], which intercepted the exiting electron cloud generated by the incident photons to provide an analog signal corresponding to their location. Both spatial resolution (i.e., the location of the photons) and, more importantly for PICA, temporal resolution (in the form of a voltage drop or current pulse correlating to the arrival time of the incident photons) were thus provided. This (x,y,t) dataset was then used to create a two-dimensional emission image with a corresponding 'time histogram' for each photon. The important principle central to the technique, and in fact key to the development of PICA, was known as time-correlated photon counting (TCPC), described by Hungerford and Birch (originally developed for measuring fluorescence lifetime decays) [22].

A distinct advantage of using an imaging detector for PICA is that emission from a relatively large area can be spatially imaged *and* time-resolved with picosecond resolution simultaneously (jitter-limited to 90 – 100 ps in practice at the time of development). So a single acquisition over one field of view provided parallel acquisition of timing information from many individual transistors, shown by the example in Figure 4. Additionally, noise is greatly minimized as the dark count is distributed among the many elements (x, y, and t) of the PICA dataset, resulting in a low per-element dark count. The limitation of the imaging approach however is significant – a quantum efficiency of much less than 10^{-3} at 1.3 microns in wavelength – resulting in acquisition times for typical tester loops lasting many hours.

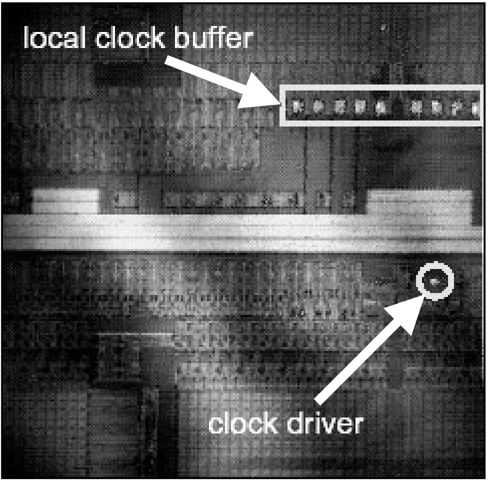


***Figure 4**: PICA data taken with an imaging detector on a 130 nm device at 100X magnification, after 60 minutes of acquisition [30].*

Single Point Detectors

Single point detectors for PICA offer the advantage of quantum efficiencies (at the infrared wavelengths of interest) that are as much as three orders of magnitude greater than that of microchannel plate imaging detectors. Such devices take the form of either an avalanche photodiode such as the one described by Bruce, et al [23], or the superconducting single photon detector (SSPD) developed by Zhang, et al [28].

APDs are widely used in a variety of photon counting applications. In single photon mode they are reverse biased above their breakdown voltage (in what is known as Geiger mode) such that a single carrier can trigger avalanche and result in an easily measurable signal (such *single photon* APDs are also called SPADs). Combination InGaAs / InP devices in particular [24] are well suited to detecting hot carrier emission from FETs, with a quantum efficiency of >50% between 1.1

[vi] Quantum efficiency (also called detection or system efficiency) is a measure of signal strength at the output of an optical path relative to the input, usually shown as a percent or an *arbitrary unit* (a.u.), vs. a range of wavelengths (or energies). Timing jitter, expressed in seconds, measures the variability or uncertainty in time of a transient event. Dark count is simply the accumulation of false or unwanted signal caused by thermal events or statistically invalid photons, typically shown as *counts per second* or *cps*.

and 1.6 microns, and 56 ps of jitter [25]. APDs do have high dark count however which can range from about 10,000 to 20,000 per second or more [25], within the gating window of the detector. Because of the unusual operating mode, avalanche can be triggered by thermal as well as photo-generated events, and by carriers emitted from trapping centers, all of which contribute to the dark count [26]. Fortunately, the very high quantum efficiency of these APDs overcomes the dark count problem, as satisfactory performance is reported (even with much higher dark count than observed above) [27]. Note that single point detectors use a selected-area aperture over the device of interest to block illumination from adjacent devices, and in the case of APDs, may be time-gated to reduce dark noise.

The SSPD offers high detection efficiency in addition to significantly lower dark count and timing jitter [28]. The active portion of the device is a 3-10 nm thick, 100-200 nm wide serpentine of niobium nitride that is cooled to well below its ~10K transition temperature, and current-biased near its critical current of about 10-30 microamperes. An incident photon causes the break-up of weakly bound Cooper pairs such that a nanometer-scale resistive ‘hot spot’ is formed, ultimately resulting in a rapid series resistance change which is detected as a voltage pulse and subsequently amplified. A cooled fiber-optic cable and cooled optics are used to couple photons from the DUT to the SSPD. Background counts are partially reduced by a 1.0-1.5 micron bandpass filter. The SSPD has the potential for detection efficiency on the order of 20% over the spectrum of interest (about 400-2500 nm), 30 counts per second (cps) dark count, and approximately 30 ps system timing jitter [29].

PICA Output and Optimization

Photons acquired by the PICA detector are counted and ‘time-binned’ into histograms, as shown in Figure 5a. Figure 5b shows a processed SSPD PICA waveform and best-fit Gaussian curve, while Figure 5c illustrates data taken after a ten minute acquisition from two different inverters, where the nominal supply voltage was 1.2 V with the chip running at 1.6 GHz, and the tester loop was 64 ns, with a 50 ps low pass filter applied to the data [30]. Based on the formulas after Lo, et al, it is possible to estimate a time resolution better than 20 ps with one minute acquisition, or better than 10 ps if the measurement time is increased to five minutes [31]. Repeatability measurements showed that over many hours there is less than 2.5 ps-rms jitter between waveforms from different acquisitions taken in the same region of the chip. Figure 5b, for example, shows a time resolution of about 50 ps full-width half-max (FWHM) achieved on a minimum size inverter (with the on-chip PLL bypassed to exclude its jitter from the measurement).

Obtaining optimal PICA data depends on manipulation of a number of parameters critical to the quality of the waveform (as defined generally by peak amplitude, background noise amplitude, bin width, and FWHM – see Figure 5a); and the time it takes to acquire it. In turn, the variables contributing to waveform quality and acquisition time are governed by the particular instrument available, the physical and electrical nature of the device being tested, and by its optical accessibility for imaging.

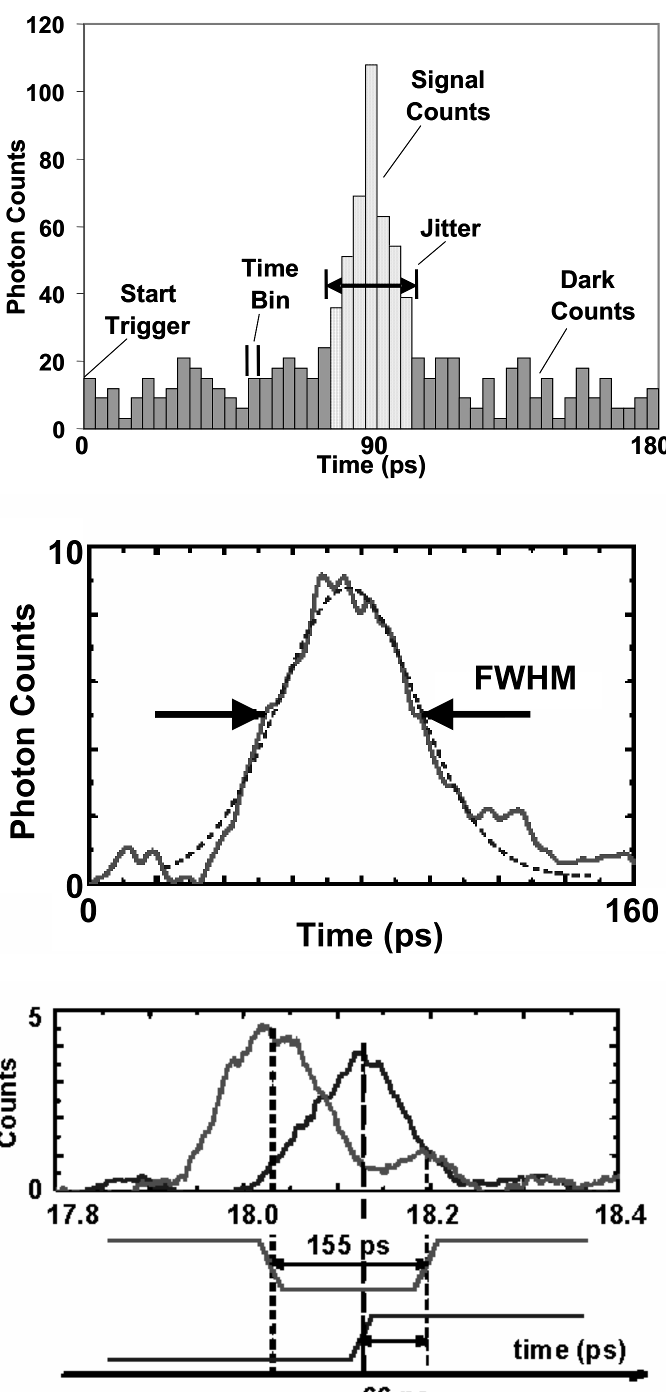


__Figure 5__: a.) Histogram of photon counts vs. time; b.) Processed SSPD waveform and best-fit Gaussian curve (dashed line) for a 130 nm device at 1.2V [30]; c.) SSPD waveforms for two devices in the same technology as in (b) [30].

Instrumental parameters key to waveform quality are system detection efficiency (or optical throughput), dark count, and timing jitter. Detection efficiency is primarily dependent on the solid angle of light collected by the objective lens, the throughput of the optical imaging system, and most importantly, by the quantum efficiency of the detector. Figure 6 shows quantum efficiencies for the detectors reviewed herein versus the emission spectra for hot carriers, where the vastly superior performance of single point detectors versus the S25 imaging photomultiplier, over the bandwidth of interest, is readily apparent. Dark count is thermal in nature, mainly a function of the detector, and is primarily responsible for noise observed in the final waveform. Jitter measures the uncertainty in the arrival time of a photon counted by the detector and output by the system. It contributes to the spread or width of the waveform in time (as does the finite dwell time or width of the native emission pulse at the FET being measured).

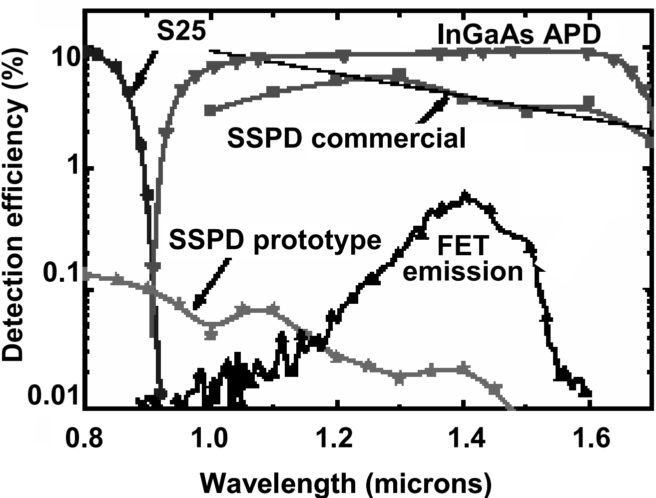


***Figure 6**: Detection efficiency vs. wavelength of PICA detectors, compared with FET emission spectrum [30].*

Device-specific variables are tester loop length, supply voltage, transistor type (NFET or PFET) and design width, and the voltage drop across its channel in-use (e.g. single vs. multiple stack, pass device, etc.). Tester loops can be shortened and supply voltage increased somewhat within allowable operating margins for the circuits under test and the process technology. Transistor type, voltage drop (governed by the number of transistors in a series stack and the fan-out), and width are obviously fixed for a given circuit. Optical accessibility for imaging (for backside analysis) is a function of silicon substrate doping, thickness, and surface reflection, with thinner, lightly-doped substrates having anti-reflection coatings obviously more desirable.

Vickers, et al, have thoroughly examined the combined effect of the parameters above [32] and show that the peak signal count N_{sig} is directly determined by: transistor width (W_{FET}), optical collection solid angle (Ω), acquisition time (T_{acq}), tester loop time (T_{loop}), emission strength (R_{FET}), and detection efficiency (α), with (E) being energy:

$$N_{sig} = W_{FET}\,\Omega\,(T_{acq}/T_{loop}) \int_{E=0}^{\infty} R_{FET}(E)\alpha(E)dE\,.$$

They also signify acquisition time T_{acq} as being a direct function of tester loop length, jitter (Δt_{total}), signal count rate (R_{sig}), and timing resolution (Δt_{res}), (for negligible dark count):

$$T_{acq} = \frac{T_{loop}\,\Delta t_{total}^{2}}{R_{sig}\,\Delta t_{res}^{2}}\,.$$

Note that the probability of a photon being generated for a given switching event is quite small, and is dependent on the spectral density of emitted photons, bandwidth, number of electrons with sufficient energy to emit, and switching duration. Stellari has calculated and measured this value, N_{ph}, equal to approximately 10^{-5} photons per switch (with a 4% collection efficiency) [26]. This very low level of luminescence illustrates the need to thoroughly understand and optimize the many parameters discussed above to attain usable signal strength in reasonable acquisition time.

Future Challenges

Scaling will create a number of challenges for PICA. The most serious is decreased emission intensity with supply voltage. Aside from improvements in sample preparation and instrumental capability, the central question is whether hot-carrier emission will occur with sufficient strength to be detectable with adequate signal to noise and in reasonable acquisition time. As discussed earlier, the strength in turn depends upon the kinetic energy distribution of the carriers near the FET's drain in saturation. While the emission is a linear function of switching current, the dependence of hot carrier kinetic energy upon electric field and hence device supply voltage is exponential.

Rusu, et al, experimentally validated this relationship and also showed emission intensity decreasing about 2X for each technology node from 0.25 micron to 0.18 micron to 0.13 micron (with supply voltages of 1.8 V, 1.5 V, and 1.3 V, respectively) [33]. Tsang and Fischetti have performed Monte Carlo simulations for 65 nm FETs with V_{DS} from 1.5 V down to 0.5 V [34], with Rowlette, et al, showing experimental data for 50 nm devices [35], both validating that the bulk of the kinetic energy distribution remains below 1.1 eV, meaning that qualitatively, NIR emission greater than one micron in wavelength can still be expected, albeit reduced in strength.

But it remains unclear how the drop in intensity will quantitatively affect overall signal detectability and quality. Recall that device-related parameters governing emission strength (e.g., FET width, circuit configuration, supply voltage,

etc.) cannot be controlled by the user. So without the use of specifically designed emission test points or similar dedicated devices, the burden falls to the instrument manufacturer (as Vickers [32] points out), with Lo predicting system detection efficiencies on the order of 20% attainable, in the case of the SSPD [31].

Another challenge to PICA (and backside optical techniques in general) is the diffraction-limited spatial resolution using NIR light – about 600-1000 nm best-case. With the smallest CMOS features already well below this, PICA systems are challenged to locate individual minimum-sized transistors. An emerging method for improving backside image resolution is the solid immersion lens (SIL), also known as a numerical aperture increasing lens (NAIL), which effectively increases the numerical aperture of the objective with a theoretical resolution limit of 0.14 microns calculated and 0.23 microns demonstrated [36]. Technology for optimum SIL placement, optical coupling to the backside surface, translation, and removal is not yet available, and will ultimately determine the tractability of this solution. Use of a SIL will also increase the collection efficiency of emitted light and should mitigate somewhat the effects of reduced emission with supply voltage.

DUT cooling for PICA is also worrisome as power dissipation for large processors approaches 200 W/cm^2 by the end of the decade. Straightforward chilled-air cooling must give way to more elaborate methods employing diamond windows, sprayed fluids, and laminar-flowed IR-transparent liquids [37-39]. It remains to be seen whether these techniques will be viable for heat removal however without significantly affecting performance.

Finally, PICA's temporal resolution versus IC technology is also a critical question looking ahead. More accurately defined as timing precision in this case, it refers to the ability to center a given optical waveform peak accurately in time, versus a known loop start. For example, one might need to differentiate two adjacent waveforms representing the relative arrival times of a clock and a data pulse at a latch. Gate delays for 65 nm devices are now 10-15 ps, expected to reach single picoseconds in five years. Using the astronomical equation for centroiding error - FWHM divided by signal-noise ratio (SNR) - we can conservatively calculate a resolution of 10 to 15 ps, using Lo's prediction of 30 ps FWHM SSPD system jitter and an assumed worst-case SNR of 2 to 3 [31]. Indeed, Varner, et al report measured timing precisions using an InGaAs APD as low as 7 ps already [40]. Whether PICA can continue to achieve single-picosecond timing resolution in the face of decreasing signal counts is likely dependent on optimizing detector efficiency to maximize SNR, and may ultimately require dedicated test points to ensure adequate emission strength.

Summary

PICA has proven to be an essential technique in the characterization, debug, and failure analysis of subtle timing errors on advanced microelectronic products. The critical ability to localize AC defects, solve first-silicon design errors, or validate software models is provided by the technique. While hot carrier emission from FETs and further development of high efficiency NIR photon detectors is expected, significant NIR spatial resolution and DUT cooling challenges remain, as do questions regarding PICA's effectiveness on unconventional devices like dual-gated transistors and FINFETs [41]. Nonetheless, the value of time-resolved emission in improving yield, reliability, time-to-market, and performance should continue to grow, especially if the attractive parallel data acquisition and low noise advantages of an imaging detector can be combined with the high quantum efficiency and low jitter of single point detectors – a solution that could indeed revolutionize PICA.

Acknowledgments

The author thanks Kevin Condon of the IBM Systems and Technology Group in Essex Junction VT, and Peilin Song, Franco Stellari and James Tsang at IBM's T.J. Watson Research Center in Yorktown Heights NY for useful discussion and comment.

References

1. J.C. Tsang and J.A. Kash*, Picosecond hot electron light emission from submicron complementary metal–oxide–semiconductor circuits,* Applied Physics Letters 70, p.889, (1997).
2. J.C. Tsang and J.A. Kash, *Dynamic internal testing of CMOS circuits using hot luminescence*, IEEE Electron Device Letters 18, p.330, (1997).
3. J.A. Kash and J.C. Tsang, *Non-invasive optical method for measuring internal switching and other dynamic parameters of CMOS circuits*, United States Patent US5940545.
4. J.A. Kash, J.C. Tsang, D.R. Knebel, and D.P. Vallett, *Non-Invasive Backside Failure Analysis of Integrated Circuits by Time-Dependent Light Emission: Picosecond Imaging Circuit Analysis*, Proceedings of the 24th International Symposium for Testing and Failure Analysis, ISTFA 1998, pp.483-488.
5. W. Huott, M. McManus, D. Knebel, S. Steen, D. Manzer, P. Sanda, S. Wilson, Y. Chan, A. Pelella, and S. Polonsky, *Attack of the 'Holey Shmoos': a case study of advanced DFD and Picosecond Imaging Circuit Analysis (PICA)*, Proceedings of the International Test Conference (ITC'99) 1999, pp.883-891.
6. P. Song, F. Motika, D. Knebel, R. Rizzolo, M. Kusko, J. Lee, and M. McManus, *Diagnostic techniques for the IBM S/390 600 Mhz G5 microprocessor*, Proceedings of the 1999 International Test Conference (ITC'99), pp.1073-1082.
7. D. Knebel, P. Sanda, M. McManus, J.A. Kash, J. C. Tsang, D. Vallett, L. Huisman, P. Nigh, R. Rizzolo, P. Song, and F. Motika, *Diagnosis and characterization of timing-related defects by time-dependent light*

emission, Proceedings of the IEEE International Test Conference, (ITC'98), p.733-739.

8. C.F. Hawkins, J.M. Soden, E.I. Cole, and E.S. Snyder, *The use of light emission in failure analysis of CMOS ICs*, Proceedings of the 16th International Symposium for Testing and Failure Analysis, ISTFA, 1990, pp.55–67.
9. W.K. Chim, *Semiconductor Device and Failure Anlaysis Using Photon Emission Microscopy*, John Wiley & Sons, LTD, West Sussex, England, 2000.
10. N. Khurana and C.-L. Chiang, *Analysis of Product Hot Electron Problems by Gated Emission Microscopy*, Proceedings of the IEEE International Reliability Physics Symposium, IRPS, 1986.
11. S. Tam and C. Hu, *Hot-electron-induced photon and photocarrier generation in silicon MOSFETs*, IEEE Transactions on Electron Devices, Sept. 1984, vol. ED-31, pp.1264–1273.
12. J. Bude, N. Sano, and A. Yoshii, *Hot-carrier luminescence in Si*, March 15, 1992, Physics Rev. B, vol. 45, pp.5848–5856.
13. S. Villa, A.L. Lacaita, and A. Pacelli, *Photon emission from hot electrons in silicon*, Oct. 15, 1995, Physics Rev. B, vol. 52, pp.10 993–10 999.
14. M.V. Fischetti, S.E. Laux, and E. Crabbe, *Understanding hot-electron transport in silicon devices: Is there a shortcut?*, Journal of Applied Physics, July 15, 1995, vol. 78, pp.1058–1087.
15. L. Selmi, M. Mastrapasqua, D. M. Boulin, J. D. Bude, M. Pavesi, E. Sangiorgi, and M. R. Pinto, *Verification of electron distributions in silicon by means of hot carrier luminescence measurements*, IEEE Transactions on Electron Devices, April 1998, vol. 45, pp.802–808.
16. L. Selmi, *Silicon luminescence techniques for the characterization of hot-carrier and degradation phenomena in MOS devices*, Micro-electronics Engineering, June 1995, vol. 28, pp.249–256.
17. J. Kölzer, C. Boit, A. Dallmann, G. Deboy, J. Otto, and D. Weinmann, *Quantitative emission microscopy*, Journal of Applied Physics Vol 71(11) June 1, 1992, pp.r23-r41.
18. S. Polonsky, A. Weger, and M. McManus, *Picosecond Imaging Circuit Analysis of Leakage Currents in CMOS Circuits*, Proceedings of the 28th International Symposium for Testing and Failure Analysis, ISTFA, 2002, pp.387-390.
19. W. G. McMullan, S. Charbonneau, and M.L.W. Thewalt, *Simulta-neous subnanosecond timing information and 2D spatial information from imaging photomultiplier tubes*, Review of Scientific Instruments, Sept. 1987, vol. 58, pp.1626–1628.
20. S. Charbonneau, L. B. Allard, J. F. Young, G. Dyck, and B. J. Kyle, *Two-dimensional time-resolved imaging with 100-ps resolution using a resistive anode photomultiplier tube*, Review of Scientific Instruments, Nov. 1992, vol. 63, pp.5315–5319.
21. M. Lampton and C.W. Carlson, *Low-distortion resistive anodes for two-dimensional position sensitive MCP systems*, Review of Scientific Instruments, Sept. 1979, vol. 50, pp.1093–1097.
22. G. Hungerford and D.J.S. Birch, *Single-photon timing detectors for fluorescence lifetime spectroscopy*, Measurement Science Technolology, Feb. 1996, vol. 7, pp.121–135.
23. M. Bruce, R.R. Goruganthu, S. McBride, D. Bethke, and J.M. Chin, *Single point PICA probing with an avalanche photo diode*, Proceedings of the 27th International Symposium for Testing and Failure Analysis, ISTFA, 2001, pp.23-29.
24. A. Lacaita, F. Zappa, S. Cova, and P. Lovati, *Single-photon detection beyond 1 µm: performance of commercially available InGaAs/InP detectors*, Applied Optics, June 1996, vol. 35 issue 16, p.2986.
25. J.S. Vickers, R. Ispasoiu, D. Cottoii, J. Frank, B. Lee, and S. Kasapi, *Time-resolved photon counting system based on a geiger-mode InGaAs/InP APD and a solid immersion lens*, IEEE Conference Proceedings - Lasers and Electro-Optics Society Annual Meeting-LEOS, v.2 2003 pp.600-601.
26. F. Stellari, *Non-Invasive Wide-Bandwidth Analysis of Integrated CMOS Circuits With Ultra-Sensitive Photodetectors*, PhD Dissertation, Politechnico di Milano, Milan, Italy, 2001.
27. D. Bodoh, E. Black, K. Dickson, R. Wang, T. Cheng, N. Pakdaman, J. Vickers, D. Cotton, and B. Lee, *Defect Localization Using Time-Resolved Photon Emission on SOI Devices that Fail Scan Tests*, Proceedings of the 28th International Symposium for Testing and Failure Analysis, ISTFA, 2002, pp.655-661.
28. J. Zhang, A. Pearlman, W. Slysz, A. Verevkin, Roman Sobolewski, K. Wilsher, W. Lo, O. Okunev, A. Korneev, P. Kouminov, G. Chulkova, and G. N. Gol'tsman, *A superconducting single-photon detector for CMOS IC probing*, IEEE Conference Proceedings - Lasers and Electro-Optics Society Annual Meeting-LEOS, v.2 2003, pp.602-603.
29. W. Lo, S. Kasapi, and K. Wilsher, *Comparison of Laser and Emission Based Optical Probe Techniques*, Proceedings of the 27th International Symposium for Testing and Failure Analysis, ISTFA, 2001 pp.33-42.
30. F. Stellari, P. Song, A. Weger and M. McManus, *Time-Resolved Optical Measurements from 0.13µm CMOS Technology Microprocessor using a Superconducting Single-Photon Detector,* Proceedings of the 29th International Symposium for Testing and Failure Analysis, ISTFA, 2003 pp.40-44.
31. W. Lo, K. Wilsher, R. Malinsky, N. Boiadjieva, C. Tsao, M. Leibowitz, and H. Deslandes, *Next-Generation Optical Probing Tools for Design Debug of High Speed Integrated Circuits*, Proceedings of the 28th International Symposium for Testing and Failure Analysis, ISTFA, 2002 pp.753-762.
32. J. Vickers, N. Pakdaman, and S. Kasapi, *Prospects of Time-Resolved Photon Emission as a Debug Tool*, Proceedings of the 28th International Symposium for

Testing and Failure Analysis, ISTFA, 2002 pp.645-653.

33. S. Rusu, S. Seidel, G. Woods, D. Grannes, H. Muljono, J. Rowlette, and K. Petrosky, *Backside infrared probing for static voltage drop and dynamic timing measurements*, Solid-State Circuits Conference, ISSCC, 2001, IEEE International, pp.5-7 Feb. 2001.
34. J.C. Tsang and M.V. Fischetti, *Why hot carrier emission based timing probes will work for 50 nm, 1 V CMOS technologies*, Microelectronics Reliability, (UK) Vol.41, No.9-10, Sept. Oct. 2001, pp.1465-1470.
35. J.A. Rowlette, E.B. Varner, S. Seidel, and M. Bailon, *Hot carrier emission from 50 nm n- and p-channel MOSFET devices*, IEEE Conference Proceedings - Lasers and Electro-Optics Society Annual Meeting-LEOS v.2 2003 pp.740-74.
36. S. B. Ippolito, B. B. Goldberg, and M. S. Ünlü, *High Spatial Resolution Subsurface Microscopy*, Applied Physics Letters, 78, p.4071, (2001).
37. S. Ansari, T. Cader, N. Stoddard, B. Tolman, J. Frank, D. Cotton, T. Wong, and N. Pakdaman, *Spray cooling thermal management of a semiconductor chip undergoing probing, diagnostics, and failure analysis*, 2003 International Electronic Packaging Technical Conference and Exhibition, Advances in Electronic Packaging v.1 2003 pp.73-80.
38. T.M. Eiles, D. Hunt, and D. Chi, *Transparent Heat Spreader for Backside Optical Analysis of High Power Microprocessors*, Proceedings of the 26th International Symposium for Testing and Failure Analysis, ISTFA, 2000, pp.547-551.
39. Stevens, K., *IC device under test temperature control fixture*, United States Patent US6191599.
40. E.B. Varner, C.L. Young, H.M. Ng, S.P. Maher, T.M. Eiles, and B. Lee, *Single element time resolved emission probing for practical microprocessor diagnostic applications,* Proceedings of the 28th International Symposium for Testing and Failure Analysis, ISTFA, 2002, pp.P741-6 .
41. H.-S.P. Wong, *Beyond the conventional transistor*, IBM Journal of Research and Development, v.46, nos. 2/3, 2002, pp. 133-168.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Current Imaging using Magnetic Field Sensors

L.A. Knauss, S.I. Woods and A. Orozco
Neocera, Inc., Beltsville, Maryland, USA

Introduction

As process technologies of integrated circuits become more complex and the industry moves toward advanced packaging like flip-chip and stacked die, present tools and techniques are having increasing difficulty in meeting failure analysis needs [1]. With gate sizes approaching 65 nm, “killer defects” may only be a few nanometers in size. In some cases, the defects are nonvisible, i.e. there is no particle that can be imaged by optical microscope or SEM. The increasing number of transistors on a die is also requiring more levels of metal interconnect, which can limit thermal and optical techniques. The more complex devices today have 6 levels of metal, but many companies see 10 to 12 levels in the near future. Further complicating die level analysis are the trends in packaging technology. Flip-chip packaging requires that nondestructive measurements be made through the silicon substrate, and stacked die packaging can require that data is taken through multiple die and packaging materials. The package substrates for these new integrated circuits are also becoming more complex with finer line dimensions approaching 10 µm and many layers of metallization often with several ground and power planes that complicate nondestructive analysis. To meet the needs of failure analysis for some present and most future applications, techniques are needed that are not obstructed by these complications. To some extent this can be accomplished in electrical test through scan architectures once adopted. However, diagnosis of defects using such methods is limited to one logical node or wire, which can often be greater than 200 µm in length and traverse many levels. Further, such diagnostic methods are often non-existent for high current failures and faults in analog devices.

Magnetic current imaging is one such technique that can provide failure analysts with a tool to help overcome some of the hurdles involved in fault isolation of present and next generation semiconductor devices. Through the use of a sensitive magnetic sensor, currents in integrated circuits can be imaged via the magnetic fields they produce. Unlike thermal, optical, ion or electron beam techniques, low frequency magnetic fields are not affected by the materials in an IC or package. Therefore, imaging can be performed from both the front or backside of a device through many layers of metal or packaging materials. These images can reveal the locations of shorts and other current anomalies at both the die and package levels. This technique has applications in fault isolation, design verification, and defective component isolation in full assemblies. A description of this technique, the sensor technology used, and a summary of the various applications of this tool at the die, package, and assembly levels are presented in this chapter.

Physical Principle

Magnetic current imaging uses the magnetic fields produced by currents in electronic devices to obtain images of those currents. This is accomplished though the fundamental physics relationship between magnetic fields and current, the Biot-Savart Law,

$$d\vec{B} = \frac{\mu_0}{4\pi}\frac{Id\vec{\ell}\times\vec{r}}{r^2}, \qquad (1)$$

where B is the magnetic induction, $Id\ell$ is an element of the current, the constant μ_0 is the permeability of free space, and r is the distance between the current and the sensor. As a result, the current can be directly calculated from the magnetic field knowing only the separation between the current and the magnetic field sensor. The details of this mathematical calculation can be found elsewhere [2,3], but what is important to know here is that this is a direct calculation that is not influenced by other materials or effects, and that through the use of Fast Fourier Transforms these calculations can be performed very quickly. A magnetic field image can be converted to a current density image in about 1 or 2 seconds.

Once the current density image is obtained, the information can be used to localize shorts in the packaging, the interconnect, or the die. High resistance defects like cracked traces, delaminated vias and cracked or non-wet bumps can also be localized by looking for small perturbations in the magnetic field between good and bad parts. The current distributions can also be used to verify designs or hunt down I_{DDQ} leakage. In principle, the current density images have the potential to find any type of current related defect or provide any type of current related information.

System Principles

The basic components of a magnetic current imaging system are shown in Figure 1. The magnetic field produced by a sample can be imaged by rastering the sample or the magnetic

sensor in close proximity to one another. If the sample contains a permanent magnetic field, as in many land grid arrays and lead frames, the system will image this constant or "DC" magnetic field. More importantly, current in the device produces a magnetic field. Fields produced by constant currents can be imaged along with permanent magnetic materials, whereas fields produced by alternating currents can be isolated from DC fields and result in a higher signal-to-noise ratio. In AC mode, the signal from the magnetic sensor is sent through a lock-in amplifier, which detects signals with the same frequency and phase as the applied current. This removes the effect of any magnetic materials or magnetic fields produced by constant currents. Interference from ambient time-varying magnetic fields is likewise negated. Thus, the "AC" magnetic image measures only the field produced by the applied alternating current through the part.

In most systems, the magnetic sensor is oriented to detect the z-component of the magnetic field, i.e. the component perpendicular to the scanning plane. To understand the image generated by the instrument, consider the case of a long straight wire carrying a current I (see Figure 2). As the sensor moves over the wire, the z component of the magnetic field will be first negative, then zero, then positive, as seen in Figure 2. A two dimensional image of B_z (the 'z' component of the magnetic field) for a simple wire is shown in Figure 3. The two regions on either side of the wihte line correspond to the negative and positive B_z and the white line corresponds to the location of the current where $B_z = 0$.

To best locate a short in a buried layer, however, the magnetic field image is converted to a current density image. The resulting current map can then be used to determine the fault location either directly or by comparing to a circuit diagram or reference part image.

Sensors

Magnetic sensors have been around for a long time. The earliest sensor was a crude compass that used a lodestone, a magnetic rock, to sense the earth's magnetic field for navigational purposes. It is not known when this was first done, but there are references to this kind of use as early as the 12th century. Since then magnetic sensors have become much more sensitive and sophisticated. Some magnetic sensors of

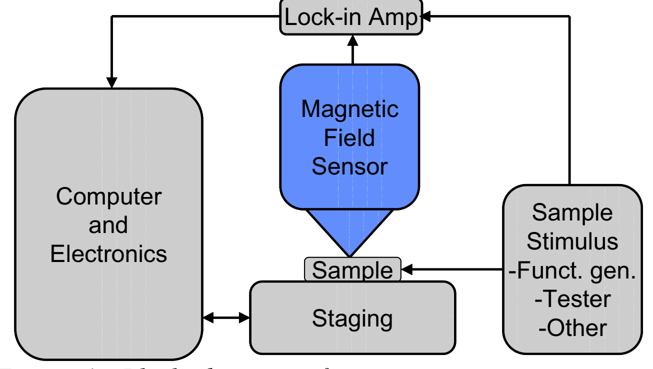


Figure 1: Block diagram of a magnetic current imaging system.

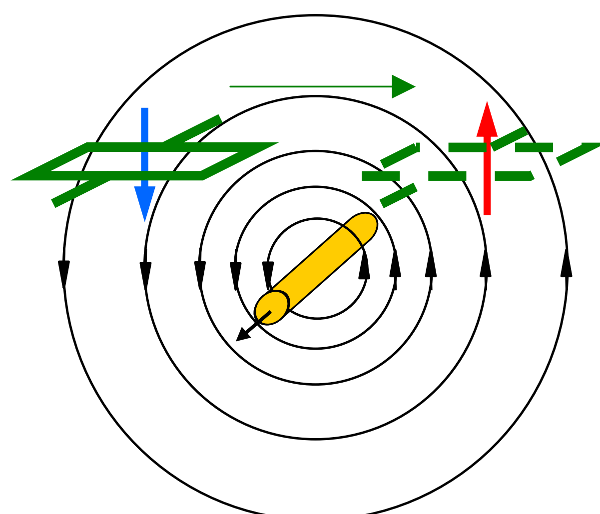


Figure 2: Illustration of a magnetic sensor in the magnetic field of a long straight wire. Wire at center is producing circular magnetic field. Large up and down arrows indicate z component of magnetic field through sensor.

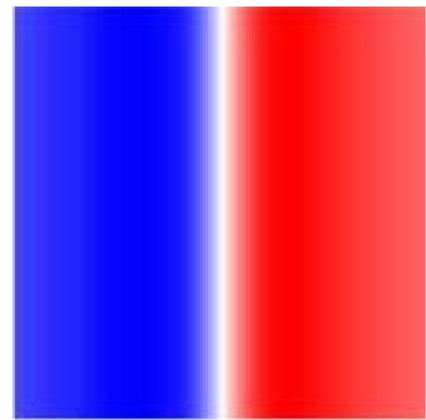


Figure 3: Two-dimensional false color representation of B_z of a long straight wire. Each side of the white line corresponds to oppositely directed flux and white corresponds to $B_z = 0$.

today include Hall effect sensors, flux gates, induction coils, magnetic force microscope tips, magnetoresistive sensors and SQUIDs (superconducting quantum interference devices). A table of the sensitivities of these sensors is shown in Figure 4 [4]. Of these sensors, only SQUIDs and magnetoresistive devices are practical for magnetic current imaging of integrated circuits today. The differences between these sensors correspond simply to sensitivity[1] and resolution[2]. In a typical IC failure, currents are often much less than 1 mA and the sensor is separated from the sample by a distance greater than 150 µm. The ability to effectively image weak currents at a such distances depends on the sensitivity of the magnetic sensor. SQUIDs are the most sensitive magnetic sensors known [5]. They can be designed to measure fields as small as 1 femtotesla (10^{-15} tesla), which is 40 billion times smaller

[1] Sensitivity for these sensors corresponds to the capability to detect weak magnetic fields.
[2] Resolution here corresponds to the ability to spatially separate two current paths in a current density image.

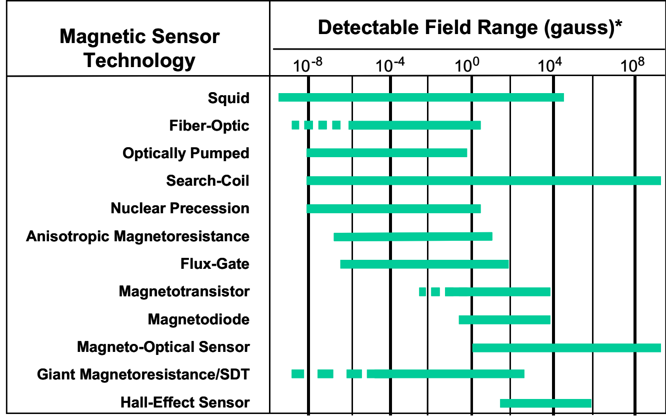


Figure 4: Sensitivities of common magnetic sensors. Solid lines represent demonstrated performance and dashed lines represent anticipated performance [4]. (Last updated by NVE 2004)

than the Earth's magnetic field. SQUIDs for electronic fault isolation are typically designed to have sensitivity around 20 picotesla (10^{-12} tesla). This provides capability to image 500 nA of current at a working distance of 400 µm under typical operating conditions. The best magnetoresistive sensors are about 2-5 orders of magnitude less sensitive, but they are easier to scale down in size to provide resolution less than 300 nm.

SQUIDs

As the name implies, SQUIDs are made from superconducting material. As a result, they need to be cooled to cryogenic temperatures of less than 90 K (liquid nitrogen temperatures) for high temperature SQUIDs and less than 9 K (liquid helium temperatures) for low temperature SQUIDs.

For magnetic current imaging systems, a small (about 30 µm wide) high temperature SQUID is used. This system has been designed to keep a high temperature SQUID, made from $YBa_2Cu_3O_7$, cooled below 80K and in vacuum while the device under test is at room temperature and in air.

A SQUID consists of two Josephson tunnel junctions that are connected together in a superconducting loop (see Figure 5). A Josephson junction is formed by two superconducting

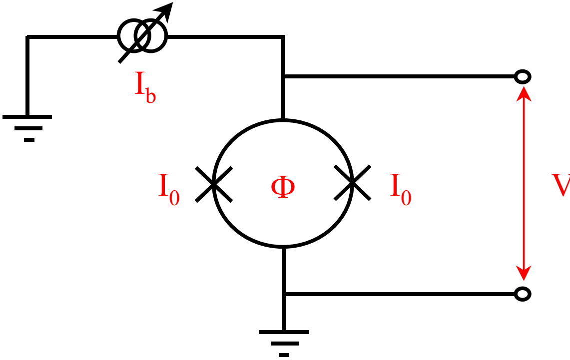


Figure 5: Electrical schematic of a SQUID where I_b is the bias current, I_0 is the critical current of the SQUID, Φ is the flux threading the SQUID and V is the voltage response to that flux.

regions that are separated by a thin insulating barrier. Current exists in the junction without any voltage drop, up to a maximum value, called the critical current, I_o. When the SQUID is biased with a constant current that exceeds the critical current of the junction, then changes in the magnetic flux, Φ, threading the SQUID loop produce changes in the voltage drop across the SQUID (see Figure 5). Figure 6(a) shows the I-V characteristic of a SQUID where ΔV is the modulation depth of the SQUID due to external magnetic fields. The voltage across a SQUID is a nonlinear periodic function of the applied magnetic field, with a periodicity of one flux quantum, Φ_0=2.07x10^{-15} Tm^2 (see Figure 6(b)). In order to convert this nonlinear response to a linear response, a negative feedback circuit is used to apply a feedback flux to the SQUID so as to keep the total flux through the SQUID constant. In such a flux locked loop, the magnitude of this feedback flux is proportional to the external magnetic field applied to the SQUID. Further description of the physics of SQUIDs and SQUID microscopy can be found elsewhere [6-8].

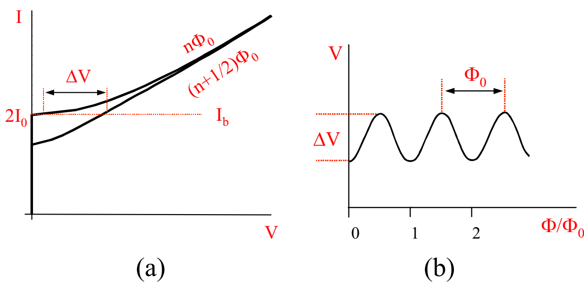


Figure 6: a) Plot of current vs. voltage for a SQUID. Upper and lower curves correspond to $n\Phi_0$ and $(n+1/2)\Phi_0$ respectively. b) Periodic voltage response due to flux through a SQUID. The periodicity is equal to one flux quantum, Φ_0.

Magnetoresistive Sensors

The magnetoresistive microscope depends upon a sensor which is intrinsically less sensitive to magnetic fields than the SQUID, but is more readily miniaturized to the nanoscale. These magnetoresistive sensors are routinely fabricated with dimensions less than 100 nm. These devices are commonly used to sense submicron bits in hard drives, and recent papers have demonstrated magnetoresistive-sensor microscopes with submicron resolution [9,10]. If such a sensor can be brought within about 1 micron of currents to be measured, it has the magnetic sensitivity to map out submicron current lines carrying about 1 mA or more of current.

Magnetoresistive devices are fabricated from materials whose resistance changes significantly in the presence of magnetic field. Before the 1990's these devices were of limited use, depending on anisotropic magnetoresistance (AMR) in thin ferromagnetic films, which only exhibit resistance changes of a few percent. In the last 15 years, magnetoresistive sensors have been revolutionized by the discovery of giant magnetoresistance (GMR) in multilayers composed of ferromagnetic and non-magnetic materials, with resistance

changes of up to 40% in fields as small as 10 Oe[3] at room temperature. The increased magnetic sensitivity of these magnetoresistive multilayers has made them the material of choice for hard drive read heads, enabling the rapid reduction of bit sizes in the last decade.

Giant magnetoresistance multilayers are composed of alternating layers of ferromagnetic and non-magnetic layers. In the simplest design, a non-magnetic layer (only several atomic layers thick), is sandwiched between two ferromagnetic layers, as shown in Figure 7. One ferromagnetic layer is made magnetically hard, often pinned by exchange coupling to an antiferromagnetic layer, and one ferromagnetic layer is magnetically soft and its magnetization direction is free to rotate in small magnetic fields. The resistance of the multilayer depends upon the angle between the magnetization directions of the ferromagnetic layers, with minimum resistance when these directions are parallel and maximum resistance when they are antiparallel, as seen in Figure 8a. To make a device with high sensitivity and a linear, bipolar response, the hard layer is pinned perpendicular to the free layer. Then the response of the device to a magnetic field is approximately linear over the largest range with saturation at fields where the magnetizations of the layers become parallel or antiparallel, as shown in Figure 8b. Applying a constant current through such a device allows a simple measurement of the voltage across the device to yield its resistance change, which is directly proportional to the field at the device's free layer. Noise levels as low as 1 $nT/Hz^{0.5}$ have been reported in magnetoresistive devices [11].

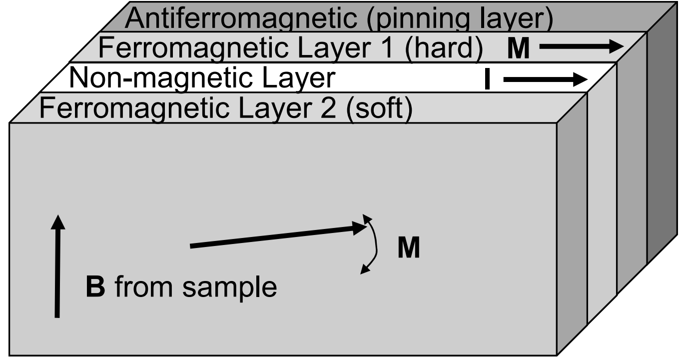


Figure 7: Typical construction of a giant magnetoresistance sensor. The magnetization directions of the magnetic layers are shown with arrows (M) where the soft layer is free to move under the influence of the external field, B. The indicating current (I) is shown on the non-magnetic layer.

(a) (b)

Figure 8: (a) Resistance change of the sensor as a function of angle between the magnetization of the ferromagnetic layers. (b) Resistance change of a linearized sensor as a function of applied magnetic field.

Applications

Magnetic current imaging began as a technique to image power shorts in integrated circuits and packages. While it is still heavily used for this, the capability has expanded to include I_{DD} leakage, in-line test structure leakage, I/O leakage, internal circuit (isolated from any I/Os) leakage, hard logic fails, high resistance defects (sometimes called resistive opens), power distribution mapping, and isolating defective components in assembled devices. In the following examples section, we will show a few of these defect types.

Magnetic current imaging is always used after electrical test as with most other fault isolation techniques. It can be used to isolate any electrical defect involving any anomalous current. This can be measured as a leakage or a resistance change in the device. For a resistance change, it can be either higher or lower than normal. At the time of this writing, magnetic current imaging is only applicable to static defects (i.e. those that can be held in their failing state). It is also possible to use this technique to image current in a good device for the purpose of verifying a design. This review, however, will focus on fault isolation only.

The next step in the process is to decide which sensor type to use. This choice is made based on allowable working distance and how much current the circuit or defect will tolerate. If the device is packaged, the working distance between the sensor and the current is usually > 100 μm. In this situation, the sensitivity of the SQUID will produce the best results and require the least amount of deprocessing (potentially none at all). Starting with the SQUID can also be helpful in preserving a defect that might be lost through any deprocessing. If the device is not packaged or the front side of the chip is accessible, then a magnetoresistive sensor can be used to obtain higher resolution. The higher resolution is accomplished by the smaller sensor size and the closer working distance. It is always important to remember that for a near-field scanning technique like this, the resolution is limited by the sensor size or the working distance, whichever is larger. The sensitivity for magnetoresistive sensors is lower, but since they are used much closer to the source current, images of currents as small as 300 μA have been obtained [10]. This capability is still improving.

In the most general case of a fully assembled device with a short, the SQUID would be the chosen sensor and used initially to isolate the defective component. Once the

[3] 1 Oe (oersted) = 1 G (gauss) = 10^{-4} T (tesla) This relationship is only true for the oersted when the magnetic permeability is equal to μ_0. This is usually true for fields in and around semiconductor devices.

defective component is isolated and the location is determined to be in the die, interconnect, or packaging, the part would be partially deprocessed to enable imaging closer to the defect. The SQUID sensor or a magnetoreistive sensor would then be used to obtain a higher resolution image and more precise localization. At this point there may be enough information to move directly to cross-sectioning or parallel deprocessing. Alternatively, other fault isolation techniques may be used to get complementary information before final physical failure analysis.

After physical failure analysis, there may be situations that require further proof that the observed defect is actually causing the electrical failure. If the defect is still electrically intact, magnetic current imaging can be used on the cross-sectioned or parallel deprocessed part to verify that current is actually present in the defect. This can be very useful in situations where various parties are in dispute over ownership of the defect or when there is some liability involved for the defect. The last example will show this, with a more detailed discussion found elsewhere in this book.

One of the primary advantages of magnetic current imaging over other physical fault isolation methods is that the analysis can begin on a fully packaged/assembled device without any deprocessing, which can minimize any risk of losing the defect and simplify the initial analysis. It is applicable to advanced packaging like stacked die. Also, one gets information that was previously not available, that is, a direct image of current in the device.

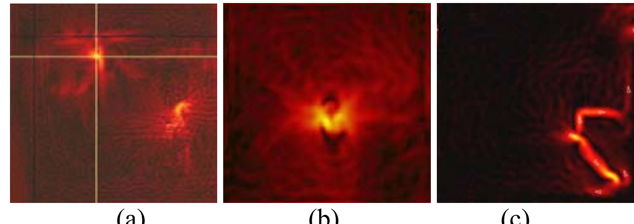
Examples

The following examples are a summary of typical applications of magnetic current imaging. There are many other applications for this technique and more being developed. In principle, the technique can be applied any time that an image of the current in a device would be useful.

Isolating a Defective Component

One of the first challenges in the isolation of a defect is to determine whether it is in the packaging, the die, or the interconnect. Magnetic imaging with a SQUID can be very useful for this due to its high sensitivity and ability to image through packaging materials without deprocessing.

There are three primary ways to determine the defective component. The first involves recognizing the current distribution associated with the die, interconnect or package. Figure 9 shows three examples of shorts in a flip-chip device with the short in a) the die, b) the interconnect, and c) the package. For the short in the die, the current can be seen clearly distributed on the die (even die edges can be seen in current distribution) and concentrated at the point of the short. In the interconnect example, the current can be seen concentrated at the short location, but there is no current distributed throughout the die. Also, there is no current seen in the package because the short in the interconnect is much closer to the SQUID sensor and thus dominates the image at



(a) (b) (c)

Figure 9: Current density images of shorts in a flip-chip device at (a) the die, (b) the interconnect, and (c) the package.

the chosen gain setting. In the package short example, current is seen in the package without any significant intensity variation between the short location and the current in the package [12]. This indicates that the short is in the package. This kind of analysis can be taken a step further and compared to CAD for the package or die to get more precise localization.

The second approach involves calculating the depth of the current in the device from the shape of the magnetic field. This is a little less accurate and depends on the details of the current distribution and the complexities of the magnetic field distribution; however, it is usually possible to calculate the depth of the current in this way with enough accuracy to determine the defective component.

If the second approach is unsuccessful because the current distributions are too complex to accurately calculate the current depth, then a third approach can be used. For this technique a good device is also imaged under the exact same conditions as the bad device. The signal-to-noise ratio can then be calculated for both images. Since the current level is the same in both measurements, the signal-to-noise ratio can be used to determine which current is closer to the SQUID sensor. Depth changes of a few hundred microns will make noticeable changes in the signal-to-noise. This can then be used to estimate if one current distribution is in the package or die.

Isolating Low Resistance Shorts

While magnetic current imaging can be applicable to all types of shorts, low resistance shorts can be difficult or impossible to isolate by many techniques due to the low power that they dissipate. For magnetic current imaging, only current is needed. The actual power that is generated by a defect is irrelevant, making this technique the only option in some cases for low resistance shorts. To demonstrate this capability, two results have been included from work published elsewhere in more detail [13]. The first example is a SRAM standby current failure and the second is an ASIC power short.

The 4 MB SRAM is 0.4 μm device with 4 levels of metal. It was flip-chip mounted on a ceramic ball grid array and thinned to approximately 70 μm of silicon. The part was biased with 130 $mVAC_{pp}$ and drew 100 μA of current (DC average power P_{eff} = 6.4 μW). Figure 10 shows the current image obtained through the backside overlaid on the CAD layout. This matches the defect location and agrees with the

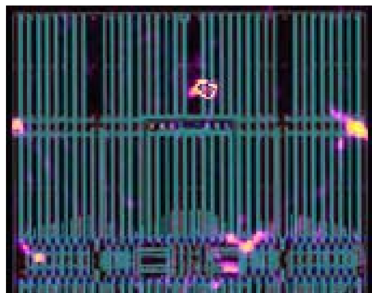


Figure 10: Backside current image of a standby current failure on a SRAM. The defect is indicated by the arrow and the whole image is overlaid on a CAD layout.

results from other techniques like Thermaly Induced Voltage Alteration (TIVA), liquid crystal, and Photon Emission Microscopy (PEM), but acquired with more than 2 orders of magnitude less power.

The ASIC is a 0.1 µm product with 6 levels of metal, which had a power supply short. The device was flip-chip mounted on a CBGA substrate. It was biased with 250 $mVAC_{pp}$ and drew 6.5 mA of current (DC average power P_{eff} = 800 µW). Figure 11(a) shows the current image overlaid on a backside optical image with a close-up of this region and current direction shown in Figure 11(b). The four lobes in the current image indicate current coming together on one layer and shorting in the middle to another layer where the current separates into opposite directions. This constitutes a plane-to-plane short at the center of the region where the current seems to have disappeared, indicated by the cross in Figure 11(b). The current is missing here because it is moving vertically between the layers, which is a magnetically silent direction for the SQUID. The location agrees with the data obtained with a Schleiren Thermal Mapper, which required about two orders of magnitude more current, also shown at the center in Figure 11(b). The resulting deprocessing at this location shows considerable shorting damage, as shown in Figure 12.

Isolating High Resistance Defects (Resistive Opens)

Typically, a high resistance defect is the result of a geometrical change in some circuit element such as a delamination process, crack, void, etc. Previously, the main approach for localizing these defects has been time domain reflectometry (TDR). TDR sends a short electrical pulse into the device and monitors the time to receive reflections. These reflections can correspond to shorts, opens, bends in a wire, normal interfaces between devices, or high resistance defects. Ultimately anything that produces an electrical impedance change will produce a TDR response. These signals are compared to a good part and require time consuming layer-by-layer deprocessing and comparison to a standard part. When complete, the localization is typically at best to within 200 microns. Clearly, the current distribution will be affected by such geometric alterations and correspondingly affect the magnetic field distribution as sketched in Figure 13 for a failing flip-chip bump.

In this situation, one expects to see a small change in the magnetic field distribution around the defect as compared with

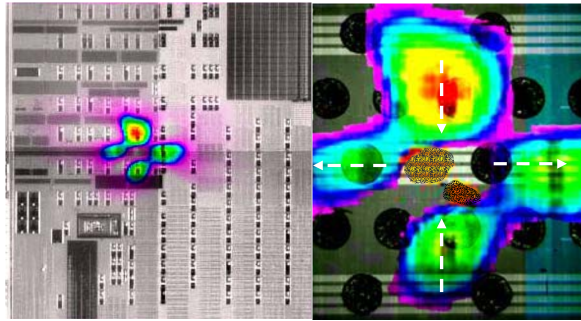


Figure 11: Backside current image of ASIC power short overlaid on backside optical image. (b) Close-up of defect location. White arrows show current converging from top and bottom and diverging from left and right with plane-to-plane short at the center of this current cross.

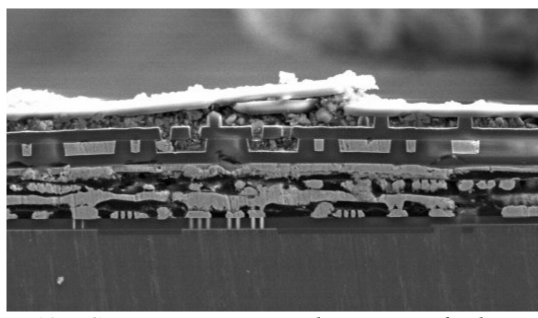


Figure 12: SEM cross-sectional image of shorting site identified in Figure 11.

that in a good part. The localization of high resistance defects through current imaging is accomplished through a detailed comparison of good and failing parts. The differences in magnetic field explained above are small and therefore require a very careful analysis between the good and failing parts. This requires improvements over conventional technology in two areas. First, the instrumentation for current imaging requires more precise automated control of the sample setup and data acquisition. The scan conditions must be as similar as possible between the good and failing parts, so that an effective comparison can be made. Secondly, even with careful sample setup and data acquisition, there will still be misalignments between the two images, and potential signal differences due to different working distances, or even part

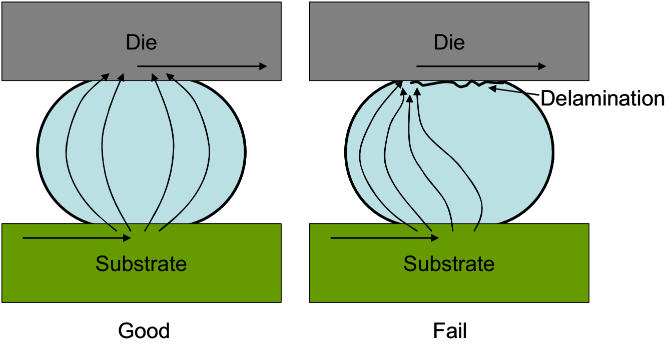


Figure 13: Illustration of current distribution in a good and a delaminated (failing) flip-chip bump.

deformations (e.g. warping). These differences need to be sorted out from those due to the high resistance defect. For this, advances have been made in image difference analysis (IDA) to assist in the identification of failing defects.

The following example is a flip-chip bump failure, which can be found elsewhere along with other examples [14]. TDR was used to determine that the failure was likely in the interconnect since the signal in the failing part did not get far beyond the substrate, as shown in Figure 14. This failure was between power and ground for which there were several bumps. TDR did not definitively indicate that the problem was in the bumps nor did it identify which bump had the problem. Magnetic current imaging was used to further isolate the problem. Figure 15 shows the region of interest from the IDA image overlaid on the CAD for the particular structure involved.

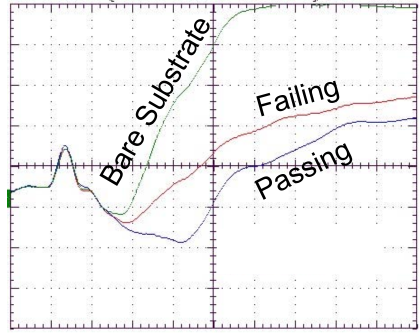


Figure 14: TDR results indicate a resistive open that is close to, but beyond the package substrate.

The metal trace connecting to the failing bump is marked with the dashed yellow line. The flip-chip bump is connected at the end of it before the green grid. The centroid for the magnetic anomaly is the black dot, aligning very well with the position of the bump.

The IDA plot shown in Figure 15(a) does not present adequately the relative intensity of the magnetic anomaly. A better way of doing that is by using a 3D representation as in Figure 15(b), where the z axis is the magnetic field intensity corresponding to the IDA results in Figure 15(a). We plot the absolute value of the magnetic field zoomed-in around the anomaly location for the sake of clarity where the two peaks associated with the defect are clearly visible.

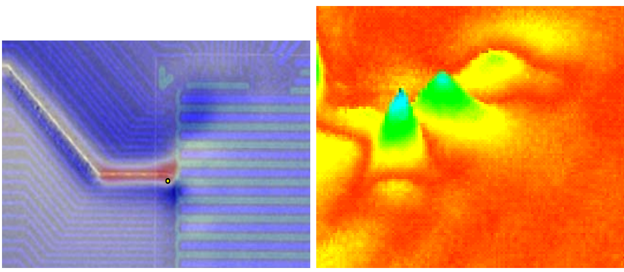


Figure 15: (a) IDA results corresponding to the HR flip-chip bump damage overlaid on the CAD. (b) Zoomed-in 3D version of the IDA results presented in (a).

The defect was isolated with an accuracy of 30 µm and verified by FA as shown in Figure 16. IDA on magnetic field images has successfully localized high resistance defects to this same level of accuracy on a variety of defects, including cracked traces and delaminated vias. This 30 µm localization by IDA is entirely non-destructuve and represents about an order of magnitude improvement over TDR.

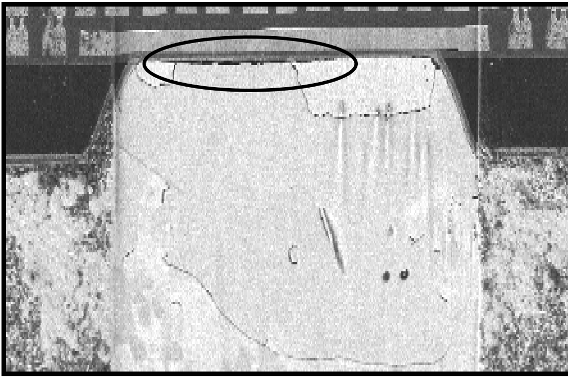
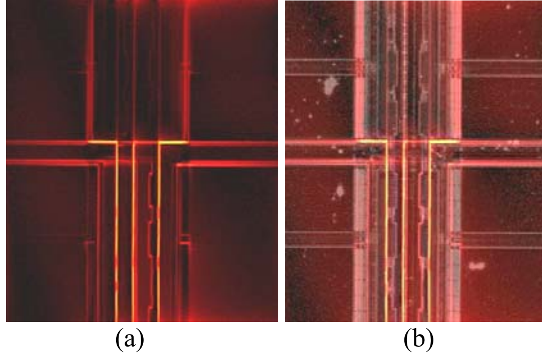


Figure 16: Optical image of cross-sectioned failing bump. Non-wet part of bump is circled.

High Resolution Imaging of Current Paths

As was discussed in the applications section, the resolution of magnetic current imaging is dependent on sensor size and working distance. In many imaging situations, the working distance is the limiting factor to resolution, but in situations where the sensor can be brought close to the current source, submicron imaging of current is possible and can provide critical localization of die level defects.

Two examples of the resolution attainable are given here using a magnetoresistive sensor. In Figure 17 (a), a current image is shown for a front-side accessible chip. A few milliamps of current were applied for this image, which was made at a working distance of about 3 µm. The detailed features in the current image can be used as references to do a two-point alignment for the overlay shown in Figure 17 (b). Figure 18



(a) (b)

Figure 17: (a) High resolution current density image of front-side wire bonded device. (b) Overlay of current density image on optical image.

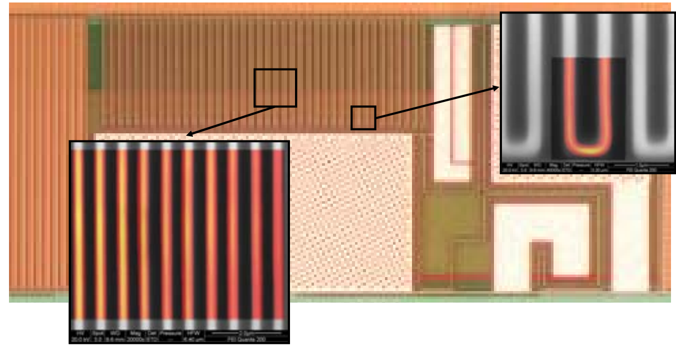


Figure 18: Optical image and high resolution current density images of a serpentine process monitor. High resolution current images are overlaid on SEM images. Line widths are 300 nm with 300 nm spacing.

shows a current image of a serpentine process monitor. The metal lines in this image are 300 nm wide with 300 nm spacing. The current used was 1 mA. The zoomed-in image easily resolves these features and could clearly resolve even finer structures.

Defect Validation

Once a defect has been localized and physical failure analysis has identified it, the failure analyst's work is usually done. In some cases, it is necessary to go one step further and verify that the defect identified is actually the electrical fault under investigation. This final validation can be necessary when ownership of the corrective action is in question or when there is legal liability involved.

Magnetic current imaging can be used to do this final validation (assuming that the electrical defect is still intact) by allowing one to "see" the electron flow through the physical defect. This can be irrefutable proof that the physical defect is also the cause of the electrical fault.

For example, Figure 19 (a), shows an optical image of a PCB with a current image overlay. This was originally a large board that had been sectioned to an area-of-interest as seen in the optical image. The bright spot in the current image corresponds to the location of a plane-to-plane short. After cross-sectioning at this point, a physical defect can be seen between the two shorted layers in Figure 19 (b). However, additional proof was needed to show that the physical defect was the source of the electrical short. Figures 19 (c) and (d) show a current image of this defect taken in cross-section registered with the optical image. The cross-hair positioned on the peak in the current image lines up with the location of the defect in the optical image. This data proves that the observed physical defect is in-fact the electrical fault in this PCB. The full case history of this part is discussed elsewhere in this book.

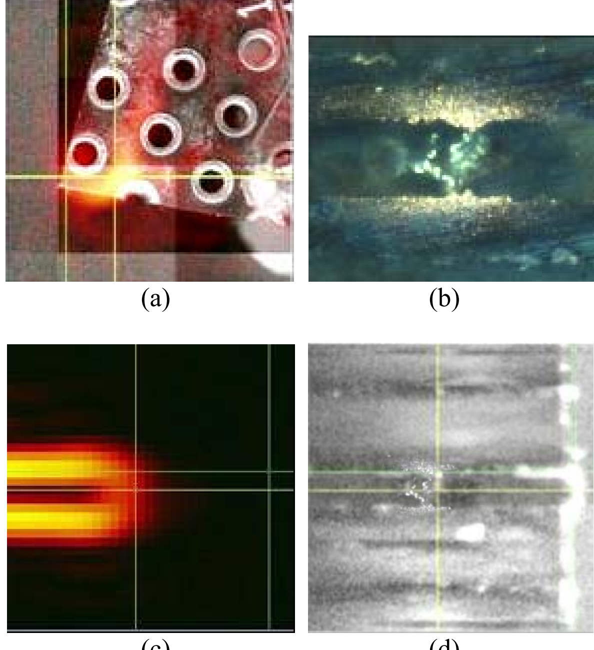


Figure 19: (a) Optical image of PCB with current overlay. (b) Optical image of defect cross-section. (c) Current image obtained from cross-section. (d) Optical image registered with current image where the crosshair marks and validates the current through the defect.

Summary and Future Directions

The examples presented here demonstrate that magnetic current imaging can be a very effective means of fault isolation in die, packages and assemblies. One of the factors making this technology unique is its ability to image current, regardless of voltage or power dissipation, in an integrated circuit or package in a noncontact, nondestructive way. Since low frequency magnetic fields are not affected by any of the materials in an integrated circuit assembly, this technique is especially beneficial in situations where the defect is buried under layers of metal, silicon, or encapsulation materials where other techniques are very limited. Many defects can be imaged for coarse localization without any deprocessing of the sample at all using the high sensitivity of a SQUID sensor. High resolution scans with a magnetoresistive sensor may require deprocessing, but submicron resolution can be achieved when the sensor is close to the source.

The capability of SQUIDs to work at large distances due to their sensitivity makes them ideally suited for isolating defects in packaged devices. They can be used to isolate defects between package, die and interconnect. This high sensitivity allows them to detect currents in the 100's of nanoamps even in packaged assemblies, and enables the detection of subtle differences in current paths between good and failing devices. The latter capability enables isolation of high resistance defects like cracked flip-chip bumps and delaminated vias in packages.

After physical analysis reveals the defect, magnetic current-imaging by SQUIDs or magnetoresistive sensors can be used

to uniquely verify the exact current path and validate the electrical fault.

There is ongoing work to expand the capabilities of magnetic current imaging. High-speed data acquisition for waveform extraction could provide a means of probing switching currents rather than voltages. Quantitative current measurements on test structures could provide valuable data to designers for improving models that would ultimately speed new circuit design.

With the combination of SQUID and magnetoresistive sensors, magnetic current imaging has the capability to meet the fault isolation needs of today's complex structures, including stacked die packaging, and densely populated die with numerous levels of metal. Coupling this technology with advancements in other techniques will provide the failure analyst with a complementary set of tools to isolate the defects of today and tomorrow.

Acknowledgements

The authors would like to thank the suppliers of the samples and data presented here. The authors also thank J.O. Gaudestad, A. Gilbertson, C. Hillman, D. Vallett, and Z. Wang for their assistance and insightful discussions.

References

1. David P. Vallett, "From Microns to Molecules—Can FA Remain Viable Through the Next Decade?," *Proceedings of the 28th International Symposium on Testing and Failure Analysis (ISTFA)*, pp. 11-20 (2002).
2. J. P. Wikswo, Jr. "The Magnetic Inverse Problem for NDE", in H. Weinstock (ed.), *SQUID Sensors: Fundamentals, Fabrication, and Applications*. Kluwer Academic Publishers, pp. 629-695, (1996).
3. S. Chatraphorn, E.F. Fleet, F.C. Wellstood, L.A. Knauss and T.M. Eiles, "Scanning SQUID Microscopy of Integrated Circuits," *Applied Physics Letters*, vol. 76, no. 16, pp. 2304-2306 (2000).
4. Michael J. Caruso, Tamara Bratland, C. H. Smith, and Robert Schneider, "A New Perspective on Magnetic Field Sensing," *Sensors Magazine*, vol. 15, no. 12, pp. 34-46 (1998).
5. H. Weinstock (ed.), *SQUID Sensors: Fundamentals, Fabrication, and Applications*, (Kluwer Academic Publishers, The Netherlands), (1996).
6. E.F. Fleet, S. Chatraphorn, F.C. Wellstood, S.M. Greene, and L.A. Knauss, "HTS Scanning SQUID Microscope Cooled by a Closed-Cycle Refrigerator," *IEEE Transactions on Applied Superconductivity*, vol. 9, no. 2, pp. 3704 (1999).
7. J. Kirtley, IEEE Spectrum p. 40, Dec. (1996).
8. F.C. Wellstood, *et al.*, *IEEE Transactions on Applied Superconductivity*, vol. 7, no. 2, pp. 3134 (1997).
9. S.Y. Yamamoto and S. Schultz, "Scanning magneto-resistance microscopy (SMRM): Imaging with a MR Head," *J. Appl. Phys.*, vol. 81, no. 8, pp. 4696-4698 (1997).
10. B.D. Schrag, X.Y. Liu, M.J. Carter and Gang Xiao, "Scanning magnetoresistive microscopy for die-level sub-micron current density mapping," *Proceedings of the 29th International Symposium on Testing and Failure Analysis (ISTFA)*, pp. 2-5 (2003).
11. Xiaoyong Liu and Gang Xiao, "Thermal annealing effects on low-frequency noise and transfer behavior in magnetic tunnel junction sensors," *Journal of Applied Physics* vol. 94, no. 9, pp. 6218-6220 (2003).
12. K.S. Wills, O. Diaz de Leon, K. Ramanujachar and C. Todd, "Superconducting Quantum Interference Device Technique: 3-D Localization of a Short within a Flip Chip Assembly," *Proceedings of the 27th International Symposium on Testing and Failure Analysis (ISTFA)*, pp. 69-76 (2001).
13. David P. Vallett, "Scanning SQUID Microscopy for Fault Isolation," *Proceedings of the 28th International Symposium on Testing and Failure Analysis (ISTFA)*, pp. 391-396 (2002).
14. A. Orozco, E. Talanova, A. Gilbertson, L.A. Knauss, Z. Wang, L. Skoglund, D. Smith, and R. Dias, "Fault Isolation of High Resistance Defects using Comparative Magnetic Field Imaging," *Proceedings of the 29th International Symposium on Testing and Failure Analysis (ISTFA)*, pp. 9-13 (2003).

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Thermal Defect Detection Techniques

Daniel L. Barton, Ph.D. and Paiboon Tangyunyong, Ph.D.
Sandia National Laboratories

INTRODUCTION

The need for a technique that would produce high spatial and thermal resolution images of microelectronic devices has existed for many years. This became particularly true with the advent of multiple level metallization on IC's. The addition of a second and subsequent levels of metallization significantly reduced defect observability and node access. Many defect types result in higher current flows which generate heat during operation. This is due to the power dissipation associated with the excess current flow at the defect site. Systems to detect this power dissipation can be characterized by their sensitivity to thermal changes and spatial resolution. IR (Infrared) thermal techniques were the earliest available that calculated the object's temperature from it's infrared emission. IR techniques have a fundamental spatial resolution limitation. Liquid crystals have been used with great success since the mid 1960's. Liquid crystals provide a binary response, indicating if the hot area is above the crystal's transition temperature or not. Two factors have contributed to a recent reduction in the effectiveness of liquid crystals. Smaller feature sizes have made the spatial resolution of liquid crystal a factor. In addition, the reduction in power supply voltages and hence power dissipation have served to make the thermal sensitivity of liquid crystals an issue. The fluorescent microthermal imaging technique (FMI) was developed to overcome these issues. This technique offers better than 0.01 ºC temperature resolution and is diffraction limited to 0.3 μm spatial resolution. While the temperature resolution is comparable to that available on IR systems, the spatial resolution is much better. The FMI technique provides better spatial resolution by using a temperature dependent fluorescent film that emits light at 612 nm instead of the 1.5 μm to 12 μm range used by IR techniques. This chapter reviews the background material, operating principles, and image characteristics for the these three thermal imaging techniques.

BLACKBODY RADIATION [1]

Blackbody radiation physics describes the process by which all heated objects emit radiation to their surroundings. It is well known that heated objects emit radiation. . The wavelength distribution of this radiation is dependent on the temperature of the object. For example, as objects are heated beyond the red hot temperature, approximately 700 ºC, the amount of radiation being emitted in the visible range increases and the objects begin to turn orange, then yellow, and eventually to a bluish-white color. What is not readily apparent to us is that most of the light being emitted by hot objects is infrared.

The connection between the peak radiation wavelength and temperature forms the basis for infrared (IR) temperature measurement. Physicists in the late 19th century knew the relationship between the temperature of a blackbody, a body that absorbs all of the radiation incident upon it, and the peak wavelength of the radiation being given off by that body, but could not describe that relationship mathematically. The most important observation was that all blackbodies at the same temperature emitted radiation with the same spectral distribution, regardless of their composition. The spectral distribution of the radiation emitted by a blackbody is known as the spectral radiancy, $R_T(\nu)$, where ν is the frequency of the radiation. The quantity $R_T(\nu)$ is defined so that $R_T(\nu)d\nu$ is the energy emitted per unit time in the frequency range ν to $\nu + d\nu$ from a unit area on a surface at temperature T.

The integral of the spectral radiancy $R_T(\nu)$ over all frequencies is known as the radiancy, R_T, or

$$\mathrm{R_T} = \int_0^\infty R_T(\nu)d\nu .$$

The relationship between the radiancy and temperature was first formulated in 1897 and is known as Stefan's Law,

$$\mathrm{R(T)} = \sigma \mathrm{T}^4 .$$

Though Stefan's Law was defined empirically in its original form, the constant σ came to be known as the Stefan-Boltzmann constant with a value of

$$\sigma = 5.67 \times 10^{-8} \frac{\mathrm{W}}{\mathrm{m^2 \cdot K^4}} .$$

Finally, the relationship between wavelength peak and temperature is known as Wein's displacement law, which can be written,

$$\lambda_{MAX} = \frac{2.898 \cdot 10^{-3}}{T}$$

where T is the temperature in degrees Kelvin. The result of the equation is λ_{MAX} in meters.

Classically, a small hole in a cavity is the best approximation of a blackbody. From this approximation, we can easily find the number of modes present in the cavity and assign an energy to each of these modes which should describe blackbody radiation. If we realize that the spectral radiancy of a blackbody is directly proportional to that of a cavity,

$$\rho(\nu) \propto R(\nu)$$

and use the equapartition of energy theory where each mode of the cavity is assigned an energy, $\bar{\mathcal{E}} = kT$, we end up with the Raleigh-Jeans formula for blackbody radiation,

$$\rho_T(\nu)d\nu = \frac{8\pi\nu^2 kT}{c^2} d\nu .$$

This formula proved to be accurate at low frequencies, but failed miserably at high frequencies in the ultraviolet end of the spectrum. This theory was ultimately known as the ultraviolet catastrophe. The reason for the failure of this theory is simple. The equapartition theorem is only valid for a continuous distribution of energies. However, the energy of an electromagnetic wave is quantized in units of hν. The theory of energy quantization led Max Planck to the correct equation to describe the blackbody radiation spectrum

$$\rho_T(\nu)d\nu = \frac{8\pi\nu^2}{c^3} \cdot \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1} d\nu .$$

In most applications, the object under examination will not be a perfect blackbody. As such, the radiancy will have to be corrected for the emissivity of the material. The emissivity, e, is defined through Stephan's Law which relates the radiancy to temperature,

$$R(T) = e \cdot \sigma T^4 .$$

By definition, the emissivity of a material is the ratio of the energy radiated by a given object to the energy radiated by a blackbody at the same temperature. The emissivities of many materials have been measured and are generally available.

Returning again to figure 1, we see that even for objects at 2000 K, there is only a small portion of the radiated energy in the visible portion of the spectrum. In fact, the majority of the information relating to the object's temperature is well into the infrared region of the electromagnetic spectrum. In the semiconductor industry, there are few devices that operate at such high temperatures. If we expand the scale shown in figure 1 to temperatures around room temperature, we begin to understand the difficulties in performing high spatial resolution thermal imaging on semiconductor devices. Figure 2 shows blackbody radiation spectra for objects between 250 K and 350 K. Of particular interest in this figure is the curve for 300 K. It is clear from this curve, that very little energy is emitted at wavelengths less than about 3 μm and most of the energy is emitted at wavelengths greater than 5 μm. In order to collect the infrared radiation information from objects near room temperature, a detector sensitive well into the infrared range is needed.

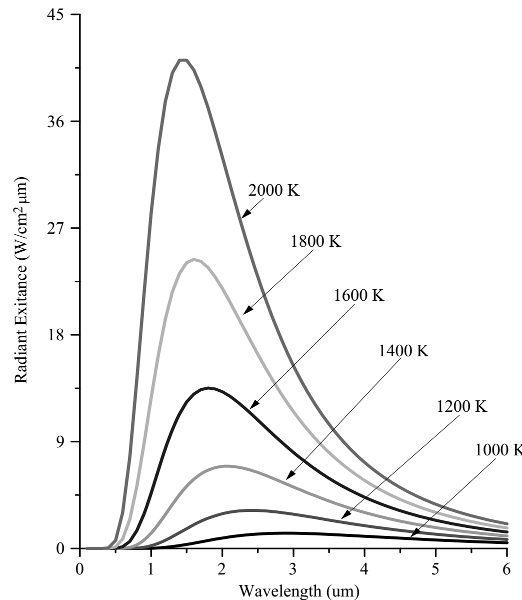


Figure 1 - Spectral radiance of blackbodies at 1000 K to 2000 K.

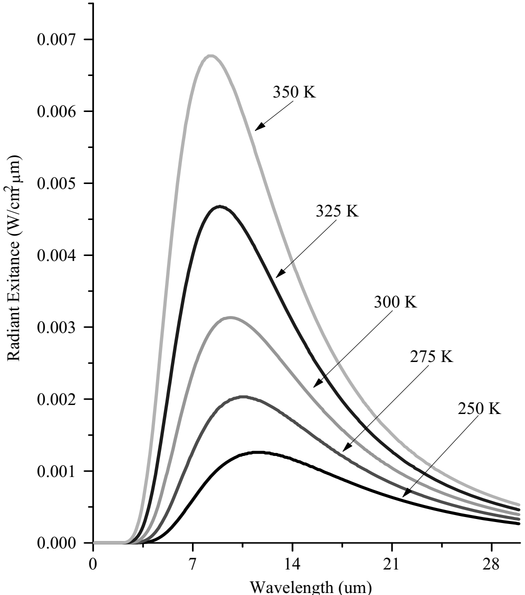


Figure 2 - Spectral radiance of blackbodies at 250 K to 350 K.

INFRARED (IR) THERMOGRAPHY [2-5]

It is evident from figure 2 that the radiation emitted by the sample in the 3 μm to 12 μm range must be analyzed to measure typical device temperatures. Most IR thermography systems use one of two types of detectors; indium antimonide or mercury cadmium telluride. Indium antimonide (InSb) detectors are sensitive in the wavelength range 1.5 μm to 5.5 μm. Mercury cadmium telluride (HgCdTe) detectors are sensitive over the range of 8 μm to 12 μm. Both detectors offer similar temperature sensitivities and ranges, but InSb operates at shorter wavelengths and should have somewhat better spatial resolution.

IR thermal imaging systems have excellent potential for temperature resolution but they suffer from a fundamental limitation on spatial resolution. In general, the spatial resolution of a microscope is limited by the wavelength of light. The relationship between resolution and wavelength as given by Lord Rayleigh's criteria can be written [6]

$$\text{Resolution} = \frac{0.61 \cdot \lambda}{\text{N.A.}}$$

where λ is the wavelength of light being imaged and N.A. is the numerical aperture of the microscope. This result although very simple, provides a good estimate of the resolving power of a microscope system. This relation clearly shows that IR thermal systems that gather radiation in the 1.5 μm to 12 μm range will not be able to resolve sub-micron structures that are found on modern VLSI integrated circuit technologies.

IR thermal systems rely on directly sensing the emitted infrared radiation from objects to measure their temperature. In general, there are several ways to accomplish this measurement. The simplest method is measure the radiance and, if the emissivity of the material is known, directly compute the temperature of the sample. Systems of this type use a very simple photovoltaic type detector that is sensitive in the IR wavelengths. Cooled InSb photovoltaics are perhaps the most common. These detectors are generally sensitive in the 1.5 μm to 5.5 μm range.

To directly measure the temperature of a sample, the radiance is measured first. The radiance from a sample at a given temperature is related to the radiance that would be collected from a blackbody at the same temperature by the emissivity, or

$$R_T = e \cdot R_{TBB}$$

where R_T is the radiance from the sample, e is the sample's emissivity, and R_{TBB} is the radiance of a blackbody at the same temperature. Ideally, if the sample's emissivity is known, its temperature can be calculated if the relationship between the radiance collected by a given system from a blackbody and its temperature is known.

In order to increase the accuracy of the temperature measurement, the radiance that is reflected by the sample must be accounted for. Thus, the total radiance collected by a system from a sample would be

$$R_{Total} = R_T + (1 - e)R_0$$

where R_0 is the radiance emitted by the ambient background. Combining the last two equations, the total radiance collected by the system is

$$R_{Total} = e \cdot R_{TBB} + (1 - e)R_0 .$$

If the emissivity is known and the ambient radiance can be accounted for, the temperature of the sample can be found.

Lastly, the spectral response of the system must be considered. In general, the system response will not be constant over the spectral range on interest. The introduction of an "effective" blackbody radiance representing a convolution of the blackbody radiation spectrum and the IR microscope/detector response must be known. This information has usually been characterized by the manufacturer and is incorporated into their blackbody radiance to temperature conversion algorithm.

Figure 3 shows a series of example images of an integrated circuit test structure taken with an IR thermal imaging system. The structure is an electromigration test

structure that has 4 μm wide aluminum lines separated by 4.5 μm. The lines are 3470 μm long yielding a total resistance of about 30 Ω. The images in figure 3 show that IR systems can easily sense hot areas on integrated circuits at relatively low power densities but will have difficulty resolving features less than about 15 μm. The image in figure 3(c) was taken after the surface of the chip was painted with a black non conductive paint with an emissivity value of 0.96. With the growing use of flip chip packaging technology, the need for backside localization of thermal features may create a resurgence in the use of IR thermal analysis for defect localization within integrated circuits. Regardless, IR systems will continue to be used in areas such as multi-chip modules, circuit boards, and IC packaging issues as they have been for years where absolute, non-contact measurements are essential and sub-micron spatial resolution is not needed.

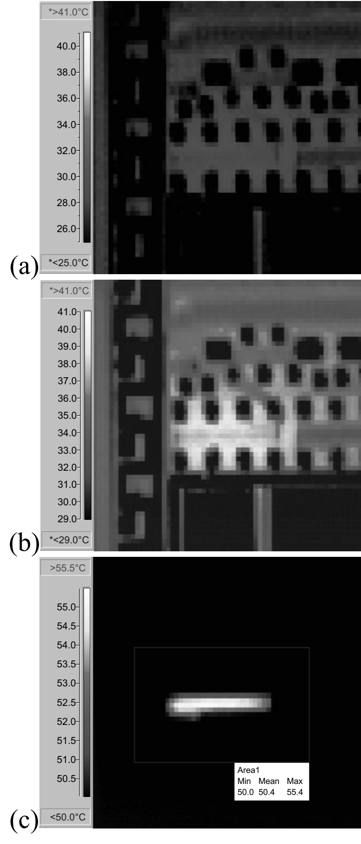


Figure 3 - IR thermal images of an electromigration test structure. (a) I = 70 mA (actual surface temperature is 36.7°C), (b) I = 100 mA (actual surface temperature is 52.0°C), (c) sample coated with black paint, I = 100 mA (actual surface temperature is 55.4°C).

HISTORY OF LIQUID CRYSTALS [7]

The history of liquid crystals dates back more than a century. The first documented observance of a material that changed from a crystalline solid to an opaque liquid and then to a clear liquid as its temperature was increased was in 1888 by Friedrich Reinitzer. Reinitzer was an Austrian botanist and chemist at the University of Graz who was synthesizing esters of cholesterol from plants and animals and noticed the state changes in these materials which he called "double melting". Although he is credited with the discovery of the liquid crystal state, he could not explain the opaque liquid state and the "double melting" phenomenon which he had observed. Reinitzer turned to Otto Lehmann, a professor of physics at the Technical High School of Karlsruhe. Professor Lehmann was noted as being Germany's leading crystallographer at the time. Lehmann discovered the optical anisotropy in Reinitzer's esters which lead him to create the terms "fluid crystals" and "liquid crystals". Based on his observations, he was able to argue that the optical anisotropy which he had observed was due to elongated molecules in the opaque liquid.

Over the next decade, approximately fifteen new liquid crystal compounds were discovered even though there was no knowledge of the connection between the molecular shape and the liquid crystal state. It was not until about 1900 when Daniel Vorländer, a professor of chemistry at the University of Halle, started systematic research into the connection between molecular structure and the liquid crystalline state. In 1908 he established the rule that liquid crystal materials must have molecules with a linear shape. When he retired in 1935, approximately 1100 liquid crystal materials were known. Vorländer has come to be known as the father of liquid crystal chemistry. Between 1935 and about 1960, little significant research was done on liquid crystals. After 1960, a large resurgence in research was brought about by interests in liquid crystal displays and thermography. The Liquid Crystal Institute was founded at Kent State University in the 1960's. From the 1960's on, there have been a large number of developments in liquid crystal technology inspired by growth in the electronics industry. At present, liquid crystal displays have matured to where they are on the verge of replacing cathode ray tubes as our primary source of full color, high resolution displays. It is estimated that there are currently more than 20,000 liquid crystal compounds which have been developed.

LIQUID CRYSTAL MATERIALS [7-13]

Liquid crystals are known by a variety of names such as fluid crystals, crystalline liquids, mesogenic compounds, mesophases or mesomorphic phases. Whatever the name, they all refer to materials which possess at least one liquid crystalline state. A liquid crystalline state (of which there are many types) is one which the material has some characteristics which are associated with crystalline solids and some which are associated with liquids (hence the name liquid crystals).

The characteristics which are like crystals are usually associated with optical anisotropy while those of liquids relate to areas like molecular mobility and other fluid-like properties. Compared to isotropic liquids, liquid crystals have a much higher state of order while compared to solid crystals, they have a higher intermolecular and intramolecular mobility.

In order to introduce materials that have liquid crystalline states, an example is shown in figure 4. The molecule in figure 4 (4,4'-azoxyanisole) is typical of many of the calamatic types of liquid crystals in that it has a long, rod-shaped molecule which is the classic shape molecule for which liquid crystals are known. Other types of molecules are discotic (disc shaped molecules) and sanidic (lath-like or board-like molecules). The materials used for failure analysis purposes have always been calamatic types. The material in figure 4 is a solid crystal at temperatures less than 118 °C, is a liquid crystal from 118 to 135 °C and becomes an isotropic liquid above 135 °C. Most references would use the shorthand notation "cr 118 N 135 is" to describe the temperature dependence of this material. The "N" in the notation refers to the type of liquid crystalline state that this material usually obtains. In this example, the material has a nematic liquid crystal state. Nematic comes from the greek νημα or thread referring to the thread-like defects often found in nematic crystals. Figure 5(a) shows a representation of a nematic crystal which illustrates the rod-like structure of the molecules and the order associated with these materials. Evident in figure 5(a) is the alignment of the molecules along an axis usually referred to as $\vec{n}$, the director. The order parameter, *S*, describes how well the molecules are aligned with the director and is calculated by using the angle, β, between the long axis of the molecule and the director.

$$S = \frac{1}{2}\left\langle 3\cos^2\beta - 1\right\rangle$$

For a highly ordered material, the value of *S* will approach unity. For an isotropic liquid, as shown in figure 5(b), the value of *S* will be small. Other types of liquid crystals phases are smectic (figure 5(c)) and cholesteric (figure 5(d)). These types of materials are similar to nematic types in that they have a relatively high order parameter but they have systematic differences which affect their properties.

The term smectic comes from the Greek σμηγμα, which means soap, and is used because the first smectic phases were observed in ammonium and alkali soaps. The similarities between the order of smectic and nematic liquid crystals is readily apparent but the difference is subtle. The smectic types have a second level of order which aligns the molecules in rows or layers. Cholesteric types were named after the cholesterols in which the phase was first identified and have a helical change in the alignment axis for the molecules. This difference gives cholesteric liquid crystals a unique ability to change colors by having reflection properties which are sensitive to wavelength.

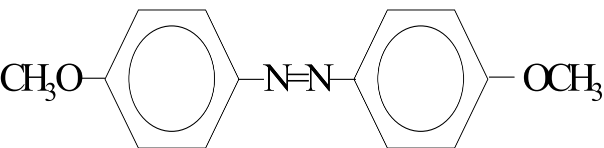


Figure 4 - 4,4'-azoxyanisole.

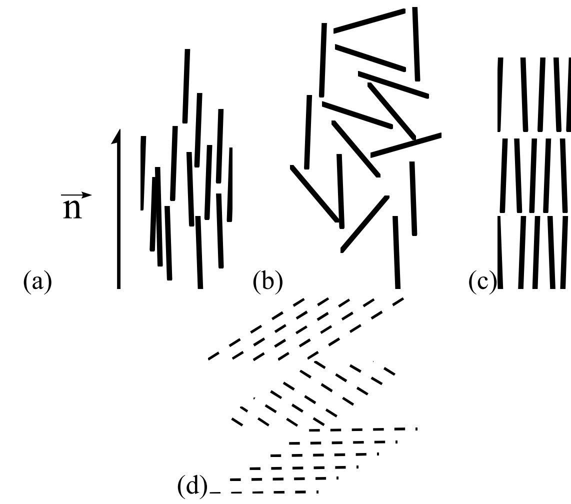


Figure 5 - (a) Nematic, (b) Isotropic, (c)Smectic, and (d) Choloesteric phases of liquid crystals. The vector $\vec{n}$ *is the director for (a), (b), and (c). (d) has a helical change in the director with depth.*

Of the thousands of known liquid crystal materials, all of the calamatic types have a similar molecular structure; that of a long rod-like structure. The structure can be described as shown in figure 6. The presence of two aromatic (also, heterocyclic or alicyclic kinds) rings which are linked together, either directly or through a linkage group, and a side chain and terminal group are the common elements. The material properties and types of liquid crystal phases are determined by the chemical structures of the parts of the molecule in figure 6.

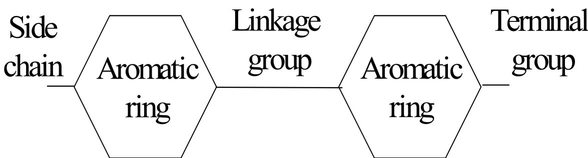


Figure 6 - General calamatic liquid crystal strtucture.

Changes in the various components of a given liquid crystal molecule affect the transition temperatures of each of the phases. Since most of the liquid crystals commonly used for hot spot detection on microelectronic devices are in the biphenyl family, examples from this family are shown in figure 7 to demonstrate transition temperature differences. The effects illustrated by the examples in figure 7 are that an increase in the number of rings in the aromatic core and an increase in the size of the terminal groups tend to increase the clearing point temperature. While the trend is clear in figure 7, changes in aromatic ring type and terminal group type, although well documented, can make some trends difficult to predict.

The use of liquid crystals for thermal mapping was introduced in 1963 and used cholesteric type materials. The technology was remarkable in that many thermally active objects, including integrated circuits, could be non-destructively tested in a fairly simple manner. Cholesteric liquid crystals have an interesting property of being able the change the wavelength of maximum scattering with temperature while they are in their cholesteric temperature range. Many references are available which describe the procedures for using cholesteric liquid crystals for thermal mapping ([14-18] for example). Since nematic liquid crystals have become the materials of choice for thermal mapping of modern integrated circuits, their optical and thermal properties will be reviewed.

Two of the molecules shown in figure 7 are the most common liquid crystals used in failure anlaysis. The second molecule from the top is the familiar, commercially available, K-18 liquid crystal. K-18 is chemically known as $C_{19}H_{21}N$ or 4-cyano-4'-n-hexyl-1, 1'biphenyl and has a state transition temperature from its nematic to isotropic state just above room temperature. The transition temperature of K-18 has undoubtedly lead to its popularity. The molecule on the bottom of figure 7 is K-21 which has a transition temperature about 15 degrees higher than K-18 making it more suitable for integrated circuits whose temperatures tend to rise above room temperature during normal operation. The use of K-21 is more convenient than cooling the sample below K-18's transition temperature in more applications.

CN–[ring]–[ring]–[ring]–C_5H_{11}

N 239 is

CN–[ring]–[ring]–C_5H_{11}

N 27 is

CN–[ring]–[ring]–C_6H_{13}

N 35 is

CN–[ring]–[ring]–C_7H_{15}

N 42 is

Figure 7 - Effects of changes in aromatic core and terminal groups on clearing point temperatures. The nomenclature "N 35 is" indicated a transition from a nematic liquid crystal to an isotropic liquid at 35 °C.

OPTICAL PROPERTIES OF NEMATIC LIQUID CRYSTALS [7, 9]

The use of nematic liquid crystals for failure analysis is based on their ability to give a detectable change when they are heated above a known temperature. Liquid crystals are well matched to failure analysis because their change with temperature is easy to detect with common optical equipment. Figure 8 shows the thermotropic optical birefringence of a nematic liquid crystal with a transition temperature of about 35 °C. Figure 8 illustrates the difference in refractive index between the ordinary ray (n_o) and the extraordinary ray (n_e) at temperatures below the transition temperature and total loss of birefringence above the transition temperature.

The optical anisotropy of liquid crystals is due to the parallel order of the molecules while in the nematic state. The order parameter, *S*, was defined earlier and its relationship to the relative positioning of the molecules is important to an understanding of the birefringence of these materials. The birefringence of the nematic state is defined as

$$birefringence = n_e - n_o .$$

To be more specific in terms of the physical parameters of liquid crystals, the birefringence can also be written

$$\frac{n_e^2 - n_o^2}{\overline{n}^2 + 2} = \frac{\rho N_A}{3\varepsilon_0 M}\left(\alpha_l - \alpha_t\right)S$$

where n_e and n_o are the indices of refraction, ρ is the molar mass, N_A isAvogadro's number, ε_0 is 8.86x10-12 As/Vm, and *S* is the previously defined orientational order parameter. The remaining parameters are the longitudinal

and transverse polarizabilities, α_l and α_t respectively. The average polarizability and average index are given by

$$\overline{\alpha} = \frac{1}{3}(\alpha_l + 2\alpha_t)$$

$$\overline{n^2} = \frac{1}{3}\left(n_e^2 + 2n_o^2\right).$$

The equations above indicate that the birefringence is determined by the order parameter, the molecular polarizability $(\alpha_l - \alpha_t)$, and the reciprocal molar volume, ρ / M.

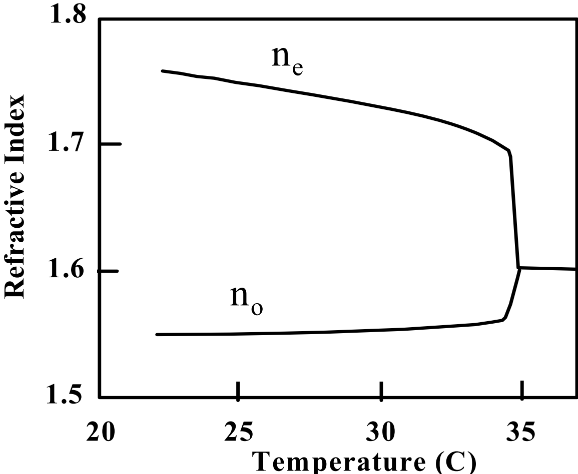


Figure 8 - Thermotropic birefringence of a nematic liquid crystal with a clearing point of 35 °C.

USE OF LIQUID CRYSTALS FOR FAILURE ANALYSIS

The application of liquid crystals to integrated circuit failure analysis can be achieved in a number of different ways and with quite a few types of materials. The first applications used cholesteric liquid crystals [14-18]. In most of these applications, the circuit had to be covered with a thin, black material. The easiest coating was the carbon emitted by a common wax candle. The black layer is needed to allow the color changes which occur in cholesteric liquid crystals to be observed under oblique illumination. The disadvantage of this method is that the opaque coating obscures the image of the circuit being tested. The references cited indicate spatial resolutions of 15 to 20 μm and temperature resolutions of 0.1 K at best.

Because of the spatial and thermal resolution limitations observed with cholesteric liquid crystals, they have largely been replaced by nematic types [19-24]. The main difference between the two applications is that cholesteric types can generate color isotherms of the area being tested while the nematic types only indicate temperature changes by crossing the nematic to isotropic transition in the material. The ability to only sense that a given region on a circuit has heated the liquid crystal beyond the transition temperature is a limitation of the technique but is outweighed by the ease of use and sensitivity of this method.

Infrared thermal imaging, as discussed in the previous section, is the only available non-contact thermal mapping technique. Liquid crystals and fluorescent microthermal imaging (to be discussed next) both rely on the presence of a thermal sensing film placed on the sample being tested to provide an indication of thermal activity. The goal with both of these contact type methods is to apply a film which is thick enough that the observable change is detectable but thin enough to achieve maximum thermal sensitivity.

Liquid crystals are typically diluted with solvent (such as freon-TF or pentane) and applied to the surface with en eye dropper or syringe. The mixture is allowed to spread and the solvent allowed to evaporate. The resulting film should be about five to seven microns thick. When this film is viewed under un-polarized light, the surface should look slightly cloudy. When viewed through crossed linear polarizers, the appearance should be mottled. Figure 9 illustrates the hot spot detection principle using nematic liquid crystals. With the polarizers crossed, the presence of the nematic liquid crystal changes the polarization of the incident light enough that a reasonable image of the circuit can be made. Over the hot spot, the crystal goes into an isotropic phase where the polarization of the incident light is no longer altered. As a result, these areas appear dark in the microscope.

The entire equipment setup is illustrated in figure 10 where the coated sample is placed on a heated stage under a microscope which has two linear polarizers, one in front of the light source and one in front of the camera or microscope eyepieces (typically called the analyzer). The most common technique is to heat the stage as close as possible to the transition temperature of the liquid crystal being used. Typically, the stage can be heated to within 0.1 K from the transition temperature with little difficulty. More elaborate procedures have been reported [20] to achieve stable temperatures much closer to the transition temperature. A variation is to heat the sample above the transition temperature and look for the last clear spot which identifies the hottest portion of the circuit. By using observing the liquid crystal while varying the temperature, the best thermal resolution can be achieved but only for a short period of time. Articles have been published [21] which use a combination of stage heating and additional heating from the illumination source to achieve very high temperature sensitivity.

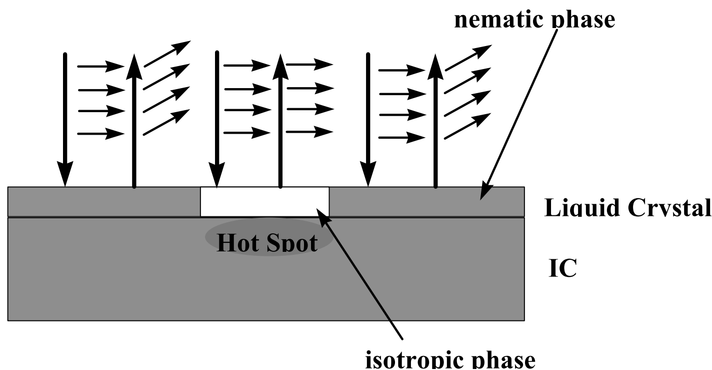


Figure 9 - The use of nematic liquid crystals for failure analysis.

An example image of an electromigration test structure is shown in figure 11. The image in figure 11 was made using nematic liquid crystal K-18 while holding the stage at room temperature, approximately 23 °C and applying 100 mA of current through the structure. The image shows the desired mottled appearance of the nematic phase of the film and the desired black color in the isotropic phase areas. This example demonstrates that liquid crystals has the capability of resolving features around 4 µm wide (the metal line width in this example) without difficulty. Higher spatial resolution is easily achieved with better temperature control.

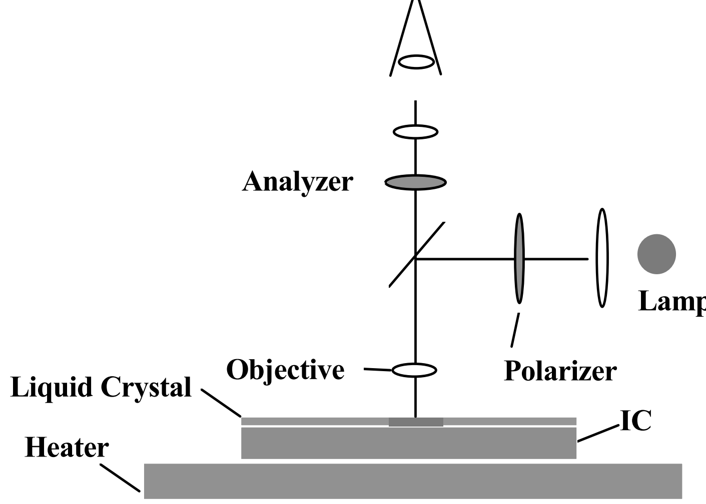


Figure 10 - Equipment setup for liquid crystal hot spot detection.

Figure 11 - Example image made using K-18 liquid crystal.

FLUORESCENT MICROTHERMAL IMAGING [25-29]

HISTORY

The influences of lower power supply voltages, increased number of thin film layers and smaller transistors have made liquid crystals much less effective. However, the need for an alternative hot spot detection technique with greater thermal sensitivity and better spatial resolution has been largely filled by Fluorescent Microthermal Imaging (FMI). The concept of using a film with a temperature dependent fluorescence quantum yield to generate high resolution thermal maps of integrated circuits was first published in 1982 [25]. The first work described a technique which could yield thermal images with a thermal resolution of 0.006 ºC and a spatial resolution which was equipment limited to 15 µm. A follow-up article demonstrated the spatial resolution capability of the technique, which was measured to be 0.7 µm [26]. Subsequent research [27-29] has developed the understanding of the fundamental limitations of FMI and the operational procedures needed to insure maximum performance when applied to modern IC technologies. This research has also refined the hardware used for FMI to improve its usability.

EUTTA COMPOUND SPECIFICS: [30-33]

During the late 1950's and early 1960's, rare earth chelates were identified for use in lasers because of their well known fluorescence responses to UV or near-UV excitation sources. One of these compounds, EuTTA (europium thenoyltrifluoroacteonate) has been the focal point for the fluorescent microthermal imaging technique. The chemical structure of EuTTA is shown in figure 12.

Figure 12 - Chemical structure of EuTTA.

EuTTA is not the only compound available for FMI. In fact, there are chelates of all of the rare earth elements which include La, Sm, Eu, Gd, Tb, Dy, Tm, Yb, and Lu. The europium system was ultimately selected as the most suitable because of its temperature characteristics, emission/absorption characteristics, availability, and other qualities. There are several other europium compounds which might be suitable for FMI. Other β–diketone chelates of europium are available such as europium benzoylacetonate, europium dibenzoylmethide, and europium hexfluoroacetonate in addition to EuTTA. EuTTA however, has the best fit for temperature dependent fluorescence quantum yield in the temperature range near room temperature.

The temperature dependent fluorescence quantum yield is the key property of EuTTA in this process. figure 13 shows the molar extinction coefficient (or loosely, the absorption spectra) versus wavelength for EuTTA in an ethanol solution. While FMI requires that EuTTA be suspended in a solid matrix, the data in figure 13 indicates the excitation wavelengths of interest. First, there is a broad absorption peak centered around 335 nm. This is where the TTA ligand absorbs energy. After about 360 nm, the amount of absorption falls off strongly. The two peaks at 460 nm and 525 nm are consistent with Eu^{3+} levels and are not of interest from an excitation viewpoint. What is of interest is the lack of absorption for wavelengths much above 500 nm. This allows for a strong separation between the excitation source and the fluorescence emission.

The ultraviolet radiation used to excite the EuTTA fluorescence does so through a series of intramolecular energy transfers as illustrated in figure 14. The TTA ligand absorbs the UV light then transfers the energy to the europium ion. While several fluorescence lines are excited, the transition from the Eu^{3+} $^5D_0$ energy level to the $^7F_2$ level, as is shown in figure 15, is the most efficient. This transition generates the bright fluorescence line at 612 nm which is used for FMI. Figure 16 shows the emission spectrum for crystalline EuTTA at 25 ºC.

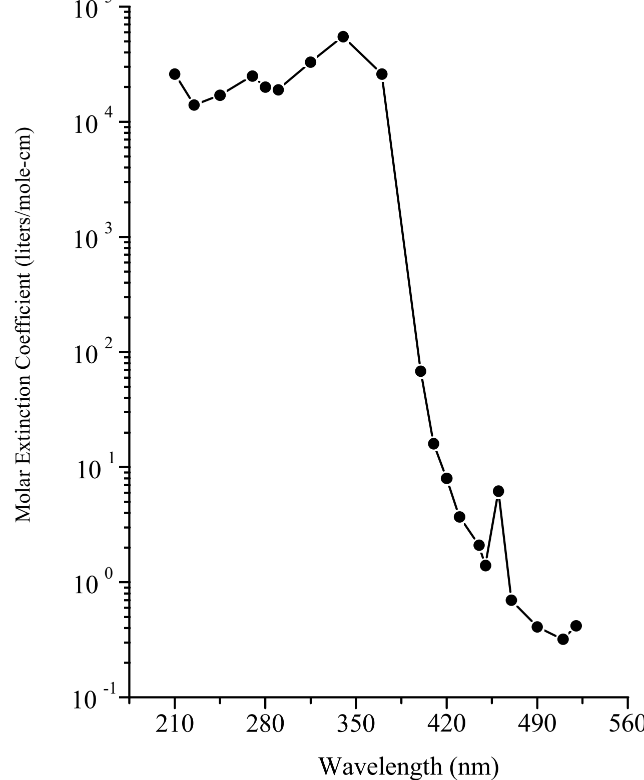


Figure 13 - Absorption spectra for EuTTA (in solution) [30].

For thermal imaging applications, we need to know how the intensity of the fluorescent emission from the compound changes with temperature. Figure 17 shows the measured absolute quantum yield (fluorescence intensity) versus temperature and figure 18 shows the decay time of the fluorescence yield versus temperature. Both of these plots were generated for EuTTA in an ether:isopentane:ethanol (5:5:2) solution. For the FMI application, a curve will need to be generated for each compound mixture that is used. These data have been included to illustrate the temperature dependence of EuTTA.

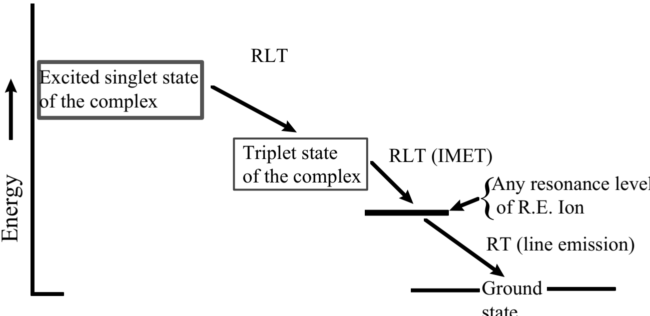


Figure 14 - Intermolecular energy transfer processes. RLT denotes non-radiative transitions while RT is the radiative transition.

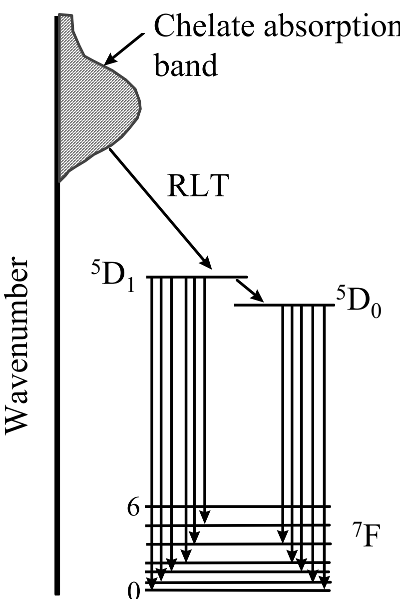


Figure 15 - Europium chelate fluorescence

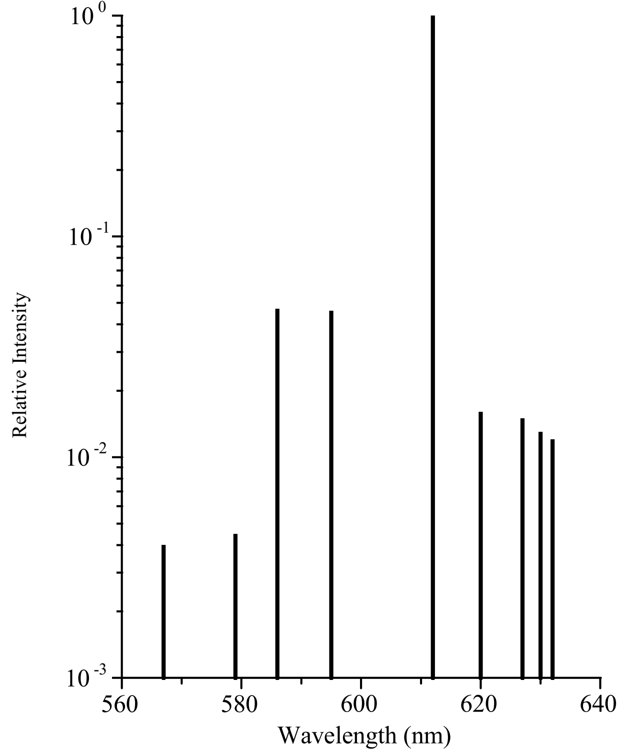


Figure 16 - Emission spectra of EuTTA (Crystalline) at 25 ºC [30].

where Q(T) represents the quantum efficiency of the compound. The behavior of this compound over this broad temperature range is very predictable and provides a simple way to calculate the temperature change on a surface by imaging its change in quantum yield.

The information in figures 17 and 18 illustrates the intermolecular energy transfer between the TTA ligand and the europium ion. First, notice that the quantum efficiency falls off faster than the decay time with increasing temperature. This shows that we may have a low quantum efficiency even at low temperatures due to a loss of excitation energy of the Eu^{3+} ion. The change in quantum efficiency with temperature is an indicator of the quenching of the whole system, while the fluorescence decay time is an indication of quenching of the fluorescence in the europium ion itself.

The standard application technique for using EuTTA for FMI is to incorporate the chelate into a PMMA (polymethylmethacrylate) matrix. A typical starting point is a solution consisting of 1.2 wt% EuTTA, 1.8 wt% PMMA, and 97 wt% MEK (methylethylketone). The MEK evaporates rapidly leaving the EuTTA/PMMA mixture on the sample [25]. Typically this mixture is spun on the sample and allowed to cure in an oven at 125 ºC for about 30 minutes. Ideally, the film should only be several optical absorption lengths thick. At an excitation wavelength of 365 nm, a 300 nm film is approximately 3.5 optical absorption lengths thick. The idea is to have the

film thick enough that most of the UV light is absorbed, but thin enough that the thermal profile of the sample surface is not distorted. As we will find out in later sections, the image processing required to create a thermal image reduces the influence of film non-uniformity on image quality. As such, the film should be as uniform as possible, but great pains to achieve perfect uniformity of film composition and thickness are not necessary.

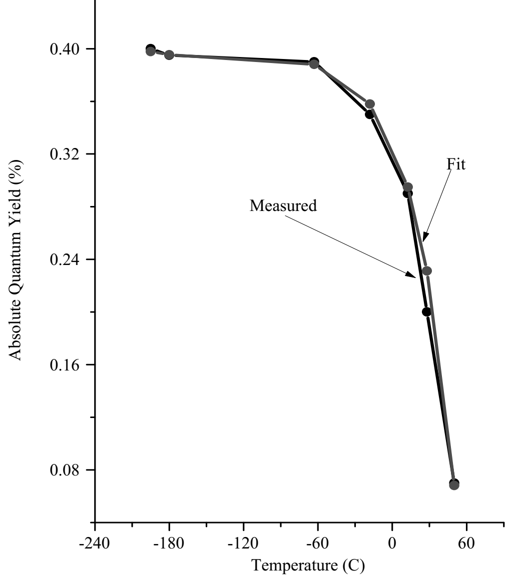


Figure 17 - Absolute quantum yield for EuTTA [31]

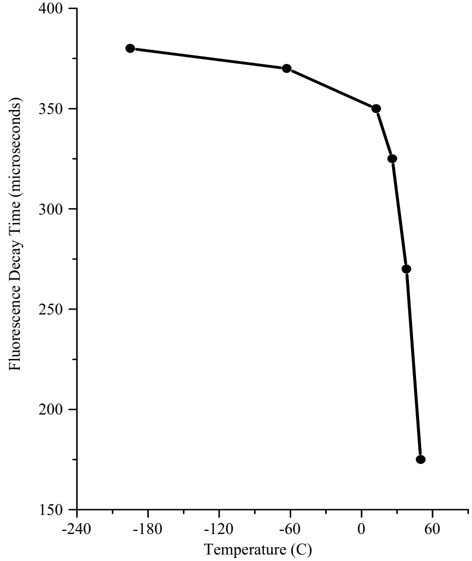


Figure 18 - Fluorescence decay time for EuTTA [31].

The EuTTA/PMMA composition can be varied as needed for any specific application. Adjusting the EuTTA content will change the amount of fluorescent light emitted from the coated sample and changing the amount of MEK in the mixture will thin the solution out for applications where spinning the sample is not practical. For instance, integrated circuits in packages are sometimes difficult to mount on a photoresist spinner. The use of a thinned out mixture would allow a thin film to be deposited without spinning the IC. Usually, spinning a packaged IC will cause the mixture to accumulate around the ball bonds leaving a thick film in these areas. The thick film is often not a problem, unless the signal input structures are the areas of interest. For these applications, a thinner mixture or a higher spin rate would be in order.

The advantage of PMMA is that it can easily be removed once the thermal analysis is completed. Rinsing the sample in acetone will dissolve the film in several minutes. The use of other polymers such dPMMA, (perdeutero-poly-methylmethacrylate) will provide a stronger temperature dependence, but the additional cost of dPMMA is not justified. Other matrixes, such as spin-on glass have been successfully used, but make film removal more difficult.

Regardless of the film type used, accurate absolute temperature measurements are possible but, because of the differences in the logarithmic slope of the quantum efficiency versus temperature curves for different materials, an accurate film calibration should be done for each type of pre-mixed solution. The calibration curve can be easily obtained by using a hot/cold stage, a calibrated thermocouple, and the camera to be used for FMI. Simply record the total emission observed by the camera in a given time period with the sample at a given temperature. Varying the hot/cold stage temperature over as large a range as possible will give the best results. Samples such as blank wafers would be good for this process since, during the measurements, they will be close to the hot chuck temperature. The emission versus temperature data can easily be plotted and a logarithmic slope can be found. Unless the composition of the mixture changes drastically, this measured slope need only be done once, especially if only relative temperature measurements are needed.

For higher temperature measurements, another europium compound, perdeutero-(tris-6,6,7,7,8,8,8-heptafluoro-2,2-dimethyl-3,5-octandionato) europium (dEuFOD) may be used up to about 200 ºC. dEuFOD has a much weaker temperature dependence both for fluorescence quantum yield and fluorescence lifetime than EuTTA [34].

IMAGE PROCESSING

Now that we have an understanding of the fluorescent film properties which allow us to use this technique to generate thermal images, we need to cover how that information is converted to temperature data. Returning to figure 17, we recall that the quantum efficiency versus

temperature can be represented as an exponential. Fitting an exponential function to the data in figure 17 gave us,

$$Q(T) = 0.398 - 0.07 \cdot e^{0.031 \cdot T}$$

as the quantum efficiency versus temperature relationship. The light intensity at a given point, (x,y), on the image can be represented by

$$S(x,y) = I(x,y) \cdot \eta(x,y) \cdot r(x,y) \cdot Q(T(x,y))$$

where I(x,y) is the illumination intensity, η(x,y) is the optical collection efficiency, r(x,y) is the sample reflectivity, and Q(T(x,y)) is the quantum efficiency. In order to remove all spatial artifacts included in the I, η, and r terms, and leave an image containing only thermal information, we can divide an image taken with the sample under bias, i.e. a hot image, by one without bias, i.e. a cold image. The result is a map of the ratio of quantum efficiencies between hot and cold images

$$S_R(x,y) = \frac{S_H(x,y)}{S_C(x,y)} = \frac{I(x,y) \cdot \eta(x,y) \cdot r(x,y) \cdot Q(T_H(x,y))}{I(x,y) \cdot \eta(x,y) \cdot r(x,y) \cdot Q(T_C(x,y))} = \frac{Q(T_H(x,y))}{Q(T_C(x,y))}.$$

If we were working with a pure exponential, this ratio would be directly proportional to the difference in temperature, T_H-T_C. The problem with doing this directly is the leading constant (equal to 0.398 in this example) in the fit for Q(T) given above.

Once way around this problem is to create a carefully measured calibration curve for a given film mixture, using a given illumination intensity, on a given optical setup, etc. and use that curve as Q(T) above. This method allows accurate *absolute* temperature measurements to be made, but adds a great deal of difficulty to the FMI process. Problems with this process arise from small changes in equipment creating changes in light collection from a sample at a given temperature. For example, since the fluorescence intensity *decreases* as the film gets hotter and the film degrades (loses fluorescence intensity) under exposure to UV light, the sample will appear hotter after repeated imaging sequences. The degradation will require repeated removal and re-application of the film for accurate results.

Since in most applications, relative rather than absolute temperature measurements are needed, the image math can be simplified greatly with only a slight loss in accuracy. We need to stress that the slight accuracy change does not decrease the sensitivity, or the smallest change in temperature that can be resolved, of the technique.

To modify the process to allow for relative temperature changes, continue with the equation for the ratio of quantum efficiencies that we had before, take the natural logarithm, and plot the result for temperatures around room temperature. This gives us,

$$\ln(S_R(T)) = \ln\left[\frac{Q(T_H)}{Q(T_C)}\right] \propto \delta T .$$

This equation is plotted in figure 19 with the cold temperature, T_C, set at 28 ºC.

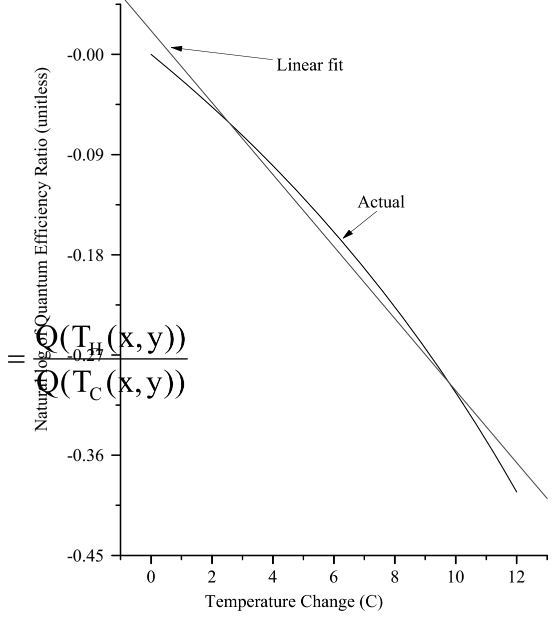


Figure 19 - Plot of quantum efficiency ratio around room temperature.

The linear fit in figure 19 can be represented by the equation,

$$y = 0.02132 - 0.03233 \cdot x.$$

The standard deviation in the slope is 6.9×10^{-4}, or about two percent. The slope of the linear fit is what we need for temperature conversion. Once we have the slope, the rest is easy. Simply divide the natural log of the quantum efficiency relation by the slope. The result is the relative temperature change at a given pixel location.

$$\delta T \approx \alpha^{-1} \cdot \ln\left[\frac{Q(T_H)}{Q(T_C)}\right] = \alpha^{-1} \cdot \ln\left[\frac{S_H(x,y)}{S_C(x,y)}\right].$$

By simply taking the natural log of the ratio of the light intensities from a hot and cold image, we can divide the result by a constant and have a relative temperature measurement. This is the method presented in the literature [25,26] for use in FMI.

The examples that have been used here are based on the quantum efficiency versus temperature from EuTTA in a solvent solution, as described in the previous section. For EuTTA combined in a polymer or glass matrix, the fluorescence quantum efficiency will have a similar temperature dependence, as this is dictated by the intermolecular energy transfer process, but they will be different. Published values for the slope of the linear fit in figure 19 for EuTTA in a dPMMA matrix are approximately -0.047 /ºC [26]. This number will be different for EuTTA in PMMA or spin-on glass, as well as for EuFOD in dPMMA. Whatever chemistry is chosen, a calibration curve as outlined above must be made.

As an example, if a camera with 16-bit gray scale resolution is used and assuming the published value of the logarithmic slope of -0.047 /ºC, the best possible thermal resolution would be:

$$\delta T \approx \left(\frac{1}{0.047}\right) \cdot \ln\left(\frac{65535}{65534}\right) = 0.324 \times 10^{-3} C .$$

Comparing this with an 8-bit gray scale camera:

$$\delta T \approx \left(\frac{1}{0.047}\right) \cdot \ln\left(\frac{255}{254}\right) = 83.6 \times 10^{-3} C .$$

This result will be referred to in the next section where hardware requirements are discussed.

Standard techniques can be used to reduce noise in the resulting thermal image. Image averaging on both the hot and cold images can be done prior to processing. Various techniques can be used after processing to enhance the thermal artifacts.

SYSTEM HARDWARE

There are essentially three main system components that are required for FMI. These components include a light source, a camera system, and an optical platform. This section discusses each of these areas and indicates differences between each of the possible choices.

Fluorescence Excitation Source. The light source used for fluorescence excitation has many possible choices. The absorption spectra for EuTTA (figure 13) indicates that we are interested in ultraviolet sources in the approximate range of 210 nm to 365 nm. Sources with wavelengths much greater than about 400 nm will require higher intensities to excite the fluorescence and thus are not practical.

Most of the FMI systems currently in operation use arc lamps as their excitation sources. The most common lamps for UV applications are mercury, xenon, and mercury/xenon types. Mercury bulbs have well known spectral peaks in the UV range and several that extend into the visible range. Xenon lamps have peaks in the upper end of the visible range, but have a broad continuum of output that extends usefully into the UV. Mercury/Xenon lamps combine the two spectra to create a broad range, general purpose light source. Stability of the light source, as evident from the last section, is one of the limiting factors in creating high temperature resolution images.

The other obvious choice for excitation source is a laser. Lasers differ from arc lamps in that all of their optical energy is in one narrow wavelength band rather than being spread over a continuum. Lasers are more efficient since all of the light output is within the wavelength range of interest. The problem with using lasers for FMI is the weighting of the cost versus performance. A laser will be significantly more expensive than a good arc lamp and will not add any significant performance advantage.

Camera Systems. The next step for system design is to choose a camera system. The only requirement for the camera system is to be able to quantitatively measure the small changes in fluorescence intensity caused by surface temperature variations. The better the camera is at performing this task, the better the thermal resolution of the will be. Existing systems, without exception, use slow-scan, CCD cameras. In this case, slow-scan refers to the frame rate at which data is read out of the CCD array. Television cameras adhere to the NTSC (or other) video standard where the CCD array in a CCD camera would be read at a rate of 30 frames per second. While this frame rate is good for television cameras, for quantitative analysis of the image content it is relatively poor. High quality TV cameras can approach 400 lines of information in about 500 fields with about 8-bits of dynamic range.

As we saw in the previous section, using a camera system with 8-bits of dynamic range would limit system sensitivity to roughly 0.4% change in quantum efficiency, or 1 part in 256. Slow-scan cameras are available with 12 to 16-bits of dynamic range and can thus image changes in intensity from 1 part in 4096 to 1 part in 65536 in an image of size ranging from 512 by 512 to more than 4096 by 4096 pixels. This translates into an order of magnitude gain in *possible* difference in temperature resolution.

Slow-scan cameras, since they do not adhere to TV standards, are designed to stare at a field of view and integrate for a variable length of time. For a situation where there is a very small amount of light being emitted, the camera can stare at the field of view for several minutes to several hours or until the detector becomes saturated. In contrast, when using TV cameras for low light situations, it is necessary to grab and add video frames together, which also adds noise. Image averaging can be used to help remove noise, but it does not boost signal.

Noise in collected images is a second factor that will limit temperature resolution. Slow-scan cameras are almost always either peltier or liquid nitrogen cooled.

Cooling helps to eliminate thermal generation of electron-hole pairs which can fill up CCD charge wells with noise instead of image signal. Peltier, or liquid/peltier, cooled systems generally operate at -39 ºC or so and, as a result, generate only several electrons per second of noise with a well capacity of several hundred thousand electrons. Liquid nitrogen cooled systems, by cooling to a much lower temperature, keep the noise down to several electrons per hour. With these systems, the readout electronics also add a small, but predictable amount of noise to the image, typically several electrons or tens of electrons per pixel.

In general, virtually any camera which can yield an image in digital format, either directly or by frame grabbing, can be used for FMI. The use of a slow-scan CCD camera will insure that the camera is not the limiting factor in system performance.

Optical Platform. Lastly, the optical platform to house the excitation source and camera must be decided upon. In order to electrically bias the sample, probe stations provide an obvious choice for an FMI system. Most systems used for probing fine geometries have optics boxes that have TV camera ports. Since they are used during probing, the lenses on these systems are extra-long working distance, but have to sacrifice some optical quality. Standard metallographic microscopes generally have superior optical systems, but suffer from short working distance objectives and limited facilities for electrical biasing of the sample. Depending on the application, whether it is primarily packaged ICs or wafer level analysis, the optical platform that best suits the most frequent use of the system should be used.

Early FMI systems input the excitation source through a UV grade fiber optic cable onto the sample at an oblique angle [xx, 4]. This does remove many headaches from the optical system, but tends to limit the amount of light that can be easily be sent to the sample, especially when using high magnification, shorter working distance lenses. Even on a probe station, a 50x lens will have a short enough working distance to complicate sample illumination. The use of a "through-the-lens" type of illumination removes the problems of sample illumination, but adds the problems encountered with non-UV transparent optics found in most microscopes. Generally, standard optical components offer transparency for light with wavelengths greater than about 370 nm. UV grade components are available because of the market for people doing fluorescence work, but the components, such as lenses, tend to be very limited in application and type, (e.g. magnification, working distance, numerical aperture, etc.).

In order to minimize the amount of time that UV light from the excitation source is focused on the fluorescent film, a shutter must be placed in the beam path between the UV source and the sample under examination. The shutter is usually synchronized with the camera shutter. As will be shown in the next sections, the UV light which excites the fluorescence also degrades the film which can generate noise in the thermal images. Although FMI can and has been done without this shutter, its inclusion is a simple way to control one of the dominant noise sources in an otherwise well optimized system.

The final system component needed is an interference filter to filter out all of the fluorescence except the dominant line at 612 nm as shown in figure 16. A filter with a bandwidth of about 2 - 4 nm will be sufficient.

The most common FMI system design is shown in figure 20. This has proven to be the easiest to assemble and best performing system design. The basic premise is to use light from a Hg arc lamp in the wavelength range from 365 - 390 nm as the excitation source. This, at first glance, is a contradiction of the desirable system design criteria for excitation source selection. The absorption spectra shown in figure 13 indicates that light at wavelengths longer than 365 nm will be stimulate the fluorescence but with much less efficiency than at shorter wavelengths. While this is true, sufficient fluorescence to perform FMI can be generated without special optical equipment needed for UV wavelengths. The most important components in the system design are the filters and beamsplitter. The excitation filter should be a short wavelength pass (SP) filter with a 390 nm cutoff wavelength. The beamsplitter needs to reflect the UV light passed by the excitation filter but allow the visible fluorescence to pass with minimal attenuation.

PHOTON SHOT NOISE AND SIGNAL AVERAGING

In absence of all other noise sources, the signal-to-noise ratio in high photon flux imaging applications is limited by photon shot noise. The statistics describing photon shot noise follows a Poisson distribution. In order to obtain any meaningful thermal information, it is crucial that we can separate the thermal signals from the photon shot noise. Figures 21 and 22 illustrate this point. Figure 21 is a histogram of pixel intensity for a ratio of two, consecutively acquired images. The histogram in figure 21 contains no thermal information meaning that the peak observed is mostly due to photon shot noise. As expected, the histogram shows a relatively good fit to the Poisson distribution, except on the right-hand side of the curve. The slight deviation is the result of ultraviolet bleaching that will be subject of discussion in the next section. Figure 22 shows a series of histograms of FMI thermal images for the same test structure biased at currents of 0 mA, 35 mA and 50 mA, respectively. In both the 35 mA and 50 mA histograms, an extra peak on the right hand side of the Poisson peak was observed. This peak contains thermal information associated with the heating of the metal test structure. As expected, this peak becomes more pronounced as the current in the test structure increases .

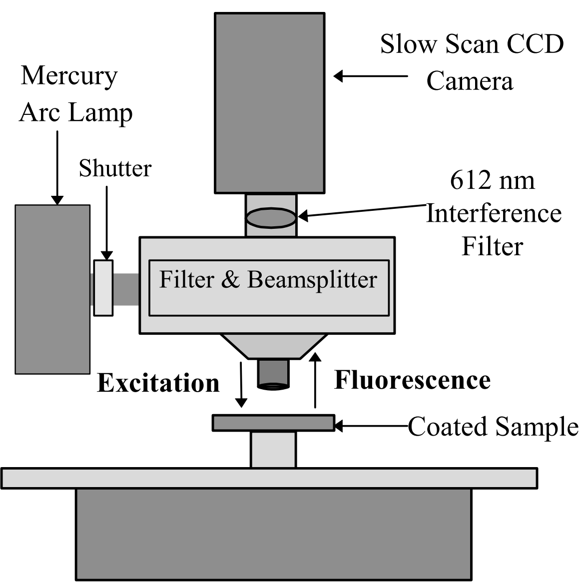


Figure 20 - Implementation of FMI using an arc lamp and coincident illumination.

Another interesting feature in figure 22 is the shift in the Poisson-peak position as the current increases. We attribute this shift to the increase in the background temperature surrounding the test structure. We can easily convert the position shift to the relative increase in temperature using the equation for δT derived earlier, with the slope (α) for the base-mixture film being obtained experimentally (see the section on the film dilution). For the histograms in figure 22, a background temperature increases of 0.6 ºC and 0.9 ºC are calculated from the 35 mA and 50 mA histograms, respectively. Interestingly, the relative increase in background temperature varies linearly with respect to the input power of the test structure (figure 18).

Since photon shot noise is due to the quantum nature of light, there is no way to eliminate it totally; however, it is possible to reduce its effects through signal averaging. figure 24 shows a comparison of two FMI images, one from a single cold/hot image pair and the other from an average of ten cold/hot image pairs. Both of the images were made with a current of 50 mA flowing through the metal test structure during the hot image exposure and a five second per frame exposure time on a UV-stabilized film. The image made from a single cold/hot image pair shows a grainy texture away from the test structure while the averaged image has a very smooth appearance. The effect of signal averaging can also be illustrated through the image histograms. Figure 25 shows histograms of a single cold/hot image pair and of an average of ten pairs. The Poisson peak from the averaged image shows a narrower width than the single image one, while there is virtually no change between the two thermal peaks. Clearly, signal averaging has improved the signal to noise of the FMI image by reducing the photon shot noise.

ULTRAVIOLET FILM BLEACHING

Inorganic-based films such as EuTTA are known to gradually lose their ability to fluoresce under UV illumination leading to decreasing fluorescent intensity with increasing UV exposure time. This gradual loss in fluorescent intensity is known as bleaching. Since bleaching is unavoidable in EuTTA films, we must characterize its behavior to identify methods to minimize its effect. Figure 26 shows the bleaching behavior of three EuTTA/PMMA films of various dilutions to continuous exposure to UV light. In all three films, fluorescent intensity decays rapidly initially and stabilizes after approximately 20 minutes of UV exposure. The diluted films, however, show a larger decrease in intensity before stabilization than the base mixture film.

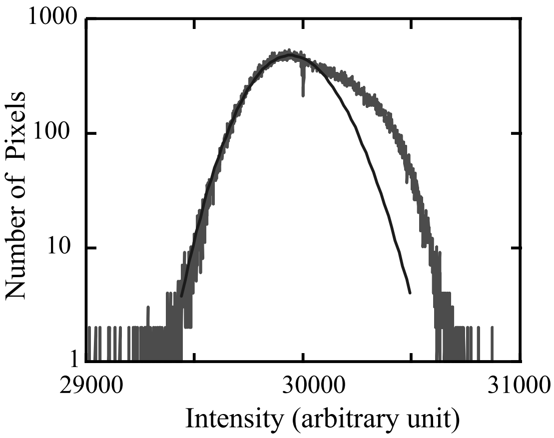


Figure 21 - Histogram of two successive fluorescence

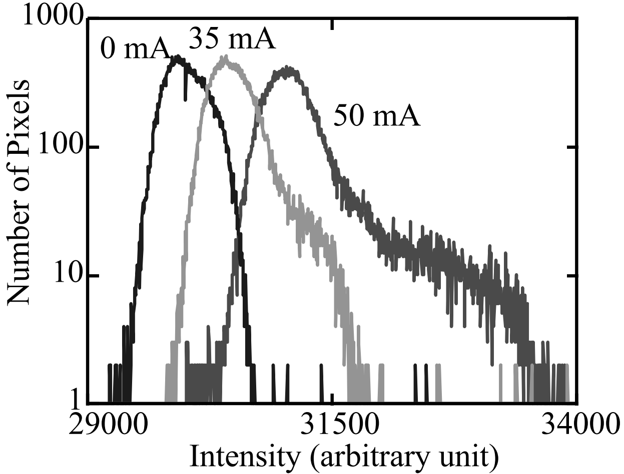


Figure 22 -Histograms of FMI images for images processed using the FMI image math sequence; a test structure biased with currents of 0 mA, 35 mA, and 50 mA, respectively

Figure 23 - Change in background temperature calculated from the position shift in the Poisson peak as a function of input power to the test structure.

In FMI, a thermal image is generated either by dividing the signals of a cold image by those of a subsequent hot image or vice versa. Ideally, the fluorescent signals of the hot and cold images should be identical in areas where no temperature changes occur. With bleaching, this ideal condition is not possible, resulting in the generation of unwanted, non-thermal signals which we usually refer to as spatial artifacts. To demonstrate the effect of UV bleaching on FMI image quality, a series of FMI images were taken after 20 seconds, 10 minutes and 20 minutes of UV exposure, respectively. In all three images, a metal test structure was biased with a current of 50 mA using a base-mixture film (figure 27). The spatial features are very pronounced in the left image (taken after 20 seconds of UV exposure) due to initial rapid decrease in fluorescent intensity and less noticeable in the center and right images (after 20 minutes of exposure) when the fluorescent intensity stabilizes.

To better understand how bleaching creates spatial artifacts, we made a series of histograms using the same image processing mathematics used for FMI but without applying any current through the test structure. The histograms were taken after 20 seconds, 10 minutes and 20 minutes of UV and are shown in figure 28. It is clear from figure 28 that bleaching causes a shift in the position of the Poisson peak and has the same appearance in the histogram as changing the background temperature as was previously described. The maximum shift occurs between the 20 second exposure and 10 minute exposure histograms, giving an apparent change in temperature of 0.60 ºC. The *apparent* thermal signals resulting from bleaching manifest themselves as the spatial artifacts observed in figure 27.

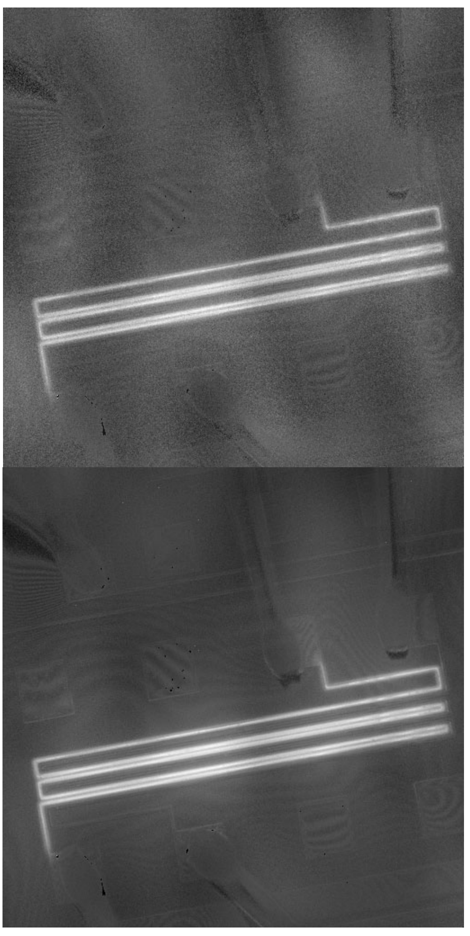


Figure 24 - A comparison of thermal images made with single cold/hot image pair(top) and an average of ten pairs (bottom).

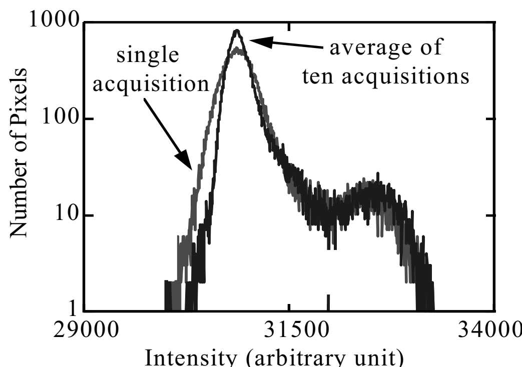


Figure 25 - Histograms of FMI images of a test structure, showing an improved signal-to-noise ratio after signal averaging

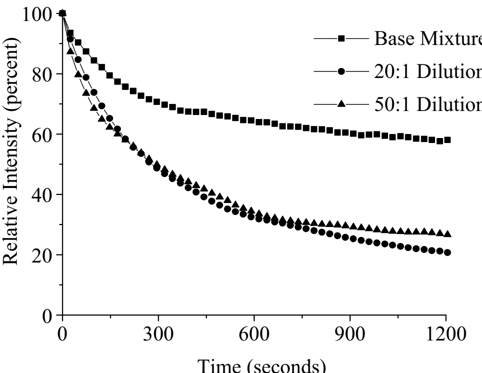


Figure 26 - Changes in fluorescent yield versus UV exposure for three different mixture dilutions.

Figure 27 - Fluorescent microthermal images illustrating the reduction in spatial features after UV film stabilization (top to bottom: t = 20 sec, t = 600 sec, t = 1200 sec).

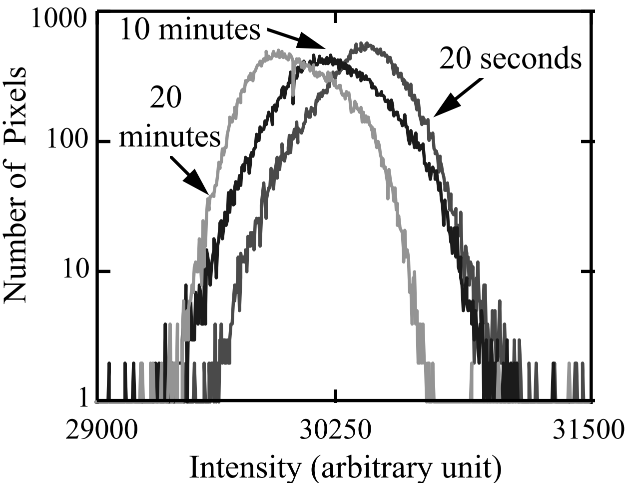


Figure 28 - Histograms of FMI images of an unbiased test structure, showing the effect of UV bleaching creating non-thermal signals.

EXAMPLE APPLICATION OF FMI:

To illustrate the application of the FMI technique to an integrated circuit failure, the following example is presented. The example is a 1-megabit SRAM. This SRAM failed functional test after an elevated voltage stress and had an elevated I_{DDQ} of 5 mA at 5 V. The I-V curve of this IC is shown in figure 29. The I-V curve in figure 29 strongly suggests that the failure of this SRAM involves more than just a simple ohmic short and most likely involves other processes such as those that cause light emission. In fact, three light-emitting areas were located on this IC as shown in figure 30. The box drawn on figure 30 highlights the area where FMI was used. A heat-generating area was also located on this SRAM with FMI is shown in figure 31. The locations of the light-emitting and heat-generating areas, did not coincide. Upon further SEM examination and visual inspection, the root cause of the failure was determined to be an embedded particle that was correctly localized using FMI and is shown in figure 32. The particle produced an ohmic short between two adjacent metal lines. By design, one of these metal lines was electrically connected to the gates of several transistors in series. The ohmic short caused these transistors to go into saturation, resulting in light emission.

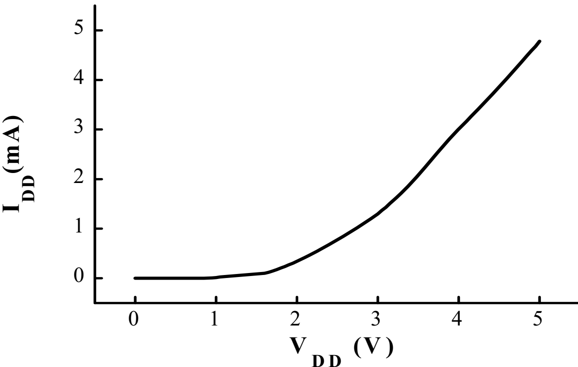


Figure 29 - I_{DDQ} versus V_{DD} for the failed SRAM.

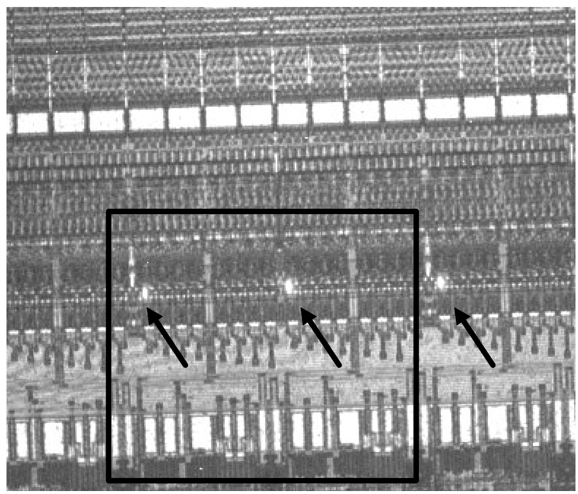


Figure 30 - overlay of light-emission and reflected light images showing light-emitting regions of the failed SRAM.

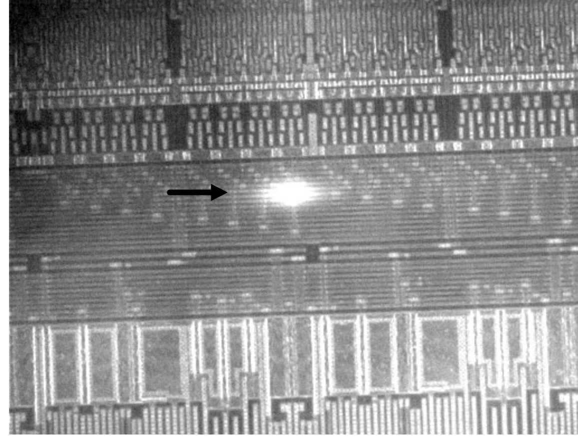


Figure 31 - Overlay of FMI and reflected light images showing the location of a hot spot.

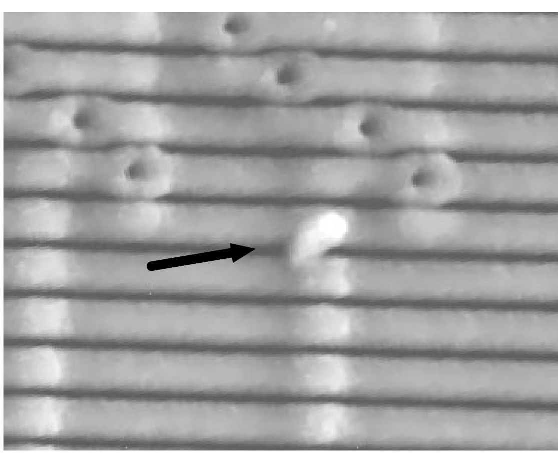


Figure 32 - SEM image showing an embedded particle that produced an ohmic short between two adjacent metal lines in the failed SRAM.

CONCLUSION

In this chapter, the subject of microthermal imaging for defect localization on integrated circuits was reviewed in detail. By introducing the physics that describes the spectral properties of heated bodies, the physical origins of the limitations of IR thermal techniques were studied. We found that the thermal resolution of IR systems can be quite good for IC applications, but the spatial resolution has become an order of magnitude too large. The use of liquid crystals for thermal defect detection was also reviewed giving insight into the reasons why this technique has been extremely popular with failure analysts for several decades. Finally, the fluorescent microthermal imaging (FMI) technique was reviewed in detail to help foster an understanding of the nature differences between FMI and the other, more common thermal analysis techniques.

In order to further demonstrate the relative differences between the three thermal analysis techniques which have been described, shown below in Figure 33 are images of the same structure at the same bias conditions shown side by side. The differences between the images support the theory described in this chapter. The infrared thermal image on the left shows good thermal resolution and is an absolute measure of the temperature of the surface of the structure under test. The image demonstrates the spatial resolution limitation that this technique must live with but, as it is the only non-contact thermal analysis technique and it's potential for backside IC analysis, infrared thermal analysis will be used for IC failure analysis for years to come.

The image in the center of Figure 33 demonstrates the spatial resolution or liquid crystals along with the technique's limited ability to map surface temperatures. By the nature of the materials used for liquid crystal technique, the technique can only indicate the areas in the

field of view which are above the nematic to isotropic state transition. Even with the imposed limitation of thermal mapping capabilities, liquid crystals will remain in the failure analysis tool repertoire because of their ease of use and excellent combination of thermal sensitivity and spatial resolution.

The final image on the right side of Figure 33 was made using the FMI technique. This image demonstrates the unique combination of thermal and spatial resolution that FMI has. The drawback to FMI is that the images are only relative temperature maps that can be difficult to accurately convert to absolute temperatures. While FMI has been used with good success for data collection for thermal modeling of transistor technologies which required absolute temperatures, this is really not the forte of FMI. FMI is intended to yield relative temperature maps of integrated circuit surfaces with very high resolution to quickly localize thermal defect signatures.

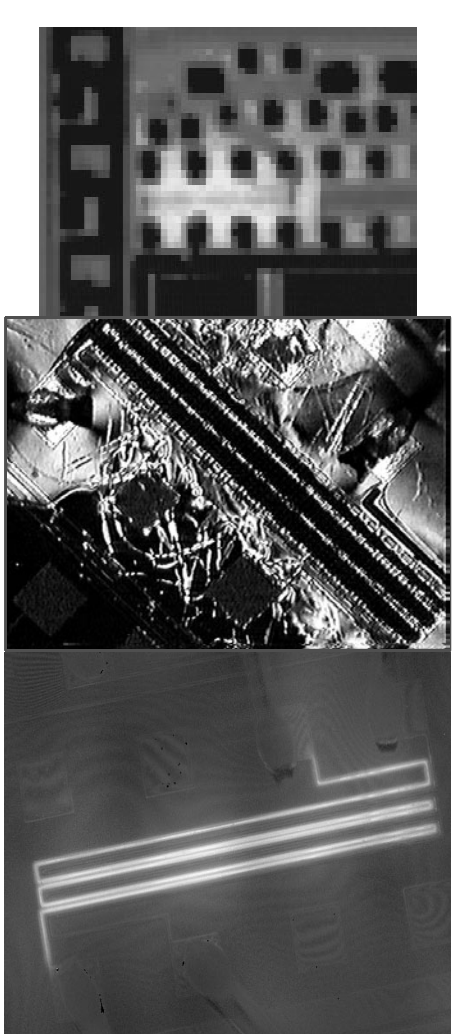


Figure 33 - Comparison of IR thermal imaging (top), liquid crystals (middle), and FMI (bottom) on same test structure.

Overall, the three techniques reviewed in this chapter are and will remain the mainstay of thermal defect detection tools for the foreseeable future. Other, competing techniques which offer promises of either better thermal or spatial resolution are either still in their infancy or have other limitations which must be overcome before they can challenge the techniques described here.

ACKNOWLEDGMENTS

The authors would like to thank Ronald P. Ruiz at the Jet Propulsion Laboratory for providing the infrared thermal images. This work was performed at Sandia National Laboratories and is supported by the U.S. Department of Energy under contract DE-AC04-94AL85000. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy.

REFERENCES

1. R. Eisberg, R. Resnick, *Quantum Physics*, John Wiley and Sons, 1974, ch. 1.
2. C. T. Elliott, D. Day, D. J. Wilson, "An Integrating Detector for Serial Scan Thermal Imaging", Infrared Physics, Vol. 22, pp. 31-42, 1982.
3. D. Pote, G. Thome, T. Guthrie, "An Overview of Infrared Thermal Imaging Techniques in the Reliability and Failure Analysis of Power Transistors", Proc. ISTFA, pp. 63-75, 1988.
4. G. J. Zissis, “Infrared Technology Fundamentals”, Optical Engineering, Vol. 15, no. 6, pp. 484-497, 1976.
5. P. Burgraaf, “IR Imaging: Microscopy and Thermography”, Semiconductor International, pp. 58-65, July, 1986.
6. E. Hecht, A. Zajac, *Optics*, Addison Wesley, 1974, ch. 10.
7. H. Stegmeyer (guest ed.), *Liquid Crystals*, Steinkopff: Darmstadt, Springer: New York, 1994.
8. P. J. Collings and J. S. May (editors), *Handbook of Liquid Crystals*, Oxford University Press, New York, 1997.
9. I. -C. Khoo, *Liquid Crystals: Physical Properties and Nonlinear Phenomena*, John Wiley and Sons, New York, 1995.
10. A. Boller, M. Cereghetti, M. Schadt, and H. Scherrer, Synthesis and some Physical Properties of Phenylpyrimidines, Mol. Cryst. Liq. Cryst., 1977, Vol. 42, pp. 215-231.
11. D. Demus, L. Richter, C. -E. Rürup, H. Sackmann, and H. Schubert, Synthesis and Liquid Crystalline Properties of 4,4’-Disubstituted Biphenyls, Journal of

Physics Colloque C1, supplement no. 3, Vol. 36, March 1975, pp. C1-349 - C1-354.
12. L. A. Karamysheva, E. I. Kovshev, A. I. Pavluchenko, K. V. Roitman, V. V. Titov, S. I. Torgova, and M. F. Grebenkin, New Heterocyclic Liquid Crystalline Compounds, Mol. Cryst. Liq. Cryst., 1981, Vol. 67, pp. 241-252.
13. G. W. Gray and D. G. McDonnell, Some Cholesteric Derivatives of S-(+)-4-(2'-Methylbutyl) Phenol, Mol. Cryst. Liq. Cryst., 1978, Vol. 48, pp. 37-52.
14. J. Hiatt, A Method of Detecting Hot Spots on Semiconductors Using Liquid Crystals, *Proc. IRPS*, 1981, pp. 130-133.
15. G. D. Dixon, Cholesteric Liquid Crystals in Non-Destructive Testing, *Materials Evaluation*, June 1977, pp. 51-55.
16. A. Geol, A. Gray, Liquid Crystal Technique as a Failure Analysis Tool, *Proc. IRPS*, 1980, p. 115.
17. J. L. Fergason, Liquid Crystals in Nondestructive Testing, Applied Optics, Vol. 7, No. 9, September 1968, pp. 1729-1737.
18. M. Laurlente and J. L. Fergason, Liquid Crystals Plot the Hot Spots, Electronic Design, Vol. 19, September 1967, pp. 71-88.
19. G. L. Hill and J. R. Agness, Practical Liquid Crystal Applications in Failure Analysis, *Proc. ISTFA*, 1983, pp. 73-79.
20. S. Ferrier, Thermal and Optical Enhancements to Liquid Crystal Hot Spot Detection Methods, *Proc. ISTFA*, 1997, pp. 57-62.
21. D. Burgess, and P. Tan, Improved Sensitivity for Hot Spot Detection Using Liquid Crystals, *Proc. IRPS*, 1984, pp. 119-121.
22. D. Burgess and O. D. Trapp, Advanced Liquid Crystal for Improved Hot Spot Detection Sensitivity, *Proc. ISTFA*, 1992, pp. 341-343.
23. E. M. Flueren, A Very Sensitive, Simple Analysis Technique Using Nematic Liquid Crystals, *Proc. IRPS*, 1983, pp. 148-149.
24. A. Csendes, V. Székely, and M. Rencz, Thermal Mapping with Liquid Crystal Method, Microelectronic Engineering, vol. 31, 1996, pp. 281-290.
25. P. Kolodner, J. A. Tyson, "Microscopic fluorescent imaging of surface temperature profiles with 0.01 ºC resolution", Appl. Phys. Lett. 40, 782 (1982).
26. P. Kolodner, J. A. Tyson, "Remote thermal imaging with 0.7 mm spatial resolution using temperature dependent fluorescent thin films", Appl. Phys. Lett. 42, 117 (1983).
27. D. L. Barton, "Fluorescent microthermographic imaging," Proceedings of the 20th ISTFA, 1994, pp. 87-95.
28. D. L. Barton and P. Tangyunyong, "Fluorescent Microthermal Imaging - Theory and Methodology for Achieving High Thermal resolution Images", Microelectronic Engineering, volume 31, Numbers 1-4, February 1996, pp. 271 - 280. (Proceedings of the Fifth European Conference on Electron and Optical Beam Testing of Electronic Devices, August 27 - 30, 1995, Wuppertal, Germany).
29. P. Tangyunyong, D. L. Barton, "Photon Statistics, Film Preparation, and Characterization in Fluorescent Microthermographic Imaging", Proc. ISTFA, 1995, pp.79 - 86.
30. H. Winston, O. J. Marsh, C. K. Suzuki, C. L. Telk, "Fluorescence of Europium henoylfrifluoroacetonate.Evaluation of Laser Threshold Parameters", J. Chem. Phys., vol. 39, no. 2, pp. 267-270, July, 1963.
31. M. Bhaumik, "Quenching and Temperature Dependence of Fluorescence in Rare-Earth Chelates", J. Chem. Phys., Vol. 40, (3711), 1964.
32. G. Crosby, R. Whan, R. Alire, "Intramolecular Energy Transfer in Rare Earth Chelates. Role of the Triplet State", J. Chem. Phys., Vol. 34, (743), 1961.
33. E. Bowen, J. Sahu, "The Effect of Temperature on fluorescence of Solutions", J. Phys. Chem., Vol. 63 (4), 1959.
34. P. Kolodner, A. Katzir, N. Hartsough, "Noncontact surface temperature measurement during reactive-ion etching using fluorescent polymer films ilms", *Appl. Phys. Lett.* 42 (8), 15 April 19

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Thermal Failure Analysis by IR Lock-in Thermography

O. Breitenstein
Max Planck Institute of Microstructure Physics, Halle, Germany

C. Schmidt, F. Altmann
Fraunhofer Institute for Mechanics of Materials, Halle, Germany

D. Karg
Thermosensorik GmbH, Erlangen, Germany

Abstract

Thermal infrared (IR) microscopy has experienced a decisive technical improvement by the application of Lock-in Thermography (LIT), which is commercially available for failure analysis by different vendors. Due to its averaging nature, this technique allows the detection of local heat sources at the surface of a few µW corresponding to a local temperature modulation of a few µK. Thus it outperforms other thermal imaging methods, like liquid crystal imaging, fluorescent microthermal imaging, Raman IR-thermography, and steady-state IR thermography, by 2 to 3 orders of magnitude. Thereby, LIT allows to extend the application field of thermal imaging for failure analysis drastically. The emissivity contrast, which obscures the thermal contrast in conventional IR thermography, can be avoided in LIT by displaying the phase image or the 0°/-90° image in a 2-phase measurement. Due to its dynamic nature, lateral heat diffusion (blurring) is considerably reduced in LIT compared to steady-state techniques, depending on the chosen lock-in frequency. The usefulness of Lock-in Thermography is given mainly by the fact that it allows a though relatively coarse (3-5 µm) but very sensitive localization of any leakage current or other local heat source in an IC with a very high success rate without any preparation expense. LIT is also applicable for backside inspection and for detecting sub-surface heat sources. Its spatial resolution can be improved down to 1 µm by applying a solid immersion lens. By using LIT some faults can be localized which are not visible in OBIRCH or light emission microscopy. In this contribution the technique of microscopic IR Lock-in Thermography is described, the basic principles of the interpretation of the results are reviewed, and some typical results illustrating the application of this technique are introduced.

Lock-in Thermography - technique

Lock-in Thermography was invented in 1984 [1] and has been used extensively in non-destructive testing for "looking below the surface" of solid objects [2]. Meanwhile it represents also a standard technique for investigating shunting phenomena in solar cell research and for failure analysis in ICs [3, 4]. Lock-in Thermography means that the power dissipated in the object under investigation is periodically amplitude-modulated, the resulting surface temperature modulation is imaged by a thermo camera running with a certain frame rate f_{fr}, and that the generated IR images are digitally processed according to the lock-in principle. Thus, the effect of LIT is the same as if each pixel of the IR image would be connected with a 2-phase lock-in amplifier. In principle, Lock-in Thermography can also be applied with 1-phase detection, but then some very useful display options like phase imaging are not available. The two primary results of 2-

phase LIT are the image of the in-phase signal $S^{0°}(x,y)$ and that of the out-of-phase (or quadrature) signal $S^{-90°}(x,y)$. In LIT the -90° signal is used instead of the +90° one, since the latter is essentially negative [3]. From these two signals the image of the phase-independent amplitude $A(x,y)$ and the phase image $\Phi(x,y)$ of the surface temperature modulation can easily be derived:

$$A(x,y) = \sqrt{S^{0°}(x,y)^2 + S^{-90°}(x,y)^2} \qquad (1)$$

$$\Phi(x,y) = \arctan\left(\frac{-S^{-90°}(x,y)}{S^{0°}(x,y)}\right) \qquad (2)$$

Note that for Φ the quadrant-correct arctan function has to be used, hence if $S^{0°}(x,y)$ is negative, 180° have to be subtracted from the pure arctan value. Both the 0° and the -90° images are proportional to the magnitude of the temperature modulation, multiplied by the IR-emissivity ε. Therefore in the amplitude image the contrast of a heat source is proportional to its dissipated power, multiplied by ε. The phase image, on the other hand, relies on the quotient of the 0° and the -90° image, hence for isolated heat sources it is independent of the power of the heat source and also of ε. Thus the phase signal is inherently emissivity-corrected. In fact, the phase image is a measure of the time delay of the surface temperature modulation referred to the power modulation, which is indeed independent of the magnitude of the heat source, as long as we have no superposition of the temperature fields of neighbouring heat sources. For the application of LIT in thermal failure analysis this property implies a kind of "dynamic compression" for the phase signal, so that local heat sources with different powers are displayed with a similar signal height. It will be shown below that these properties greatly simplify the interpretation of the results. Note, however, that in metallised regions having a low value of ε also the signal-to-noise ratio of both primary images decreases, hence metallised regions may appear more noisy in the phase image than non-metallised ones. This can be avoided by blackening the surface, which can be done e.g. by colloidal bismuth which is relatively easy to deposit and can easily be removed in an ultrasonic bath [5, 6]. Instead of IR thermography also fluorescent microthermal imaging (FMI) or Moiré thermal imaging can be applied in lock-in mode [7, 8]. These techniques are sometimes called "stabilized" instead of lock-in [8] to make clear that in these dynamic techniques the result can be obtained already after a few lock-in periods or even after one period and not after reaching thermal equilibrium. Note that FMI may provide sub-micron spatial resolution, but it does not allow backside inspection with good spatial resolution, it needs a foreign layer at the surface, and it is about a factor of 10 less sensitive than IR LIT [7]. Stabilized Moiré thermal imaging [8] is a pure backside inspection technique. Its spatial resolution is only in the order of the bulk thickness, and its sensitivity is also well below that of IR LIT.

As an alternative to the phase image in 2-phase LIT, also the 0°/-90° signal ratio may be displayed [5]. Since in microscopic regions the -90° signal has a very poor spatial resolution, this image can be taken as a local measure of the IR emissivity ε. Therefore the 0°/-90° image is also inherently emissivity-corrected, just as the phase image. In contrast to the latter, however, the 0°/-90° image shows a better spatial resolution, the signal height reflects the power of local heat sources, and it even can be used directly for correcting the thermal blurring by applying image deconvolution techniques [5].

Fig. 1 shows how the image data are evaluated in LIT to obtain the primary signals $S^{0°}$ (or $S^{-90°}$) for each pixel. In simplest case the supply voltage V_{CC} of the device is switched periodically on and off to a well-defined value for realizing the periodic heat modulation. If

V_{CC} is not allowed to be switched completely off, it also can be square-wave modulated between two values. Sine-modulation of V_{CC} would be also possible, but it gives no advantage to square-modulation and is even not useful, since then the device is not investigated at a well-defined supply voltage. At least 4 frames per lock-in period have to be evaluated, since according to the Shannon sampling theorem at least two samples per period are necessary, which holds for each phase position (0° and -90°). The actual 2-phase lock-in correlation consists of multiplying the information of each incoming image by weighting factors in two logical channels and summing up these products in two separate frame storages. In one channel the weighting factors are approximating a sine- and in the other one a (-cosine)-function [3] (harmonic correlation). Other correlation functions (e.g. square-shaped) are also feasible, but for them the splitting into an amplitude and a phase image according to (1) and (2) is not correct anymore. Fig. 1 (a) illustrates the general case of harmonic correlation for the in-phase signal for the case of 12 frames per lock-in period. Here the line symbolizes the basic harmonic of the in-phase part of the local temperature modulation, which is detected by this procedure, and the squares symbolize the sampling moments and the magnitudes of the weighting factors. Here, in the first half-period the weighting factors are positive and in the second half-period they are negative. After each period (dashed line) the procedure repeats. For obtaining the -90° signal a (-cosine)-function has to be used in Fig. 1 (a) for the weighting factors. For obtaining the highest possible detection sensitivity, the IR camera is always running at its highest possible frame rate f_{fr} and the lock-in frequency $f_{lock-in}$ is adjusted by choosing an appropriate number of frames per period. Since the sum of all weighting factors is zero, this correlation provides a perfect suppression of the steady-state (topography) image, which is basically governed by the emissivity contrast. Nevertheless, the primary LIT images $S^{0°}$ and $S^{-90°}$ still contain the emissivity contrast $\varepsilon(x,y)$ as a factor.

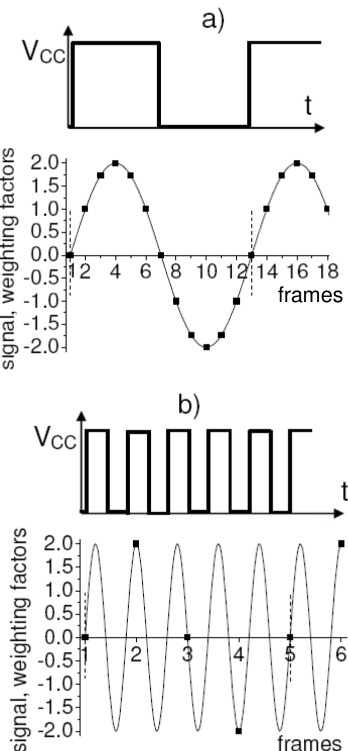


Fig. 1: Scheme of the Lock-in Thermography correlation (0° signal), (a) conventional, (b) undersampling

The limitation of the conventional LIT correlation shown in Fig. 1 (a) is that, since we need at least 4 frames per lock-in period, the maximum possible lock-in frequency is $f_{lock-in} = f_{fr}/4$, which is 25 Hz for a typical frame rate of 100 Hz. If several local heat sources are lying close to each other, it may be necessary to further increase $f_{lock-in}$ for further reducing thermal blurring. This limitation can be overcome by applying the "undersampling" timing strategy sketched in Fig. 1 (b). Here only one sample is taken in each lock-in period (or even every few periods), but its phase position varies from period to period. The evaluation occurs as for conventional LIT. By applying this technique lock-in frequencies in

the kHz-range may be realized even for full frame LIT imaging. Note that this reduction of blurring reduces the sensitivity, since the magnitude of the LIT signal always reduces with increasing lock-in frequency [3].

The effective spatial resolution

If Lock-in Thermography is applied for failure analysis of integrated circuits, a decisive limitation is its limited spatial resolution. For a point-like heat source lying at the surface of a device, the surface temperature modulation field at a small lateral distance r from the source reduces with $1/r$, independent of the thermal diffusion length [3]. Hence, even for low lock-in frequencies point-like heat sources at the surface can be localized sharply, only the extension of the halo around depends on the lock-in frequency. Lateral heat spreading is more disturbing for spatially extended heat sources [3]. This heat spreading can be considered even as an advantage since it guarantees that spatially small heat sources cannot be overlooked in a low-magnification survey image. Note, however, that the critical point is not the resolution of the lock-in images, which anyway may appear more or less blurred due to lateral heat spreading or because the actual heat source may lay at a certain depth below the surface. Even in such a case, if a heat source is point-like, the position of its center can usually be estimated up to an accuracy of 1 pixel by finding the center of gravity of the blurred spot. The main problem with spatial resolution is that the operator still needs to be able to navigate on the surface of the IC! Today the layout patterns may be so small that no details can be resolved anymore with a conventional microscope objective in the mid-IR range. Therefore the challenge for improving the spatial resolution is to get a meaningful topography image, which enables an orientation on the surface. Only then local peaks in the lock-in images can be related reliably to the layout of the IC.

The resolution Δx of any optical system is physically limited by diffraction, which is governed by the wavelength λ of the radiation used for imaging. According to the so-called Sparrow Criterion [9], the optical resolution is limited to:

$$\Delta x = \frac{0.5\,\lambda}{n\sin(\theta)} \tag{3}$$

Here θ is the half-angle of the light-cone to the objective and n is the refractive index of the medium surrounding the sample. The product $n^*\sin(\theta)$ is also called "numerical aperture" (NA). For a given magnification factor of the objective, the brightness of an image (hence the signal-to-noise ratio of the measurement) increases with the square of $\sin(\theta)$, since the number of photons reaching the detector increases with an increasing solid angle used by the objective. Even for high brilliance microscope objectives, θ can hardly be larger than 30 to 45° for technical reasons, therefore $\sin(\theta)$ is at best between 0.5 and 0.7. Hence, in air (n = 1) the optical resolution can be only slightly better than the wavelength λ used for imaging. Therefore midwave IR cameras working in the 3-5 µm range show a better spatial resolution than longwave cameras working at 8-12 µm. Unfortunately, for samples being close to room temperature, in the mid range the light intensity exponentially increases with wavelength. So the dominant part of the light is concentrated close to 5 µm and only a negligible part appears at 3 µm. Therefore, for a good microscope objective with NA = 0.7 (+/-45° light acceptance angle), according to (3) the diffraction-limited spatial resolution limit for λ = 5 µm is Δx = 3.6 µm. This limit can be improved to close to 1 µm by applying a solid immersion lens (SIL) [10, 11]. This is in simplest case a half-bowl made from silicon or germanium which is placed with its plane bottom face on the plane surface of the device.

This lens provides an additional optical magnification at least equal to the diffraction index n of the material, which is about 3.5 for Si and 4 for Ge. Since here the object is "immersed" in the SIL-material, the wavelength of the light is smaller by n there and the NA increases by this factor. Thus, if the slit between the surface and the SIL is well below 1 µm, surface structures with dimensions down to 1 µm can be seen, and for point-like heat sources at the surface the spots in the LIT image are getting correspondingly smaller. SILs can be applied both at the front or at the back surface of devices, if they are accurately flat polished.

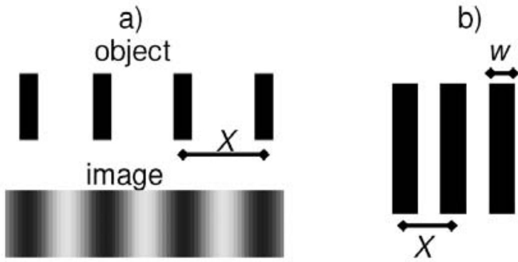


Fig. 2: (a) Line object and its image close to the resolution limit, (b) USAF pattern showing the spatial resolution X and the "line and space distance" *w*

Note that the term "diffraction-limited spatial resolution" refers to the minimum distance of two neighboring small spots or lines (i.e. a "line pair"), which can be separated from each other. If more than two parallel lines are used, the right line of the left line pair coincides with the left line of the right pair, hence e.g. "288 line pairs/mm" actually means 288 lines/mm, corresponding to a line distance (center-to-center) of $X = 3.47$ µm. If such a periodic arrangement is imaged with an objective close to its resolution limit, the brightness is sine-modulated with a spatial frequency of $f = 1/X$, which is the basic spatial harmonic, see Fig. 2 (a). All higher spatial frequencies are suppressed since we have assumed that these details are below the diffraction-limited spatial resolution. The decisive point is that the spatial frequency f is only dependent on the center-to-center distance of the lines X but not on the line width w or the line distance $X - w$. Only the intensity of the basic spatial harmonic compared to higher harmonics depends on w. It is highest if $w = X/2$ holds, hence if the lines have a distance equal to their width. This is realized e.g. in the elements of the well-known USAF resolution target, one of them sketched in Fig. 2 (b). Here 3 lines with a center-to-center distance X are displayed with a distance of $X/2$ in between. For 288 line pairs/mm this distance is about 1.74 µm. If these lines can be observed separately, the thereby proven spatial resolution is not $X/2$, but it is the center-to-center line distance X! Some authors consider the "line and space distance" $w = X/2$ as a measure of the resolution [11], but this overestimates the spatial resolution by a factor of 2.

Another interesting question is which magnification factor M of the lens must be used for making use of the diffraction-limited spatial resolution. One might think that a lens leading to an object pixel distance of Δx according to (3) should be sufficient ($M > 4.2\times$ for a pitch size of 15 µm and $\Delta x = 3.6$ µm), but this is wrong. According to Shannon's sampling theorem at least two samples are necessary per spatial wavelength in order to have at least one pixel at the maximum and one at the minimum of a periodic contrast, see Fig. 2 (a). Thus, for a pitch size of 15 µm and $\Delta x = 3.6$ µm the lens must have a magnification of at least $M = 8.4\times$ for reaching the diffraction-limited spatial resolution. An even higher magnification factor may still improve the visual image quality, but for LIT it also degrades the signal-to noise ratio, which reduces with $1/M^2$ [9].

Typical results

Fig. 3 shows the amplitude image of a Hall sensor circuit (a), the corresponding phase image (b), the 0°/-90° image (c, detail), and the

power distribution (d) numerically deconvoluted from (c).

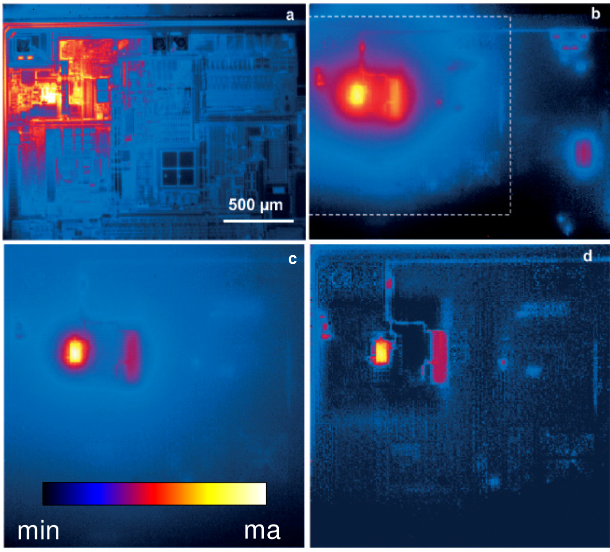


Fig. 3: Amplitude image (a), phase image (b), ε-corrected 0° image (c; 0°/-90° image from the region indicated in b), and power distribution (d), numerically deconvoluted from (c), of a hall sensor circuit; supply voltage pulsed with 22 Hz

All details visible in Fig. 3 are due to the normal operation of this circuit. The selected region displayed in (c) and (d) is indicated in (b). The measurement was performed at a lock-in frequency of 22 Hz within a few minutes. We see that the amplitude image (a) is indeed strongly affected by the emissivity contrast caused by the metallization pattern. Note that due to the lock-in technique the steady-state IR image (topography image) is already perfectly suppressed even in the amplitude image (a). Hence, the bright regions outside of the heat source positions, modulated by the local emissivity contrast, are caused by the inevitable lateral heat conduction-induced halo of the temperature modulation around the heat sources. From the amplitude image (a) it is hardly possible to judge which of the bright regions are real heat sources and which are regions of high emissivity. In the phase image (b) and the 0°/-90° image (c), however, this emissivity contrast is indeed perfectly removed. Only the signal-to-noise ratio is degraded in regions of a low emissivity. The differences between the phase image (b) and the 0°/-90° one (c) are clearly visible in Fig. 3. The phase image (b) shows a stronger halo around the heat sources, and it displays heat sources of different intensity in a comparable brightness ("dynamic compression" feature). On the other hand, the 0°/-90° image (c) shows a lower blurring and displays heat sources of different power with different brightness. The power distribution (d) which was deconvoluted from (c) reveals many more details than the original images.

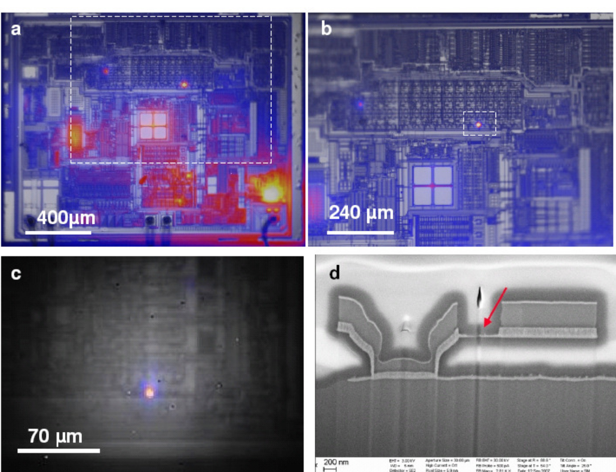


Fig. 4: (a) Survey image of the whole die (objective 2.5×), (b) detail image of the region framed in (a) (objective 5×), (c) Detail image of the region framed in (b) (objective 5× with SIL), (d) SEM image of a cross section through the fault region

In the next example overlay images of the topography image (in grey) with the actual LIT images (in color) of a faulty device are shown. Fig. 4 (a) was taken at a lock-in frequency of 10 Hz by using a 2.5× objective which is able to display the whole die. By comparing the heat transportation of the defective device to a reference, the point-like power source in the framed region could be identified as defect-

related thermal emission. In order to allow a better localization of the thermal emission to the single device components (b), a 5× objective was used and the lock-in frequency was increased to 25 Hz. Finally, (c) shows the region around the fault imaged at the same frequency through a silicon solid immersion lens (SIL), which further increased the magnification by a factor of 3.5. The localization of the fault allowed a focused ion beam (FIB) preparation of a cross-section specimen for a scanning electron microscopy (SEM) investigation, which is shown in (d). This image shows residues of a TiN barrier layer which were not completely etched away (arrow) and finally lead to a short. This case study shows the usefulness of LIT at IC investigations due to the fact that even defects which are orders of magnitude smaller than the resulting spatial resolution are detectable.

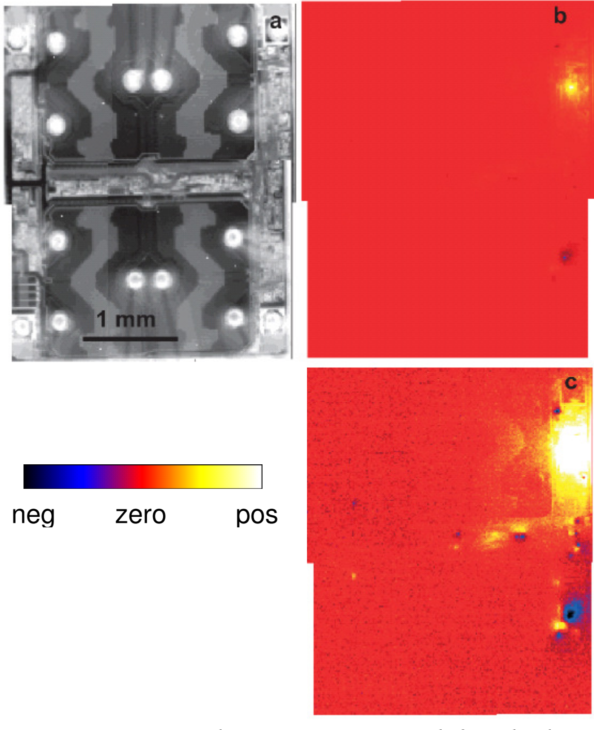


Fig. 5: Topography image (a) and fixed phase lock-in thermogram: (b) and (c) of an IC with permanently applied supply voltage and triggered control input. (c) is a contrast-enhanced presentation of (b)

In the next example in Fig. 5, which was an intact step motor controller, the supply voltage was permanently applied and the lock-in trigger was fed to a control input. In this arrangement permanently existing heat sources in the device, which are not affected by the trigger signal, do not appear in the lock-in thermogram. Only heat sources affected by the trigger signal are detected here. In a single phase image (approx. -45°) it can be distinguished whether heat is generated in the on-state or in the off-state of the trigger (positive or negative response). In Fig. 5 (b) and (c) complementary acting heat sources are made visible by bright and dark spots. In a similar way certain activities in a logic device can be switched on or off, synchronized to the lock-in correlation, which would allow an easy functional in-circuit test of complex logical devices based on lock-in thermography.

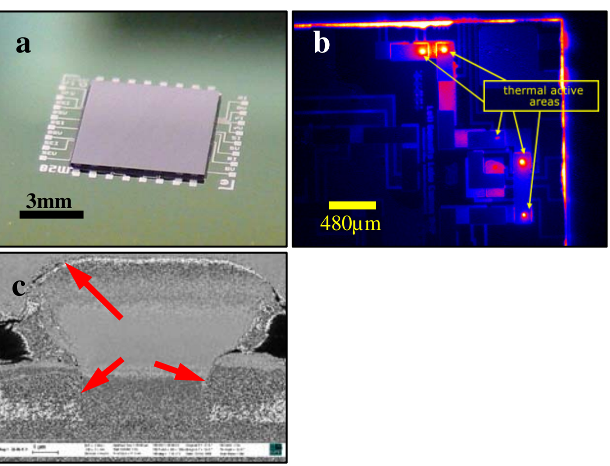


Fig. 6: Picture of the flip chip device (a), LIT result shows thermally active areas (b), SEM investigations of a contact cross section reveals additional insulator layer (red arrows in c)

In addition, Lock-in Thermography is also applicable for backside inspection. Figure 6 shows the investigation of a flip chip device with local high ohmic contacts. Due to the fact that silicon is IR transparent (dependent on its doping concentration) it is possible to

investigate the inner structure non-destructively. A mechanical cross section and SEM investigations were able to determine the root cause – an additional insulator layer infiltration in the contact area [12].

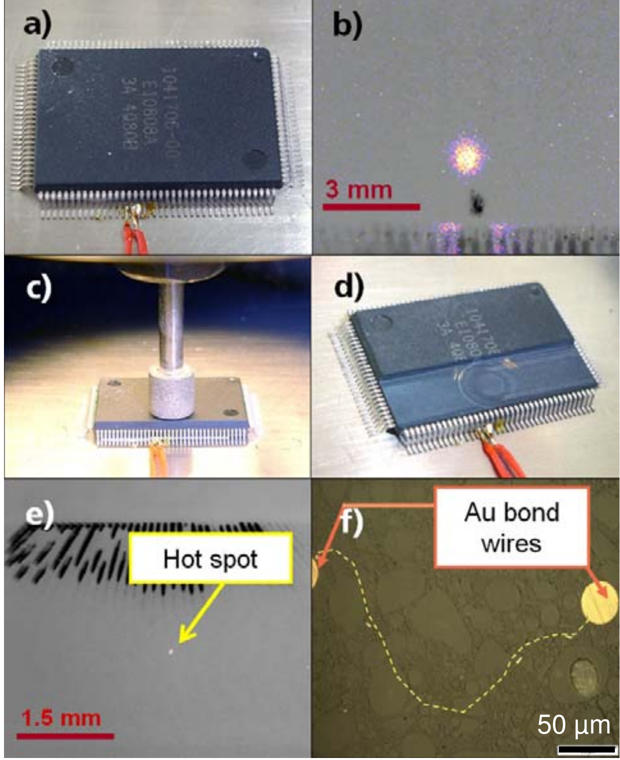


Fig. 7: Picture of the fully packaged device (a) and first LIT measurements (b), mechanical grinding of package material (c, d) allow defect allocation between bond wires (e), cross section show path of a metal splinter creating a short (f, light microscope image)

Besides the advantages of using Lock-in Thermography in case of single chip devices especially the opportunity of investigating fully packaged devices is an important advantage in comparison to other failure detection methods. Note that thermal waves penetrate even optically opaque materials like mould compounds. Due to the damping of thermal waves, using a lock-in frequency of 1 to 25 Hz allows the observation of hot spots through 100-400 µm package material with a lateral spot size of only a few tens to hundred µm. Even if the spatial resolution is not good enough to allow the detection of a single defect electrical component, this is extremely helpful especially in case of defective bond areas or stacked die devices. It gives the opportunity of deciding the next preparation steps and, most important, preserves the overall electrical functionality of the device. Such investigations cannot be done by optical methods like OBIRCH or light emission microscopy. Figure 7 gives an example for the usefulness of this allocation method where a hot spot due to a metal splinter occurred in the bonding area. After first hot spot allocation (b), the package material was grinded mechanically and a second LIT measurement (e) allocated the root cause between to bond wires which was proven by a mechanical cross section.

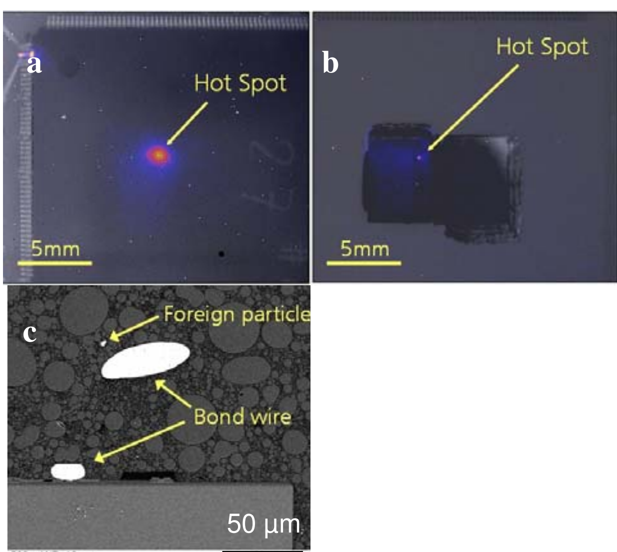


Fig. 8: LIT result at a fully packaged stacked die device (a) and after opening using chemical etching (b), SEM investigations of the cross section show foreign Al particle (c)

In case of the investigations at a stacked die device given in Fig. 8 LIT results reveal a hot spot not in the bond but in the chip area. Therefore, an opening using chemical etching was possible allowing a better spatial

resolution. Again, cross sectioning and SEM/ EDX investigations were applied determining an Al-splinter as root cause. Furthermore, a 3D defect localization at fully packaged multi chip modules by analyzing the phase quantitatively is part of research, these days.

Summary and outlook

Lock-in IR thermography, which was already established as a standard technique in non-destructive testing and solar cell research, has now successfully entered the field of IC failure analysis. In comparison with other thermal failure analysis techniques (liquid crystal microscopy, fluorescence microthermal imaging; also stabilized, stabilized moiré imaging, Schlieren imaging, Raman IR thermography, conventional IR microscopy) it shows the following advantages:

1. The thermal sensitivity may be below 100 µK (depending on the measurement time and the NA of the IR lens), which is 2-3 orders of magnitude better than that of previous techniques. Thus, LIT can be applied to investigate also weak heat sources, which have been non-accessible by thermal methods previously.

2. It requires no foreign layer at the surface, hence it also is used on a wafer scale.

3. The measurement procedure and the interpretation of the results are very straightforward: One needs no temperature stabilization and no shading of the setup, there is no degradation, and heat sources appear simply as bright spots in the images.

4. It also is used as a backside analysis technique. In case of highly doped bulk silicon LIT can also be applied, similar to investigations through opaque materials like mould compound.

5. Due to the dynamic nature of the measurement, lateral heat conduction is considerably reduced. Hence, the effective spatial resolution is improved compared to steady-state thermal imaging techniques.

6. Non-destructive defect localization at fully packaged devices is possible, though with degraded spatial resolution, which preserves the overall functionality and allows a better evaluation of the following preparation steps and failure analysis methods.

One of the general problems of conventional IR microscopy, the emissivity contrast artefact, can be overcome in 2-phase Lock-in Thermography by displaying the phase image or the 0°/-90° image, which are both inherently emissivity-corrected. The spatial resolution, which is diffraction-limited for mid-range IR cameras to about 3.6 µm, can be improved to nearly 1 µm by applying a solid immersion lens.

Acknowledgements

The authors like to thank Achim Lindner (Micronas GmbH), Jürgen Schulz (Melexis GmbH) and Volkmar Gottschalk (ELMOS Semiconductors) for providing samples and J.-M. Wagner for critically reading the manuscript. The work was supported by the European project Pidea "Full control".

References

1. P.K. Kuo, T. Ahmed, H. Jin, and R.L. Thomas, *Phase-Locked Image Acquisition in Thermography, SPIE* **1004**, 41-45 (1988)
2. X.P.V. Maldague, *Theory and Practice of Infrared Technology for Nondestructive Testing,* Wiley, New York (2001)
3. O. Breitenstein and M. Langenkamp, *Lock-in Thermography – Basics and Use for Functional Diagnostics of Electronic Components,* Springer, Berlin, Heidelberg (2003), new edition in 2010
4. O. Breitenstein, J.P. Rakotoniaina, F. Altmann, J. Schulz, and G. Linse, *Fault Localization and Functional Testing of ICs by Lock-in Thermography, Proc. 28th ISTFA,* 29-36 (2002)

5. O. Breitenstein, J.P. Rakotoniaina, F. Altmann, T. Riediger, M. Gradhand, *New Developments in IR Lock-in Thermography, Proc. 30th ISTFA,* 595-599 (2004)
6. M. Gradhand, O. Breitenstein, *Preparation of non conducting infrared-absorbing thin films*, Review of Scientific Instruments **76**, 053702 (2005)
7. F. Altmann, T. Riediger, O. Breitenstein, *Fault Localization of ICs by Lock-in Fluorescent Microthermal Imaging (Lock-in FMI)*, Poster at *30th ISTFA (2004)*; see also O. Breitenstein et al., *Lock-in Thermography – A New FA Tool*, tutorial at *ISTFA 2005*
8. J.B. Colvin, *Moiré Stabilized Thermal Imaging*, *Proc. 12th IPFA* (2005), 163-166
9. E. Hecht: *Optics*. Addison-Wesley, San Francisco, CA (2002)
10. O. Breitenstein, F. Altmann, T. Riediger, D. Karg, V. Gottschalk, *Use of a Solid Immersion Lens for Thermal IR Imaging, Proc. 32nd ISTFA*, 382-388 (2006)
11. H. Suzuki, K. Koshikawa, T. Kuroda, T. Ishizuka, F.P. Gyue, F.P. Yuan, W.J. Ji, *Improvement of Performance for Higher Magnification Thermal Imaging*, *Proc. 16th IPFA 2009,* 489 - 492 (2009)
12. C. Schmidt, M. Simon, F. Altmann, *Failure analysis of stacked-die devices by combining non-destructive localization and target preparation methods, Proc. 35nd ISTFA*, 319-323 (2009)

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor
Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Principles of Thermal Laser Stimulation Techniques

F. Beaudoin,
THALES Microelectronics, Toulouse, France

R. Desplats, P. Perdu
CNES, Toulouse, France.

C. Boit
TUB Berlin University of Technology, Berlin, Germany

Abstract

Thermal Laser Stimulation techniques such as OBIRCH, TIVA and SEI have proven to be valuable in FA laboratories. Initially intended for interconnect defect localization, these techniques are also used to detect and localize polysilicon shorts, ESD defects, soft gate oxide shorts and even on-state transistors.

The goals of this article are to shed light on the basics of Thermal Laser Stimulation and to assist failure analysts in determining how to best implement this method in their FA laboratories.

Introduction

In most microelectronic foundries, Thermal Laser Stimulation (TLS) techniques, such as OBIRCH [1], TIVA [2] and SEI [2], have been successfully implemented in their failure analysis flow. They have become essential defect localization tools along with backside emission microscopy [3]. Their combination is not accidental. It is the result of years of shared exchanges within user groups. Indeed, the use of several metal layers and the appearance of new packaging technologies challenged traditional front-side techniques such as liquid crystals requiring the introduction of new techniques. The search for new solutions applicable from the backside of the die led the failure analysis community to consider NIR optical techniques.

The extension of emission microscopy to backside failure analysis was quite straightforward [4,5]. However, the application of an NIR laser beam to thermally stimulate ICs was relatively new. Early investigations showed promising results in locating interconnect defects causing current leakage type faults [6,7,8]. However little was known of the potential industrialization of this method into a failure analysis laboratory.

In this article we will first start with the basics of Thermal Laser Stimulation. Then we will present an overview of Thermal Laser Stimulation systems, techniques and setups. Thereafter several models aimed at understanding Thermal Laser Stimulation effects on metallic lines, junctions, transistors and complex ICs will be discussed. Finally, we will present relevant case studies requiring a good understanding of the Thermal Laser Stimulation method to interpret failures.

Thermal Laser Stimulation Basics

The Thermal Laser Stimulation method uses a near-infrared (NIR) laser beam to thermally stimulate ICs. The laser energy is chosen below the silicon bandgap energy (1.1eV) to avoid generation of photocurrents in the IC substrate. Those would mask the Thermal Laser Stimulation effect [6]. In addition, the silicon substrate is relatively transparent at such NIR wavelengths. The laser beam absorption depends on the doping type and level [9,10]. Thermal Laser Stimulation is therefore also applicable to backside failure analysis. For doped silicon substrates, the die backside may be thinned appropriately (e.g. 100μm) to minimize laser power losses due to absorption [11,12].

The small portion of the absorbed NIR laser beam principally heats IC metallic elements, and to a lesser extent, polysilicon elements and highly doped substrate areas. It modifies the electrical properties of the heated medium, namely its resistivity. This change of resistivity alters the current flow (or voltage). These electrical changes inside the device may be monitored at the power supply nodes. Correlating IC power consumption changes with the laser spot location provides precise localization of thermally sensitive areas such as resistive defects.

TLS induced IC power consumption variations can be classified as follows:

i) resistance variation (OBIRCH, TIVA)
ii) thermoelectric energy conversion or Seebeck effect (SEI)

Resistance variation
A temperature increase in the heated elements induces a resistivity variation, modeled as follows:

$$\Delta\rho = \rho_0 \alpha_{TCR} (T - T_0) \quad (1)$$

where ρ_0 is the resistivity, α_{TCR} is the thermal coefficient of resistance, and ΔT is the temperature variation.

When the IC is biased, the thermally induced resistance variation is monitored through the device voltage or current consumption depending on the biasing method. This is illustrated by the following equations:

$$\Delta V = \Delta R_{IC} I_S \quad (2)$$

$$\Delta I = -(\Delta R_{IC} / R_{IC}^2) V_S \quad (3)$$

where ΔR_{IC} is the resistance variation and R_{IC} is the IC resistance, I_S is the current source and V_S the voltage source.

Seebeck effect
The Seebeck effect is based on the generation of an electromotive force by temperature gradients at junctions composed of two different materials (e.g. aluminum-tungsten) [13]. The Seebeck voltage induced when one side of a thermocouple is heated is given by:

$$\Delta V = (Q_1 - Q_2)\Delta T = Q_{1-2}\Delta T \quad (4)$$

where Q_1, Q_2 are the respective thermoelectric power of materials and $Q_{1\text{-}2}$ the relative thermoelectric power difference of the two materials.

When no bias is applied to the IC, the measured voltage variation is directly related to the Seebeck effect. This effect is also present when the IC is biased, but its contribution is often negligible compared to the Thermal Laser Stimulation signal due to the resistance variation [2].

Thermal Laser Stimulation Systems, Techniques and Setups

The Thermal Laser Stimulation method is usually implemented on commercial NIR laser-scanning microscopes. The laser wavelength is typically 1300nm or 1340nm with laser powers up to 500mW.

Thermal Laser Stimulation systems also integrate the following elements:

i) Voltage or current source to bias the IC.
ii) Low noise amplification scheme to amplify small power consumption variations.
iii) Laser scanning synchronization system to synchronize the laser beam position with power consumption variations.
iv) Imaging software to visualize power consumption variations overlaid on the IC image.

One of the key components of Thermal Laser Stimulation systems is the amplification scheme used to pick up the extremely small voltage or current variations caused by laser heating. Commercial systems are available, but specifically designed amplifiers have a bandwidth adaptable to the laser scanning speed with high adjustable gain.

Several Thermal Laser Stimulation techniques have been developed and appropriately named to reflect their novelty. They are presented in Table 1. The difference between them lies in device biasing and signal detection schemes.

Table 1: Thermal Laser Stimulation techniques

Thermal Laser Stimulation techniques	Source	Amplifier
CC-OBIRCH [14] TIVA [2]	Current	Voltage
OBIRCH [1] IR-OBIRCH [6]	Voltage	Current
TBIP [15] XIVA [16]	Voltage	Voltage
SEI [2]	Current or no bias	Voltage

The CC-OBIRCH (Constant Current Optical Beam Induced Resistance Change) and TIVA (Thermally-Induced Voltage Alteration) techniques use a constant current source with a voltage amplifier mounted in parallel to the IC as can be seen in Figure 1. The same setup, with or without the current source, is used for the SEI (Seebeck Effect Imaging) technique.

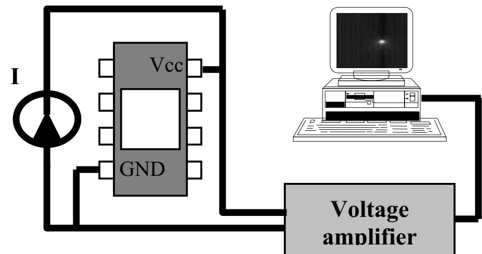


Figure 1 TIVA, CC-OBIRCH and SEI techniques

For the OBIRCH (Optical Beam Induced Resistance Change) and IR-OBIRCH (Infrared Optical Beam Induced Resistance Change) techniques, a constant voltage source is used. A current amplifier is mounted in series with the IC as shown in Figure 2.

In the TBIP (Thermal Beam Induced Phenomenon) and XIVA (Externally Induced Voltage Alterations) techniques, the current amplifier used in the OBIRCH configuration is replaced with a specific amplification scheme composed essentially of an inductor and a voltage amplifier.

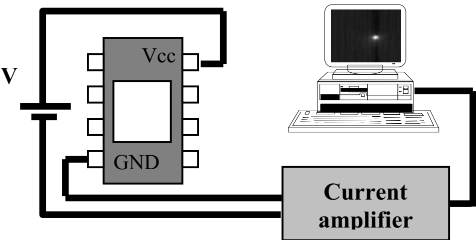


Figure 2 OBIRCH and IR-OBIRCH techniques

The sensitivity of the Thermal Laser Stimulation method strongly depends on the noise related to both the biasing and detection scheme. The biasing source should be as noise free as possible. The constant current or constant voltage source noise detected by the amplifier is more or less significant depending on the IC resistance. For failed ICs characterized by a high resistance, it is more appropriate to use a voltage source. For low resistance leakage paths, the use of a current source may reduce the noise associated with the biasing scheme. The sensitivity can also be improved by using a lock-in detection scheme. This option is proposed on some commercial Thermal Laser Stimulation tools.

Thermal Laser Stimulation Models

In this section we present several models aimed at understanding the Thermal Laser Stimulation signal when the NIR laser beam is scanned over a single metallic line.

Thermal simulation model
To gain insight into the heating dynamics, a 3D finite element model of a 1µmx0.5µm aluminum line embedded in silicon dioxide was studied. A schematic diagram of the modeled structure is shown in Figure 3. The material properties of thin film Al, bulk SiO_2 and Si, given in Table 2, were assumed to be temperature independent. A detailed description of the model is given in [17].

The Thermal Laser Stimulation effect was dynamically modeled for the two following cases:

- Longitudinal case when the laser beam is scanned perpendicular to the metal line.
- Transversal case when the laser beam is scanned parallel to the metal line.

Thermal simulations were performed for a laser beam power of 100mW, a laser beam radius of 0.65µm and an initial temperature of 25°C. Both transversal and longitudinal cases were resolved for a fast scanning speed of 1.23m/s and slow scanning speed of 0.00768m/s. These respectively correspond to a 1024x1024 frame scanning time of 2s when a 5x objective is used and to a scanning time of 16s when an objective of 100x is used.

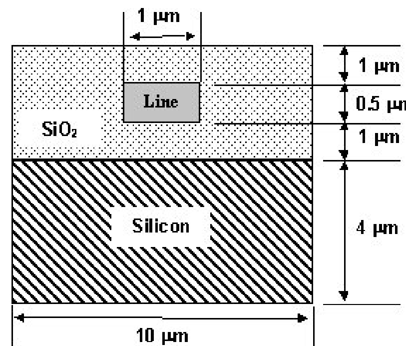


Figure 3 Schematic diagram of the modeled structure.

Table 2: Structural parameters [18,19,20].

	Al	SiO_2	Si
Density (ρ) (Kg/m³)	2702	2190	233 0
Specific heat (c) (J Kg⁻¹ K⁻¹)	896	1400	703
Conductivity (K) (W m⁻¹ K⁻¹)	246	1.4	141
Absorption coefficient (α) (cm⁻¹)	$1.3x10^6$	0	NA
Reflectivity (R) @ 1.3 µm	97%	0	NA

The model confirmed that the metal line reaches the permanent regime at the slowest scanning speed. The temperature profile for the longitudinal case is shown in Figure 4. The peak temperature is obtained in less than 10µs.

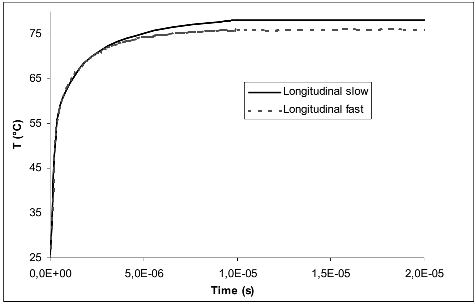


Figure 4 Temperature profile as a function of time for the longitudinal case at both scanning speeds.

The temperature profile along the metal line at the hottest spot for all cases is given in Figure 5. A permanent regime temperature of 79.9°C was obtained. However this maximum temperature was not reached for the fastest scanning speed. At low magnification, defects may be undetected if the scanning speed is not sufficiently slow. For instance, a 1024x1024 frame scanning time higher than 4s is required when a 5x objective is used.

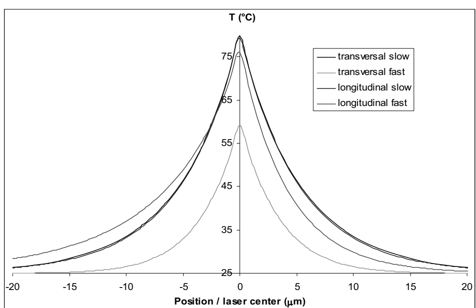


Figure 5 Temperature profile along the metal line for all the considered cases. Absolute temperature values are calculated for a 100 mW laser power.

Simulations calculated a thermal spreading limited to approximately 30 µm and a separation of less than 0.1µm between the hottest spot and the laser beam.

Finally, it was found that the temperature increases linearly with the laser power. Given sufficient time for the permanent regime to be established, the peak temperature variation induced by laser heating in a 1µm x 0.5µm aluminum line is:

$$\Delta T_{max} = 0.55\,^{o}C/mW \quad (5)$$

Resistance variation

Resistance variation of the aluminum was calculated in the permanent regime. Resistance change, ΔR, can be calculated from the average temperature along the metal line at a given time ($T_{avg}(t)$) according to the following analytical model:

$$\Delta R(t) = \rho_0 \alpha_{TCR} (L/S)(T_{avg}(t) - T_0) \quad (6)$$

where ρ_0 is the metal resistivity, α_{TCR} is the metal temperature coefficient of resistance, L the length of the line, S its cross-section and T_0 the initial temperature.

Equation 6 points out the inversely proportional relationship of the resistance change to the metal line section. The resistance change of smaller elements is favored compared to larger ones.

In the case of the 1µmx0.5µm aluminum line, maximum resistance change in the permanent regime is given by:

$$\Delta R_{max} = 1.7x10^{-3}\ \Omega/mW \quad (7)$$

This calculation was done using a resistivity (ρ_0) of $4.8x10^{-6}$Ωcm and a temperature coefficient of resistance (α_{TCR}) of $3x10^{-3}K^{-1}$. These values are typical of an Al technology CMOS IC process [21].

The voltage or current variations due to the resistance variation are calculated with equations 2 or 3. The reduction or increase in the power consumption variation is directly related to the metal temperature coefficient of resistance. Therefore, the polarity of the Thermal Laser Stimulation signal provides direct information on the resistive defect nature. For an aluminum line submitted to a constant current, a voltage increase is expected. This is not the case for polysilicon as its temperature coefficient of resistance depends on its nature such as its doping level.

Electromotive force generation

The Seebeck voltage induced at an interface such as that shown in Figure 6 can be calculated from equations 4 and 5. The Seebeck coefficients for different materials are given in Table 3.

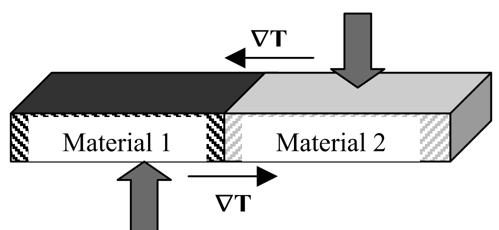


Figure 6 Thermal gradients induced at a thermocouple as a function of the laser beam position.

Table 3. Seebeck coefficients for different materials [18,22].

	Q (µV/°C)
Al	-3.4*
W	3.6*
Al / n+ Si	287**
Al / p+ Si	-202**

**measured relative to copper,*
***relative thermoelectric power for a 10^{18} cm^{-3} doping.*

The polarity of the Seebeck voltage depends on the laser beam position relative to the junction. For example, when the laser beam crosses the junction, the polarity of the Seebeck voltage is inverted since the induced temperature gradient is in the opposite direction.

Upon Thermal Laser Stimulation, the temperature can be uniformly distributed throughout the whole volume of the two materials composing the junction due to heat spreading. This thermal spreading can attenuate the Seebeck voltage. However, an interface defect such as a hole can create an imbalance between the two opposite Seebeck voltages, therefore allowing its localization.

The Seebeck effect is best observed when no bias is applied to the IC. Indeed, the power generation variations are generally small compared to induced resistance variations. Another advantage of non biased Thermal Laser Stimulation is the very small electrical power involved, which is in the order of picoWatts. Defects, as well as the device under test, are therefore unaltered, which can be very important for later physical analysis.

IC effect model

In complex ICs, the Thermal Laser Stimulation signal also depends on the IC's response to the resistance variation and/or the induced electromotive force. The IC effect was simulated with Spice for a resistive defect short-circuiting the PMOS of a CMOS inverter as shown in Figure 7. Thermal Laser Stimulation of the resistive defect was simulated by varying the resistance value.

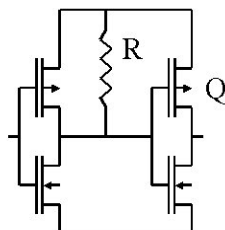


Figure 7 CMOS inverter modeled in SPICE.

The total current variation as a function of the resistance change is plotted in Figure 8.

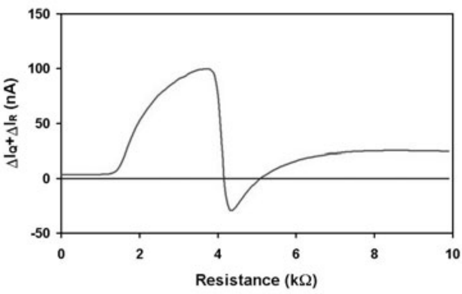


Figure 8 Total current variations in the CMOS inverter shown in Figure 7.

The magnitude and sign of the current variation strongly depend on the resistance value. The physical model predicts a current decrease upon a resistance increase when a voltage bias is applied to the CMOS inverter. This is the case only for a small range of resistance values.

Transistor heating model

The FET transistors are also affected upon Thermal Laser Stimulation. This is understood through the current-voltage characteristics of an idealized MOSFET in the saturated region [23]:

$$I_{Dsat} \cong \frac{Z\mu\varepsilon_{ox}}{2dL}\left(V_G - V_T\right)^2 \quad (8)$$

where Z and L are respectively the channel width and length, d is the thin oxide thickness, μ is the mobility, $\varepsilon_{o\xi}$ is the oxide dielectric constant, V_G is the gate voltage and V_T is the threshold voltage.

A temperature increase in the channel results in a mobility change as well as a voltage threshold change [24].

The channel region of the FET is not directly heated by the laser beam, but rather indirectly through heat diffusion from neighboring regions as indicated in Figure 9. The FET device regions efficient in converting NIR light into heat are:

- Heavily doped regions such as source and drain
- Polysilicon gates
- Buried layers

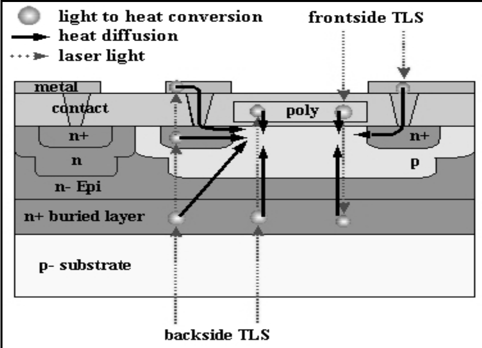


Figure 9 Light transmission and heat diffusion for front-side and backside Thermal Laser Stimulation.

Oversize FET test structures were studied in order to understand their behavior under Thermal Laser Stimulation. These devices, built in a BiCMOS technology, are characterized by a 10x10µm^2 gate area and a 15nm gate oxide. The Thermal Laser Stimulation results obtained from the front and backside of the FET when biased in the saturated region are presented in Figure 10. A high Thermal Laser Stimulation signal is observed at the gate level due to indirect heating of the channel by heat diffusion though the thin gate oxide. The source and drain region as well as the interconnects were also

found to affect the drain current. The higher thermal efficiency obtained from the backside can be attributed to heat generation at the buried layer.

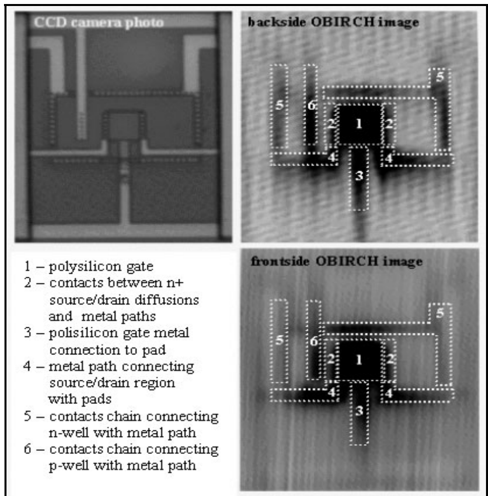


Figure 10 Front-side and backside Thermal Laser Stimulation images obtained on oversized FET test structures.

Finally, the maximum current change was measured for different drain to source voltages under Thermal Laser Stimulation. As can be seen in Figure 11, the output characteristics indicate a current change proportional to the FET drain current. Therefore the Thermal Laser Stimulation method is able to localize conducting FET transistors. The magnitude of the Thermal Laser Stimulation signal depends on the biasing condition of the FET.

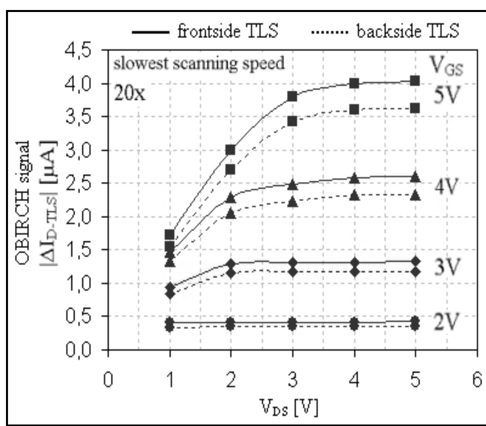


Figure 11 Front-side and backside Thermal Laser Stimulation induced current variation as a function of VDS for various values of VGS.

Discussion

The knowledge gained from the thermal laser stimulation models is manifold. The Thermal Laser Stimulation method:

- Is relatively fast. The thermal permanent regime is reached in less than 10µs.
- Can precisely localize current paths and defects. The hottest temperature occurs at the laser spot center.
- Is applicable to metallic, polysilicon and highly doped silicon structures.
- Is efficient in localizing short-circuit type defects when a current flows through them. The resistance variation of smaller elements is higher than for larger ones.
- Provides information on the nature of the material. The polarity of the Thermal Laser Stimulation signal depends on the polarity of the temperature coefficient of resistance.
- Allows the localization of material discontinuities, which can be defect induced, through the Seebeck effect.
- Can detect and localize conducting FET transistors.

The Thermal Laser Stimulation signal is not always directly interpretable in complex ICs. The magnitude and polarity of the TLS signal is strongly influenced by the location and the value of the resistive defect. In addition, junctions as well as conducting FETs may be visualized in the Thermal Laser Stimulation image.

Furthermore, the high number of metal levels used in today's ICs can result in the delocalization of the thermally sensitive area relative to the resistive defect location. This is the case if the resistive defect is not directly heated by the laser beam (e.g. due to metallization masking), but through heat diffusion from the nearest heated structure.

Case Studies

Five case studies are presented in this Section to illustrate the application of Thermal Laser Stimulation in a FA laboratory. The obtained TLS images can be understood with the models presented above.

Case study #1: Front-side TLS of shorted metal line test structures

Several shorted metal line test structures composed of two closely interlaced aluminum combs were investigated with the Thermal Laser Stimulation method. In order to localize the short with TLS, the less resistive path was biased with a constant voltage source to obtain a current in the order of a few tens of microamps.

The TLS image of the current variations and its superimposition on the laser reflected image are shown in Figure 12. The faint white signal area (indicated by the arrow) corresponds to the biased Al

lines. Indeed, a current decrease, caused by a resistance increase of the Al upon TLS, is represented in white in our setup. The higher black signal (circled) located at the edge of the current path corresponds to the short-circuit defect. Indeed, it suggests a smaller structure which cannot be attributed to aluminum. Precise localization of the suspected short was done only with Thermal Laser Stimulation using a 100X objective. Physical analysis revealed a small TiN short between the two Al combs.

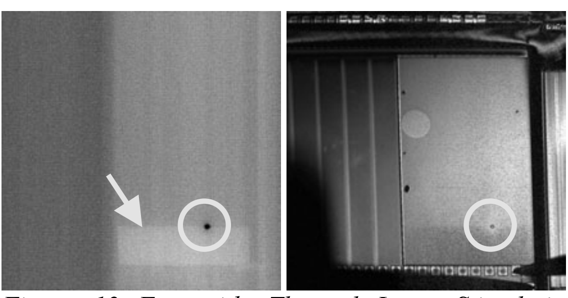


Figure 12 Front-side Thermal Laser Stimulation image obtained under 5X magnification (left) and superimposition on the laser reflected image (right).

Case study #2: Front-side TLS of EOS/ESD damaged bipolar transistors

This case study presents the application of the Seebeck effect to localize resistive vias in via chain test structures. The Thermal Laser Stimulation images obtained on non-biased via chain test structures are shown in Figure 13. Imbalances in the Seebeck signals at certain vias are observed, such as the one indicated by an arrow in Figure 13. Physical analysis of those vias revealed holes on the edge of the via's foot. Localization of the defective vias could not be obtained with TLS under biased conditions. The power variations due to resistance variations dominate in front of the Seebeck effect.

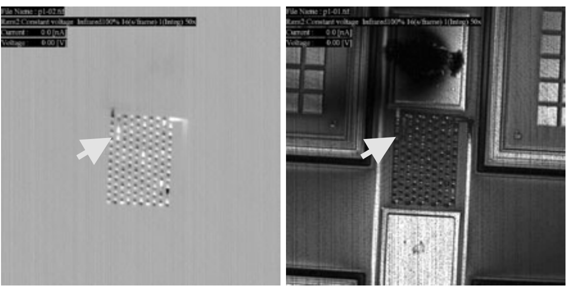


Figure 13 Front-side non-biased Thermal Laser Stimulation image obtained under 50X magnification (left) and superimposition on the laser reflected image (right).

Case study #3: Front-side TLS of EOS/ESD damaged bipolar transistors

The Seebeck effect can also be used to non-destructively localize EOS/ESD damages in silicon. Failed commercial RF bipolar transistors used in space applications were investigated. All transistor junctions were short-circuited.

Thermal Laser Stimulation images obtained under no bias by monitoring the emitter-base junction are shown in Figure 14. A Seebeck induced signal is observed at one finger of the bipolar transistor. The black and white signal indicates a polarity change in the Seebeck voltage. These signals indicate a thermocouple composed of the metal contact on one side, and the damaged silicon junction on the other side (circled). Physical analysis revealed EOS/ESD type silicon damage.

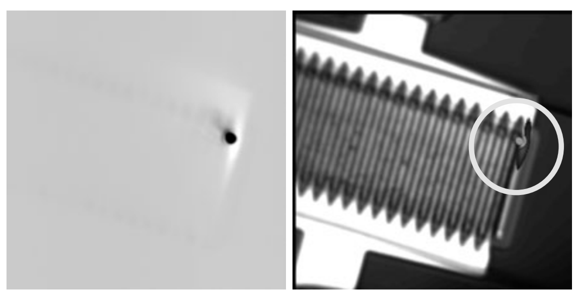


Figure 14 Front-side Thermal Laser Stimulation image of the voltage variation acquired for a 200X magnification (left) and superimposition on the laser reflected image (right).

Case study #4: Front-side and Backside TLS on failed ICs

Several failed ICs of 0.25µm technology (6 metal layers) were investigated from the front-side as well as from the backside. These failed ICs were characterized by a leakage current of 5mA under 1V bias. In the backside approach, the silicon substrate was thinned to approximately 100µm. A thermally sensitive area could be observed in both front and backside Thermal Laser Stimulation. The backside Thermal Laser Stimulation results are shown in Figure 15.

Layout investigation highlighted a 5µm position shift between the front and backside thermally sensitive area location. Physical analysis in that area revealed shorts at the polysilicon layer level (pointed out by an arrow in Figure 16). The location of this polysilicon short corresponds to the location of the backside thermally sensitive area. The short defect is masked by several metallization layers. Therefore, upon front-side TLS, the short defect is not directly heated, but indirectly through upper metallization levels which are connected to this polysilicon layer level. The sensitive area localization upon front-side TLS is therefore the directly heated metal line which is the closest to the polysilicon short, in this case 5microns away.

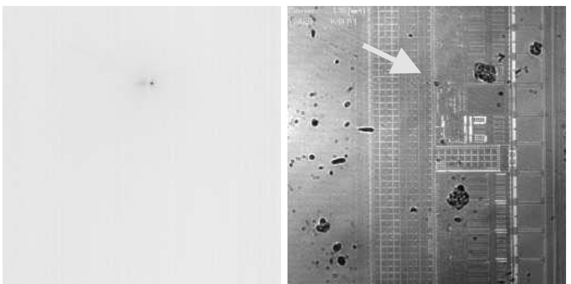


Figure 15 Backside Thermal laser stimulation image of the current variation acquired with a 5x objective (left) and superimposition on the laser reflected image (right).

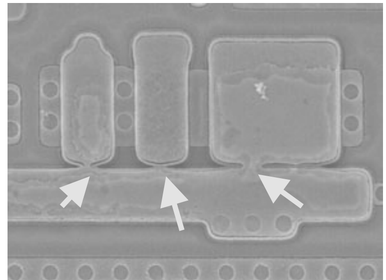


Figure 16 SEM image of the polysilicon shorts.

Case study #5: Front-side TLS on failed ICs

In the last case study the sensitivity of the Thermal Laser Stimulation method for the localization of highly resistive type shorts is illustrated. The failed devices presented a 2mA leakage current under a 3V bias. The obtained Thermal Laser Stimulation image and its superposition on the laser reflected image are shown in Figure 17. The current path created by the resistive defect can be observed. Both extremities of this current path are connected to a different CMOS inverter stage. The transistor sensitivity to Thermal Laser Stimulation indicates that these inverters are conducting.

A magnified TLS image obtained at the location of the polarity change within the current path (pointed out by an arrow) is shown in Figure 18. Physical investigation at this location revealed a metallic parasitic via, indicated by an arrow in Figure 18. The observed polarity change at the defect site is explained by the IC effect. The inverters at each side of the two short-circuited metal lines are polarized in a different state. This can be better understood in the view of the IC effect model that the induced power consumption variations are different on each side of the defect.

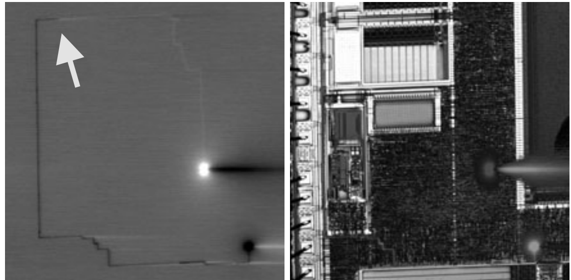


Figure 17 Front-side Thermal laser stimulation image of the current variation acquired with a 5x objective (left) and superimposition on the laser reflected image (right).

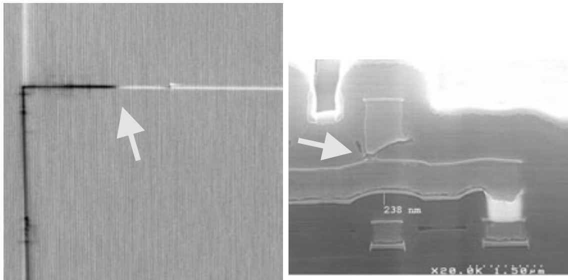


Figure 18 Front-side Thermal laser stimulation image of the voltage variation acquired with a 20x objective (left) and SEM image of the thermal laser stimulation located metallic parasitic via.

Conclusion

We have described several considerations related to understanding the thermal laser stimulation method (OBIRCH, TIVA, SEI) and its implementation into a failure analysis laboratory.

We have shown through models and case studies that Thermal Laser Stimulation methods can rapidly and precisely localize short-circuit defects and current paths. The Thermal Laser Stimulation method is applicable from the front-side of ICs as well as from the backside. In addition, the Thermal Laser Stimulation signature provides information on the nature of the defect. Metals, polysilicon and highly doped silicon substrate areas were found to be thermally sensitive to the NIR laser beam. Finally, the Thermal Laser Stimulation method can also localize material discontinuities and conducting FET transistors.

The Thermal Laser Stimulation method requires only few minutes to a few hours to localize a current leakage failure. A better understanding of the Thermal Laser Stimulation signature makes it possible to gain insight into the physical nature of the located defect.

Acknowledgments

The authors would like to acknowledge the members of the French ANADEF user group on Thermal Laser Stimulation techniques who considerably contributed to the knowledge and know-how on those techniques. Special thanks to M. Gil Gartiez, Tower Semiconductor, M. Gérald Haller and M. Abdellatif Firiti, ST Microelectronics, and to Bernard Baradat, CNES, for their case study contributions.

References

[1] K. Nikawa and S. Tozaki, "Novel OBIC Observation Method for Detecting Defects in Al Stripes Under Current Stressing", Proceedings of the 19th International Symposium for Testing and Failure Analysis, p. 303-310, 1993.
[2] E.I. Cole Jr., P. Tangyunyong, and D.L. Barton, "Backside Localization of Open and Shorted IC Interconnections", 36th Annual International Reliability Physics Symposium, IEEE, p.129-136, 1998.
[3] F. Beaudoin, R. Desplats, P. Perdu, D. Patrie, G. Haller, P. Poirier, P. Jacob, and D. Lewis, "Emission Microscopy and Thermal Laser Stimulation for Backside Failure Localization", Proceedings of the 27th International Symposium for Testing and Failure Analysis, p. 227-235, 2001.
[4] D. L. Barton, P. Tangyunyong, J. M. Soden, A. Y. Liang, F. J. Low, A. N. Zaplatin, K. Shivanandan, and G. Donohoe, "Infrared Light Emission from Semiconductor Devices" Proceedings of the 22th International Symposium for Testing and Failure Analysis, p. 9-17, 1996.
[5] N. M. Wu, K. Weaver, and J. H. Lin, "Failure Analysis from Back Side of Die", Proceedings of the 22nd International Symposium for Testing and Failure Analysis, p. 393-399, 1996.
[6] K. Nikawa and S. Inoue, "Various Contrasts Identifiable from the Backside of a Chip by 1.3μm Laser Beam Scanning and Current Change Imaging", Proceedings of the 22nd International Symposium for Testing and Failure Analysis, p. 387-392, 1996.
[7] K. Nikawa, "Failure Analysis Case Studies Using the IR-OBIRCH (Infrared Optical beam Induced Resistance Change) Method", IEEE, p. 394-399, 1999.
[8] E.I. Cole Jr., P. Tangyunyong, D.A. Benson, and D.L. Barton, "TIVA and SEI Developments for Enhanced Front and Backside Interconnection Failure Analysis", Microelectronic Reliability, Vol. 39, p. 991-996, 1999.
[9] A.H. Johnston, "Charge Collection in p-n Junctions Excited With Pulsed Infrared Lasers", IEEE transaction on Nuclear Sciences, Vol. 40, No. 6, p. 1694-1702, 1993.
[10] T.W. Joseph, A.L. Berry, and B. Bossmann, "Infrared Laser Microscopy of Structures on Heavily Doped Silicon", Proceedings of the 18nd International Symposium for Testing and Failure Analysis, p. 1-7, 1992.
[11] T.W. Lee, "A Review of Wet Etch Formulas for Silicon Semiconductor Failure Analysis", Proceedings of the 22th International Symposium for Testing and Failure Analysis, p. 319-330, 1997.
[12] P. Perdu, "Comparative Study of Sample Preparation Techniques for Backside Analysis", Proceedings of the 26th International Symposium on Test and Failure Analysis, p. 161-171, 2000.
[13] Ioffe A.F., Semiconductor thermoelement and thermoelectric cooling, Inforesearch limited, 1957.
[14] K. Nikawa and S. Inoue, "New Capabilities of OBIRCH Method for Fault Localization and Defect Detection", ATS, p. 214-219, 1997
[15] M. Palaniappan, J. M. Chin, B. Davis, M. Bruce, J. Wilcox, C. M. Chua, L. S. Koh, H. Y. Ng, S. H. Tan, J. C. H. Phang, and G. Gilfeather, "New Signal Detection Methods for Thermal Beam Induced Phenomenon", Proceedings of the 27th International Symposium for Testing and Failure Analysis, p. 171-177, 2001.
[16] R.A. Falk, "Advanced LIVA/TIVA Techniques", Proceedings of the 27th International Symposium for Testing and Failure Analysis, p. 59-65, 2001.
[17] F. Beaudoin, X. Chauffleur, J. P. Fradin, P. Perdu, R. Desplats and D. Lewis, "Modeling Thermal Laser Stimulation", Microelectronic Reliability, Vol. 41, p.1477-1482, 2001.
[18] Metals Handbook 8th Edition, "Vol. 1 : Properties and Selection of Metals", Editeur T. Lyman, American Society For Metals, Ohio, 1961.
[19] G. Hass and E. Ritter, "Optical Film Materials and Their Applications ", Journal of Vacuum Science Technology, Vol. 4, No. 2, p.71-79, 1967.
[20] R. King, C.V. Schaick, and J. Lusk, "Electrical Overstress of Nonencapsulated Aluminum Bond Wires", Proceedings of the International Reliability Physics Symposium, IEEE, p. 141-151, 1989.
[21] O. Paul, M. von Arx, and H. Baltes, "Process-Depedent Thermophysical Properties of CMOS IC Thin Films", The 8th International Conference on Solid-State Sensors and Actuators, and Eurosensors IX, p. 178-181, 1995.
[22] F. Beaudoin, "Localisation de Defaut par la Face Arriere des Circuits Integres", Ph.D. Thesis, Universite Bordeaux 1, No d'ordre 2605, 2002.
[23] S.M. Sze, "Semiconductor Devices: Physics and Technology", John Wiely & Sons, New York, 1985.
[24] C. Boit, A. Glowacki, S. Brahma, and K. Wirth, "Thermal Laser Stimulation of Active Devices in Silicon – A Quantitative FET Parameter Investigation", IRPS 2004.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Introduction to Laser Voltage Probing (LVP) of Integrated Circuits

Siva Kolachina
Texas Instruments Inc, Stafford, TX, USA

Abstract

In this article, an introductory overview of the Laser Voltage Probing (LVP) is presented. LVP is used for waveform analysis and design debug of flipchip packaged integrated circuits (IC). The basics of LVP and its implementation are presented. Variants in LVP methodologies and instrumentation are discussed. References provided allow the reader to further obtain in-depth knowledge of this subject.

Introduction

Laser voltage probing as part of the integrated circuit failure analysis represents an important paradigm shift in the failure analysis community. The advent of flipchip packaging technology was crucial in improving the speed and performance of ICs. However the flipchip packaging of ICs prompted drastic changes in how the ICs could be analyzed. Laser voltage probing emerged as a viable technique to probe diffusions in flipchip ICs.

Probing of ICs is required when an analyst seeks to obtain waveform data at various nodes in a device. The location of interest needs to be made accessible for the probe to be inserted. Conventional techniques of probing include use of mechanical probes, and electron-beam probing[1]. In recent years, AFM type probes are also explored as viable candidates for probing small regions of a device.

The waveform analysis that the analyst may perform include measurement of voltage at a specific location on the device, the switching characteristics at that location, and the comparison of both these parameters between different locations as applicable to the debug strategy. Limitations may be imposed on these measurements by the tools utilized. Mechanical probing is the most widely employed way of obtaining waveforms from a specific location. It is relatively inexpensive to setup and maintain. The main disadvantages of the mechanical probes are access to a site on the device, the impedance of the probes, bandwidth of the signals that could be measured, size of the probes, and speed of probing a large number of locations. Probing a small region on a device may require the probes to be placed in an SEM to take advantage of the higher magnifications these tools provide.

Electron-beam probing is a very powerful technique to probe small geometries in a device. This technique relies on the phenomenon of voltage contrast in a Scanning Electron Microscope (SEM). A calibrated electron beam prober can be used to obtain timing and voltage information in a waveform on a metal line in the device. The limitations of the electron-beam probers are the bandwidth of the probing system, and the ability to access a metal line in the device for probing. The multi-level metal architecture in the ICs necessitate the use of a Focused Ion Beam milling tool to mill selectively such that the electron-beam could be impinged on the underlying metal of interest. Electron-beam probers played a very significant role in ICs as long as it was possible to expose the metal region of interest. This step however became a barrier for electron -beam probing in flipchip devices.

The schematic cross-section of a flipchip device is shown in Figure 1. This design of the flipchip device impedes the ability to expose selective regions of metals for probing with electron-beam (or with the mechanical probes) from the topside of the chip as depicted in Figure 1. The only accessible route is from the backside or the silicon side of the device.

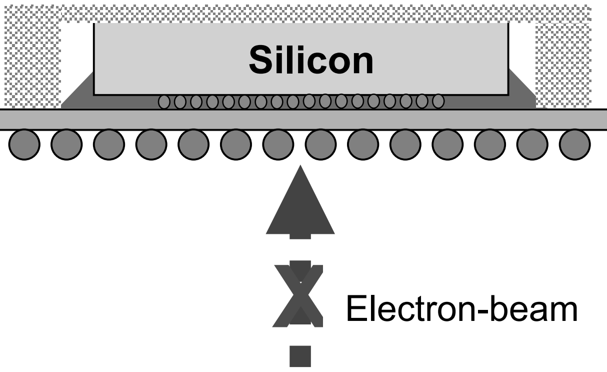


Figure 1: Schematic diagram of a flipchip packaged device. The electron-beam cannot be used to probe metal lines from the topside in such a package.

The need to preserve the functionality of the device, and hence the physical and electrical integrity of the diffusions, dissuades one from milling through these layers using a FIB tool through backside silicon. Laser voltage probing (LVP) emerged as an alternative to the electron-beam probing for applications in flipchip devices[2,3,4]. The LVP technique relies

on the interaction of an optical beam with the electric fields in the diffusions such that waveforms at the diffusion of interest could be extracted. With the LVP, one probes the diffusions and not the metal regions of the device. The immediate advantage of this difference is that a larger number of electric nodes represented by diffusions are available to an analyst, but the ability to measure differences in waveforms between a node at the diffusion level and a node at an intermediate metal level is impaired.

Principles of Laser Voltage Probing

Electro-Optic probing for measurement of waveforms in a flipchip device was reported by Heinrich et al[5]. The application of this technique for integrated circuit failure analysis was reported by Paniccia et al[2]. The technique of LVP is based on the Franz-Keldysh effect[6]. In practice two phenomena, the Franz-Keldysh Effect and the Free Carrier absorption play a role[7]. The interaction of light with silicon devices in the region of diffusions forms the basis of these effects. The implementation in LVP tools is based on the use of 1.06um laser beam.

According to Franz-Keldysh effect, the effective bandgap of semiconductors changes when an electric field is applied as a result of the mixing of the conduction and valence bands[6]. As a result, the refractive index of silicon changes. When light is incident on such a device, absorption of light is impacted by the changes in the refractive index. According to this effect, the intensity and phase of the reflected beam are modulated by the electric fields present in the junctions. In principle, such changes in absorption of light could be tracked to detect changes in the electric field at the device junctions.

The free carrier absorption also plays a role in the use of this technique. This is primarily due to the use of the 1.06um laser which is in the range of free carrier absorption for silicon.

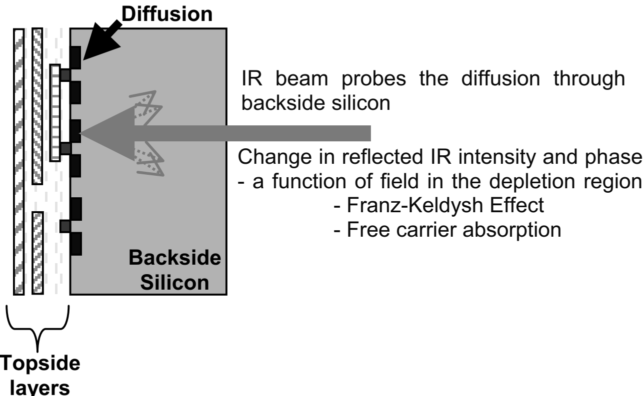


Figure 2: Device configuration for probing with LVP

Eiles et al[7] discussed the relative strengths of the above two components in the reflected laser power. They observed that the phase modulation dominates over the absorptive component. The amplitude-modulation component is predominantly from the free carrier modulation for the laser powers typically available in the LVP systems.

The laser beam is incident onto the junctions through the silicon and the reflected light is collected. The amplitude of the reflected light is a function of the silicon thickness and hence it is necessary to perform this analysis on a thinned device. The silicon thickness is typically between 50um and 100um. Additionally, the phase and amplitude of the reflected light is a function of the electric field and density of charge carriers. If the amplitude and phase of the reflected light can be detected, then one can reconstruct any voltage variations at the junction on which the beam was incident. The intensity of the reflected light is also modulated by free carrier modulation which is in turn a function of incident laser power.

Tool Description and Function

Detection and conversion of these modulations to waveforms that could be interpreted by a design debug engineer is an involved engineering feat. The IDS2xxx[8] series of optical beam probers have evolved to fill this need. The basic system consists of a diffraction-limited confocal laser scanning microscope (LSM), a compact mode-locked (ML) diode-pumped Nd:YAG laser with 1.06um wavelength and a low-power, continuous-wave 1.06um wavelength laser. Both the lasers are utilized for waveform acquisition, while the low-power continuous-wave laser is also used for imaging. During the process of signal acquisition, the incident and reflected intensities of the two lasers are measured. The reflected low-power laser intensity contains information that can be used to minimize vibration induced variations in the reflected ML laser power. The electronic boards synchronize the entire sequence of waveform measurement.

Before proceeding further, let us take a look at the evolution of the LVP technology. The initial series of tools available were capable of measuring the reflected intensity only. Because of the mode of measurement and the use of a continuous-wave laser for noise-cancellation, the absolute voltage levels within the waveforms could not be measured. From a designer point of view, this is a critical deviation from the electron-beam tester technology discussed earlier. A limitation of the intensity-measurement-only tools is that the dominant phase component in the Franz-Keldysh effect was not measured. According to this effect, the intensity and phase of the reflected beam are modulated by the electric fields present in the junctions. The measurement of phase would minimize some of the disadvantages of measuring intensity, and improve the signal to noise ratio. This was subsequently accomplished by the use of a phase-interference-detector (PID) incorporated into the LSM. The design of this detector and its integration into the system improved the vibration sensitivity. Two other critical developments in the LVP tool technology are use of oil based immersion lenses and use of diamond heat spreaders. The immersion lenses improved the image resolution and spot size, while the diamond heat spreaders allowed devices to be operated at high wattages. In the absence of effective cooling of devices using fixtures like

the diamond heat spreaders, the heat from the device would deteriorate the lens used for imaging and probing.

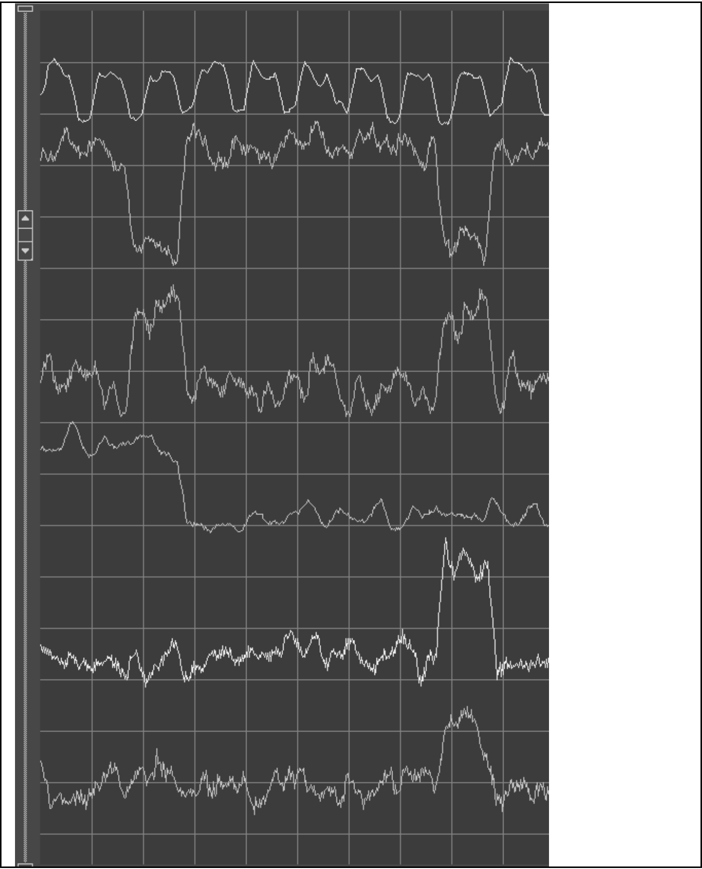


Figure 3: A set of waveforms obtained with LVP with amplitude modulation. The vertical axis is 200ppm/div.

Optical probing allows measurement of high bandwidth signals. The waveforms obtained resemble those from electron-beam testers with some critical differences. The chief difference is that the amplitude of the waveform is not calibrated with voltage and hence the only reliable information is the timing of transitions within the waveform. The second critical difference is the constraints on the stimuli from the tester.

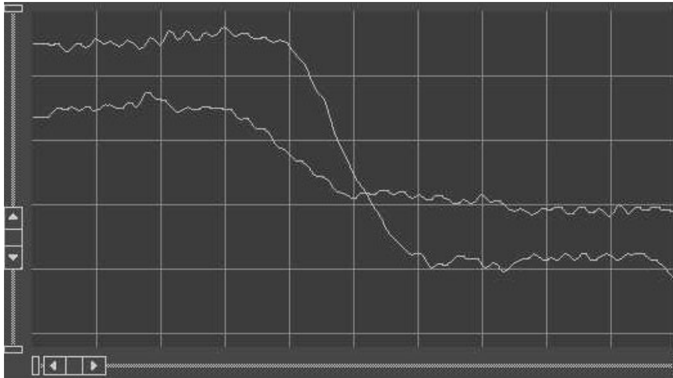


Figure 4: Waveforms acquired at two similar locations. The Y-axis is 100ppm/div. The amplitude information is not reliable as it depends on several factors including focus and incident laser power.

The lack of voltage information in the waveforms does not have to imply that timing analysis is the only one that can be performed. In one case[9] a simulation indicated resistive change of voltage with time, and the LVP signal indicated such a signal at a suspected location (see Fig.5). Based on this information successful resolution of a contact problem was solved.

The intensity-mode of signal acquisition and the PID-mode of signal acquisition follow separate routines. However both these modes rely on accurate synchronization of the tester with the LVP tool. Unlike electron-beam systems, the LVP tools utilize a mode locked laser whose natural frequency is 100 MHz. A phase-locked loop (PLL) circuit is used to synchronize the laser with the tester generating the stimulus. A single pulse from the laser is allowed to be incident on the device per loop, and the position of this pulse is varied to acquire the waveform in a stroboscopic mode[10].

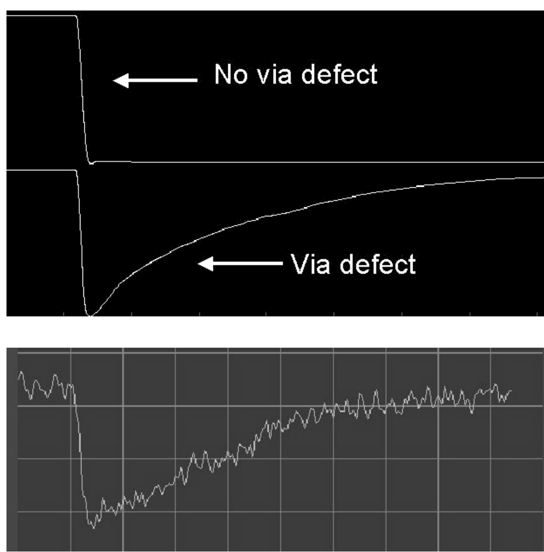


Figure 5: An example of a LVP waveform indicating its use in non-timing related failure analysis. The image above is data from simulation, and the trace below is that of the acquired waveform. Subsequent analysis revealed the suspected via defect.

Phase Interference Detection
The addition of a phase interference detector to the basic LVP system allows detection of phase modulation in the reflected laser[11]. As noted earlier, phase modulation is the dominant phenomenon as compared to amplitude modulation. Systems with PID detector exhibit improved signal to noise ratios and faster acquisition times.

The PID detector implementation involves addition of an interferometer in the LSM. Based on the Michelson-type interferometer this miniature interferometer allows detection of phase sensitivity of the reflected laser beam[11]. Essentially the optical path in the LSM contains two arms, a DUT arm and a Reference arm. The final measurement in this setup continues to be the reflected power, but the use of the interferometer configuration makes this reflected power a function of the phase difference between two beams in the two arms. The interferometer is set up such that adjustments can be made to the reference arm to match path lengths and thus maximize fringe contrast and phase sensitivity under varying reflectivity conditions of the DUT probe points.

In the standard (non-PID) mode of operation, acquisition of waveforms is dependent on focus position, and probe point location within the diffusion which are typically deduced by trial and error. However, in the PID mode, this dependency is reduced and hence the operator has a greater success in obtaining a waveform. The signal to noise ratio is also increased and acquisition times tend to be smaller. The configuration of the PID detector allows the tool to be less prone to external vibrations.

Liquid Immersion Lens
The 100X objective lens with 0.85 numerical aperture is used in the conventional LVP systems. This lens typically yields a beam with full width half maximum of 0.77um. An oil based immersion lens was introduced that reduced the full width half maximum to 0.5um[12]. While the original 100X had a working distance of 1mm, the oil based immersion lens had zero working distance. This results in improved imaging and probing resolution thus extending the use of LVP technology to tighter geometry nodes.

Diamond Window Heat Spreader
Another development in the LVP tool technology is the diamond window heat spreader. While probing devices that are operated at high wattages, there is likelihood that the optics is impacted by the heat emitting from the device. Additionally, the silicon is thinned down to facilitate probing, and this thinned down silicon poses thermal management issues when operated at higher wattages. A heat spreader would allow this heat to be dissipated. A diamond heat spreader is a thin polycrystalline diamond film that is typically thinned to 300um[13]. Such film has adequate cooling, optical and mechanical properties. High power microprocessors generation over 70W have been probed using the diamond window[13].

Sample Preparation
Sample preparation is a critical part of LVP. Typically samples thinned down from the backside with the remaining silicon thickness of 50-100um are used. A smooth surface finish is essential for imaging and probing. To account for the difference between the refractive index of silicon and air, an anti-reflective coating needs to be deposited[14]. The ability of an LVP tool to access signals at the junctions is an advantage even for non-flipchip devices. The increasing number of multi-level metal layers and large coverage of the top surface with bus lines impede the use of electron-beam probers. However in order to be able to use such devices with LVP, the devices must be repackaged such that the silicon surface can be accessed by the 100X lens[9]. The small working distance (1mm for 100X lens, and zero mm for the immersion lens) pose problems in designing boards and sockets that can accommodate them. Typically such hardware needs to be prepared and tested ahead of time.

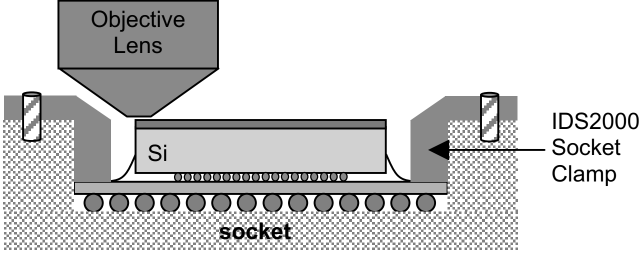


Figure 6: The geometry of the objective lens, and the small working distance necessitate custom configured sockets, clamps and board for tester stimulus.

Further Reading

This article provides a brief overview of the LVP technology and use in failure analysis. The reader is encouraged to review the articles in the references for an in-depth knowledge of the LVP tools and configurations. Flipchip packaged devices are increasingly being analyzed with emission based probe techniques[15]. A comparison of these techniques by Lo, et al[16] is a good introduction to this topic.

References

[1] Electron Beam Testing Technology, Plenum Press, New York, 1993. Edited by J.T.L. Thong

[2] M. Paniccia, T. Eiles, V.R.M. Rao, and W.M. Yee, "Novel Optical Waveform Probing Technique for Flip Chip Packaged Microprocessors," Proceedings International Test Conference ((ITC) p740 (1998)

[3] M. Paniccia, R.M. Rao, and W.M. Yee, "Optical Probing of Flip Chip Packaged Microprocessors," Journal of Vacuum Science and Technology B, 16, p3625 (1998)

[4] S. Kasapi, C-C Tsao, K. Wilsher, W. Lo, and S. Somani, "Laser Beam Backside Probing of CMOS Integrated Circuits," Microelectronics Reliability, 39, 957 (1999)

[5] H.K. Heinrich, D.S. Kent, and L.M. Cropp, “Backside Optical Measurements of Picosecond Internal Gate Delays in a Flip-Chip Packaged Silicon VLSI Circuit,” IEEE Photonics Techn. Lett. 3, 673 (1991)

[6] L.V. Keldysh, “The Effect of a Strong Electric field on the Optical Properties of Insulating Crystals,” Soviet Physics JETP 34, 788 (1958)

[7] T. Eiles, G.L. Woods, and V.R. Rao, “Optical Probing of VLSI Ics From the Silicon Backside,” Proceedings 25th International Symposium for Testing and Failure Analysis (ISTFA), p 27 (1999)

[8] Credence Systems Corporation., 150 Baytech Drive, San Jose, CA, 95134-2302

[9] S. Kolachina, K.S. Wills, T. Nagel, A. Mehta, R. Carawan, O. Diaz deLeon, J. Grund, C.P. Todd, K. Ramanujachar, S. Nagarathnam, "Optical Waveform Probing – Strategies for Non-Flipchip Devices and Other Applications", in Proceedings of International Symposium for Testing and Failure Analysis (ISTFA) (2001)

[10] M. Bruce, G. Dabney, S. McBride, J. Mulig, V. Bruce, C. Bachand, S. Kasapi and J.A. Block, “Waveform Acquisition from the Backside of Silicon Using Electro-Optic Probing”, Proceedings 25th International Symposium for Testing and Failure Analysis (ISTFA), p19 (1999)

[11] K. Wilsher, W. Lo, T. Eiles, G. Xiao, “Integrated Circuit Waveform Probing Using Optical Phase Shift Detection”, Proceedings 26th International Symposium for Testing and Failure Analysis (ISTFA), p 479 (2000)

[12] T. Eiles and P. Pardy, “Liquid Immersion Lens Objective for High-Resolution Optical Probing of Advanced Microprocessors”, Proceedings 27th International Symposium for Testing and Failure Analysis (ISTFA), p x (2001)

[13] T. Eiles, D. Hunt, and D. Chi, "Transparent Heat Spreader for Backside Optical Analysis of High Power Microprocessors," Proceedings 26th International Symposium for Testing and Failure Analysis (ISTFA), p 547 (2000)

[14] B. Davis and W. Chi, “Antireflection Coatings for Semiconductor Failure Analysis”, Proceedings 26th International Symposium for Testing and Failure Analysis (ISTFA), p 155 (2000)

[15] J.A. Kash, J.C. Tsang, D.R. Knebel, and D.P.Vallett, “Non-invasive Backside Failure Analysis of Integrated Circuits by Time Dependent Light Emission: Picosecond Imaging Circuit Analysis,” Proceedings, ISTFA, 483 (1998)

[16] W. Lo, S. Kasapi and K. Wilsher, “Comparison of Laser and Emission Based Optical Probe Techniques”, Proceedings 27th International Symposium for Testing and Failure Analysis (ISTFA) (2001)

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

CAD Navigation in FA and Design/Test Data for Fast Fault Isolation

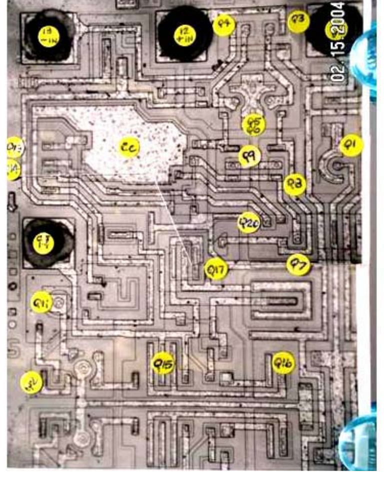
William Ng
National Semiconductor Corp., Santa Clara, California, USA

Abstract

This paper presents an overview of post-processing Computer Aided Design (CAD) data for Failure Analysis (FA). Data from Layout Versus Schematics (LVS) checkers is used to build the fully linked netlist-to-layout navigation system. The use of other design and test data to improve FA cycle time is discussed. Incorporation of logic simulation results to provide golden waveform references, schematic graphical symbols to enhance circuit realization, and Automatic Test Pattern Generation (ATPG) diagnosis for fast fault isolation are possible enhancements.

Introduction

The importance for a navigation system in device analysis cannot be overstated. Composite overlays and optical micrographs are examples of brute force apparatus used to identify circuit elements on layout with their corresponding schematic counterparts. In both cases, the correspondency is established with annotations. As shown in Figure 1, an optical micrograph of an integrated circuit (IC) and schematic were manually annotated for this purpose. While the method is an alternative for simpler devices, complex ICs require advanced systems to identify circuit elements, and hence CAD navigation systems.



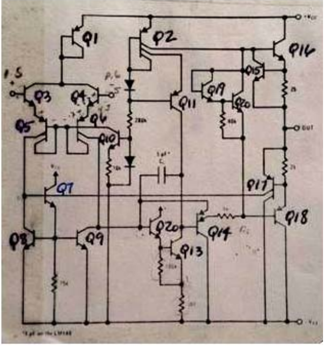


Figure 1: IC optical micrograph and its schematic. Labels were found on both sides for cross-referencing.

A good navigation system with well-planned undelying data structures sets the foundation for it to integrate with other design and test information, which is necessary for reducing FA cycle time. In the following sections, typical characteristics and caveats of CAD navigation systems are discussed. Two applications of preparing CAD navigation data are presented, followed by a discussion of the incorporation of other design and test data (schematic symbols, simulation results and ATPG tester diagnosis) as elaborate methods of fault isolation evolve.

CAD navigation system characteristics

The main purpose of a CAD navigation system is to correlate[1] device schematic to layout and vice versa. Diagnostic systems such as those used in FA can drive the device under analysis to the location of interest. Although centralized design tools that maintain native internal schematic and layout cross-references can be utilized, transporting the navigation ability across diagnositic systems is not straightforward. It also poses performance shortfall as the native data structures are not optimized for the target systems.

Typical CAD navigation systems circumvent this problem by reusing and/or converting information from LVS check during physical design verification in which the consistency between layout and schematic has already been established. The LVS process performs the consistency check by first building a layout netlist from the device physical layout and its primitive element structural information in a process technology file. The layout netlist is then checked aganist the schematic netlist for discrepancies. During the process, intermediate geometric structures, or trapezoids, containing connectivity, or nodal, information are dervied from the input layout polygon geometries. The nodal information tagged onto the trapezoids that are associated with the layout netlist provides the highlighting capability in a CAD navigation system. Post-processing LVS data for CAD navigation is used to overlay the nodal information from trapezoids onto the input displaying layout polygons. The databases produced in this

[1] This correlation between the logical and physical domains is also referred as cross-referencing, cross-mapping or cross-linking. In essence, it is the activity identifying circuit elements from the logical to physical domains or vice versa.

process become the self-contained datafiles and optimized data structures for maximized interactive response performance.

CAD navigation caveats

The information used by the post-processing is typically temporary LVS datafiles. These files are generally discarded at the end of design verification or not generated at all. Retrieving or reconstructing these files later is time consuming at best. A good infrastruture archiving or retaining this information is therefore critical to support FA during the entire product lifecycle. This can be done at product release when all pertinent information is still readily available. The process can be automated and tailored to process technologies in a company. Prepared databases can be stored and managed on centralized file servers for use with different FA tools.

Incompatible LVS netlist(s) between frontend and backend design tools due to differences in design styles is also a problem of using LVS results to build CAD navigation databases. The effect is non-conforming schematic names and design hierarchies. Although the ability of cross-referencing between the LVS netlist and layout is unaffected, correlating circuit elements across design tools could be hindered. Such design tools could be Verilog simulation or ATPG diagnositic environment from which the schematic node paths and circuit elements are structurally and semantically more descriptive.

Concina et. al. reported a software alogrithm in which an index system employs a novel counting method that can traverse design hierarchies and uniquely annotates schematic elements when building the navigation databases [1]. This approach deemed schematic names irrevalent since references were done with unique idenitfiers. This worked well for designs with well maintained hierarchy across the LVS and other design tools.

Although the advent of new generations of design tools is addressing some of these issues, until a single unified environment is utilized that delivers designs with conforming names and hierarchies, datafile ‘massaging’ is inevitable for the device navigation to extend beyond physical design verification, e.g., for FA purposes.

CAD navigation data preparation

CAD navigation databases are constructed by converting the applicable LVS output. There are several commerical LVS packages from which the output files can be converted. Two of these packages will be discussed, Cadence DRACULA and Advant! Hercules.

CAD navigation preparation with DRACULA

Henderson presented a data preparation flow with Cadence DRACULA LVS engine [2]. The flow produced a set of “index” files along with other intermediate LVS files forming a self-contained, optimized environment. The index files were built based on the aforesaid Concina et. al. counting technique [1] and used predominately in electron-beam probing[2].

Figure 2 provides examples of the necessary intermediate DRACULA LVS files from a CMOS process to perform CAD navigation data conversion. The polygon datafiles, which contain geometrical representations of the specific mask layer, are used by the graphically user interface (GUI) to display the device layout. There are also trapezoid datafiles derived from the respective polygon data, carrying nodal information, cross-reference data structures (6NXRF.DAT and 6EXRF.DAT), primitive device data (MOSXP8G.DAT and MOSXN8G.DAT and schematic netlist(s).

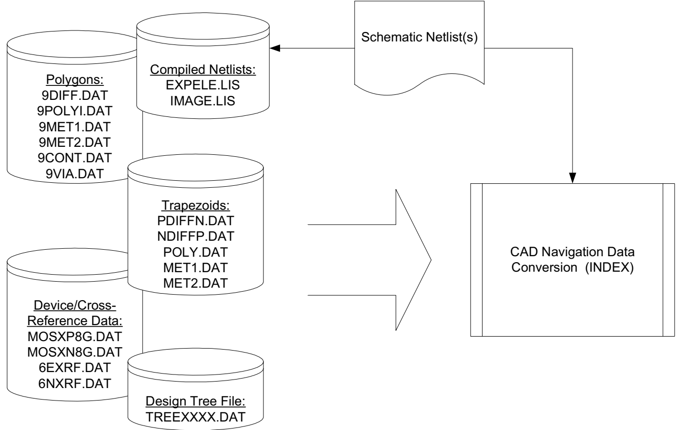


Figure 2: Examples of DRACULA LVS output files for index. Polygon files are characteristically prefixed with a numeral 9. Other file names are either default or determined by various operations within the technology file.

The polygon geometries and the target trapezoids are inter-related through Boolean and connect operations[3] within the LVS technology file. These operations decompose polygons into trapezoids as the connectivity among layout elements are established. An example of the association of trapezoids to the deriving polygons is illustrated in Figure 3. In the example, the routing POLY trapezoids are derived from the input polygons by the Boolean “NOT” operation, which is the substraction of the overlapping geometries of POLYI and RESPOLY. The polygon geometries 9MET1 are decomposed into MET1 trapezoids prior to establishing the inter-layer connectivity with POLY.

In this data preparation flow, both the original and compiled schematic netlists are read by the conversion tool for establishing the cross-referencing. The original schematic netlist(s) is read by the GUI after the data preparation for point and click type of schematic to layout navigation.

[2] Formerly Schlumberger Technologies and now NPTest used this CAD navigation environment on IDS classes of electron-beam systems.

[3] These are operations in physical verification for forming and connecting devices and routing layers, and are needed to extract the layout netlist for LVS check.

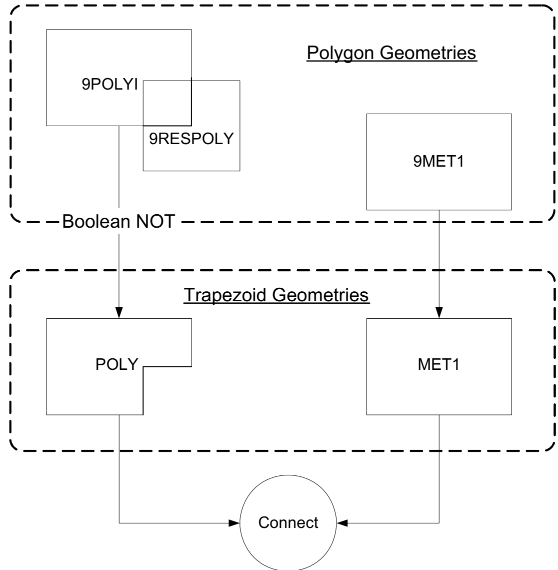


Figure 3: Association of target trapezoids to the deriving polygons.

Orchestrating the overall data crunching is a control file shown in Figure 4. This file is also read by the the layout and netlist GUI for input datafiles. The control file groups the pertinent polygon and trapezoid datafiles together and contains other information such as the layout and schematic top cell names, layout hierarchical tree structure file, inter-layer connect sequence, schematic netlist(s) and other design attributes. The applicable input files are orgainzed in an application and vendor specific directory structure for data crunching. After the data preparation, the polygon files, the tree file, the original schematic netlist(s) and the resulting index files[4] consitute the complete CAD navigation datafile set. Figure 5 is a demonstration of such a directory structure.

```
colormap some_colormap
resolution .25
design top_cell
scale .001
tree TREEXXXX.DAT
layer 9DIFF.DAT PDIFFN.DAT MOSXP8G.DAT
layer 9DIFF.DAT NDIFFP.DAT MOSXN8G.DAT
layer 9POLYI.DAT POLY.DAT
layer 9MET1.DAT MET1.DAT
layer 9MET2.DAT MET2.DAT
layer 9CONT.DAT
layer 9VIA1.DAT
layer 9VIA2.DAT
special 9DIFF.DAT 9POLYI.DAT
count 9DIFF.DAT 9POLYI.DAT 9MET1.DAT 9MET2.DAT
element MOSXP8G.DAT MOSXN8G.DAT
schemtop top_cell
spice some_spice.cdl
```

Figure 4: The control file for running CAD navigation conversion from Cadence DRACULA LVS output for the IDS class of FA diagnostic system from NPTest.

[4] There are four index files, MULTINDEX, NODEINDEX, POLYINDEX and NETINDEX.

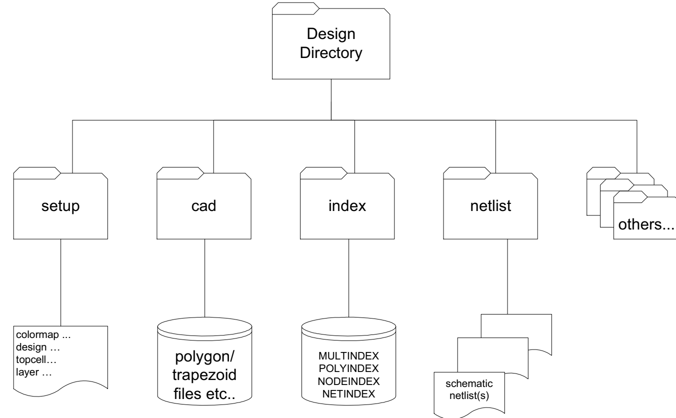


Figure 5: Vendor/application specific directory structure holding the CAD navigation data files after data crunching. This is the data preparation flow used predominately in IDS class of diagnostic system from NPTest.

Cadence DRACULA LVS output can also be post-processed by the Merlin Framework from FEI-Knights. The underlying principle of data preparation from both flows are similar. The orchestrating control file used by Merlin Framework is often referred as "mask.spec". The information contained in either control files are essentially similar and the ressemblances are easy to note after the discussion on Avant! Hercules LVS package in the following section.

CAD navigtaion preparation with Hercules

Another commerical LVS package building CAD navigation database is the Hercules LVS checker from Avant!. Figure 6 is a view of the output files and directories from the LVS checker. The core of the cross-referencing engine is the Runtime Library (RTL) in the "evaccess" directory. It contains internal references to data structures of different views of the design. Views can be thought of different representations of a design. Important views for the CAD navigation post-processing are: layout, layout-netlist, schematic-netlist and cross-reference. The meanings and purposes of these views are somewhat intuitive by names. For example, the "compare" directory holds the LVS comparison results, or cross-reference view, and the "layout" directory contains the physical layout data, or layout view, of the device. The schematic-netlist and layout-netlist views are the original schematic netlist and the derived netlist from layout during LVS check, respectively.

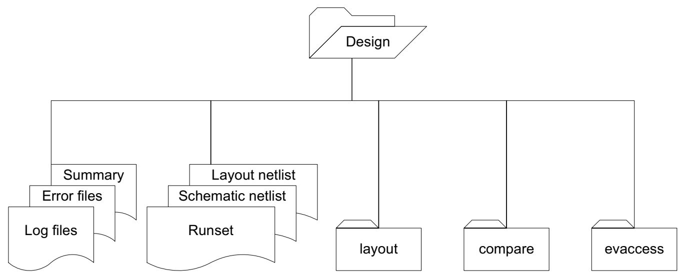


Figure 6: A directory view of LVS output from Hercules design verification tool. Different design views are stored in separate sub-directories or files.

Building the navigation database system requires Application Programming Interfaces (API) to access the encrypted evaccess RTL. Merlin Framework from FEI-Knights features the necessary accessing routine and other utilities for the navigation data conversion.

The caveat with this post-processing flow is the unsettling specification of the encrypted evaccess RTL. Periodic updating of the accessing routines in Merlin Framework is required and inevitable. In addition, some Hercules LVS releases use absolute directory paths in evaccess data structures. As a result, relocating the LVS output directories and data files after the LVS check, e.g., for archival for FA usage later, will require updating the internal referencing data structures with arcane administrative utilities.

Hence, the conversion process for fully linked navigation databases involves checking the integrity and consistence of the LVS output files, updating any path references had the output files were relocated and running the post-processing. A good automation wrapper script should be able to simplify the tasks such that a "push button" operation is possible.

The flow for the data preparation from the Hercules LVS results with Merlin Framework is illustrated in Figure 7. The core components are the accessing routine that interfaces with evaccess RTL, netlist/layout parsers to convert the data into propietary formats and a mapper for sorting trapezoid nodal information onto polygon layout geometries from the LVS check. The resulting Merlin Framework databases form the underlying data structures for the application user interfaces to display the cross-referencing results.

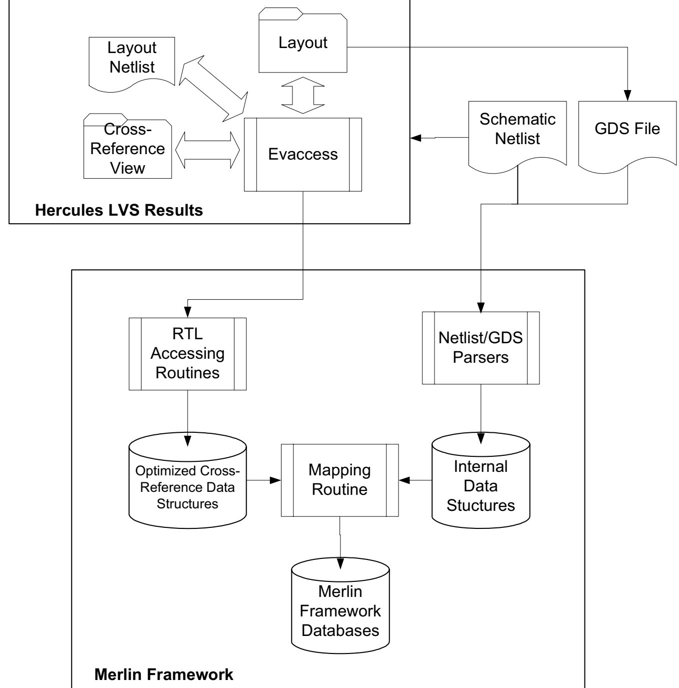


Figure 7: CAD navigation preparation from Hercules LVS.

As in the previous flow, a control file orchestrating the overall data crunching and tool application is shown in Figure 8. An arbitrary but descriptive label is included for each displaying polygon geometry preceding its layer number. For polygon layers with nodal information to correlate to the schematic counterparts, an additional "connect" column are used to list the trapezoid association for mapping the connectivity infomration. Other information in the control are intuitive by names. The layer numbers for the displaying polygons and the associated mapping trapezoids refer to the assigned numbers found in the evaccess specification file in the respective Hercules LVS view.

```
resolution
        0.001 micron;
datafile
        file.gds;
primarycell
        top_cell;
polyfile
        NDIFFP   layernum 19,
        PDIFFN   layernum 20,
        CONT     layernum 10,
        POLY     layernum 21 connect 35,
        VIA      layernum 15,
        MET1     layernum 22 connect 32,
        VIA2     layernum 17,
        MET2     layernum 23 connect 33;
map node_lvs;
```

Figure 8: Merlin Framework control file for CAD navigation data crunching and tool application.

A slight difference in the Hercules LVS flow with the former is a supplement control file for the accessing routine that extracts the evaccess data. The file is often referred as "iss.spec". An example iss.spec is shown in Figure 9. The file contains the directory path of the evaccess data, layout/schematic top cell names and the internal referenced layout layers with nodal information for building the cross-referencing. Note that the layer numbers for these nodal layers are duplicative of the "connect" layers in the "mask.spec".

```
VERIPATH
        /some/directories/evaccess
LAYER
        35 32 33
LAYTOP
        top_cell
SCHEMTOP
        top_cell
```

Figure 9: The supplement control file in Merlin Framework for extracting information from Hercules LVS evaccess data structures.

To sum up, Figure 10 is a detailed illustration of Merlin Framework CAD navigation flow with Hercules LVS output. Four utilities are shown, gdsread, issconv (or spiceconv depending on input netlist format), issread and hmapper. The

former two are converter for layout (gds2) and schematic netlist data into internal Merlin Framework format respectively. “issread” is the utility accessing the evaccess RTL for layout-schematic cross-referencing information. Temporary “np” files are created for the mapping utility (hmapper) to overlay the nodal geometries onto the layout displaying polygons. Also shown are the resulting Merlin Framework database files for the application.

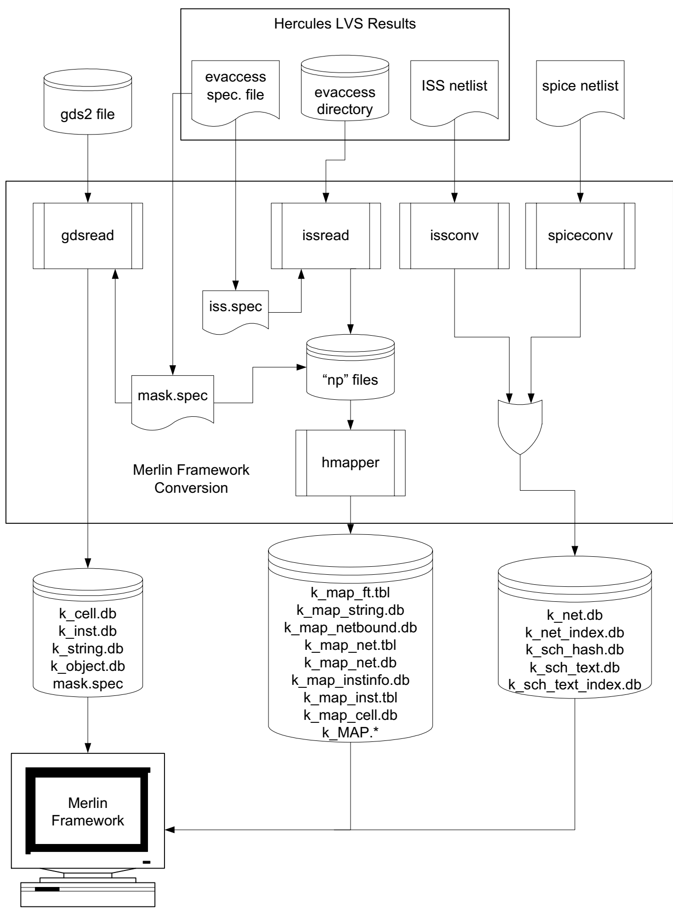


Figure 10: A detailed Merlin Framework CAD navigation data preparation flow. The intermediate files, final databases files and control files are shown. The specific utilities at each step of the flow are also shown (courtesy of Scott Shen from FEI-Knights Company).

A subtle difference between the flows presented in Figures 7 and 10 is the choice of the physical layout (GDS) input file. Figure 7 streams out the GDS file from the native LVS layout representation for the navigation data crunching while Figures 10 uses the original GDS LVS input databases. The use of LVS layout as input for data conversion is primarily a result of creating a manufacturing process ‘shrink’ from the original design dimension. Scaling and biasing factors typically apply in design verification under such a circumstance and may cause the rendering of overlapping nodal trapezoids onto displaying polygons out of place. In any case, the choice of which layout input file dictates the layer numbers to be included for the displaying polygons in the “mask.spec” control file.

Synthesis of schematic symbols

A graphical representation to realize circuit logics, i.e., a ‘schematic’, instead of a textual netlist is inarguably superior. An earlier approach to import schematic symbols into CAD navigation systems was done with the schematic view of Electronic Data Interchange Format (EDIF), an industry specification. The process to generate the necessary data files generally requires running LVS check with the EDIF schematics followed by the post-processing. The issues with this approach were threefold:

- EDIF schematic files are ASCII formatted and the file size tends to be prohibitively huge for large designs.
- EDIF schematic files are not LVS check accommodating, let alone the extraneous symbols information may hinder file content maneuvering. Designers are more inclined to use other netlist formats and generating an extra EDIF file is not always well received.
- While EDIF is a neutral specification, ambiguous interpretation of certain semantics caused problems in mapping EDIF syntax to vendor implementations.

Garyet et. al. discussed a generalization for schematic symbols generation [3]. The method took advantage of the syntactical port I/O direction specified in Verilog to synthesize connectivity between circuit elements. A rule-based algorithm to determine port direction was developed when the port I/O attribute of a signal was not explicitly known.

An interactive schematic tool can also be considered[5]. The tool uses pre-extracted schematic symbols and synthesizes on-demand connectivity with an automatic place and router. The schematic graphical symbols are typically exported from its native schematic capture environment into the EDIF format.

Figure 11 shows the scheme used by the interactive schematic generator. In this scheme the schematic symbols are exported for a given technology library into EDIF. LVS netlists with leaf cells and port names conforming to the native library reuse the EDIF symbol library by the interactive place and router to synthesize the connectivity between logics. An obvious advantage of this approach is the reusable schematic symbol library for netlists from the same process technology but it lacks the look and feel of a ‘full blown’ schematic as with the original schematic capture environment.

[5] It is an optional feature in Merlin Framework. There is also a full schematic option, which requires a name conforming schematic EDIF file.

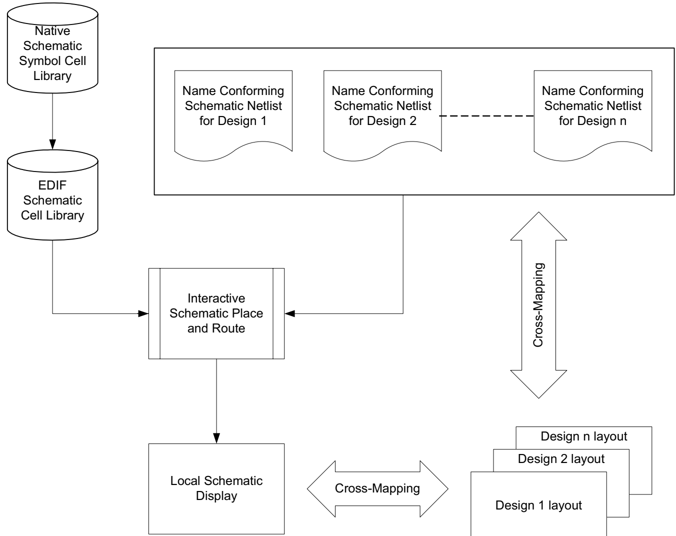


Figure 11: Interactive schematic tool in Merlin Framework. The native schematic symbols are exported into EDIF. LVS netlists by default will be name conformed to the EDIF symbol library used by the interactive schematic place and router.

Figure 12 shows a local schematic interactively generated (top right). The layout counter-part and the netlist representation are also shown. The correlated trace is highlighted in both the layout and the schematic viewers.

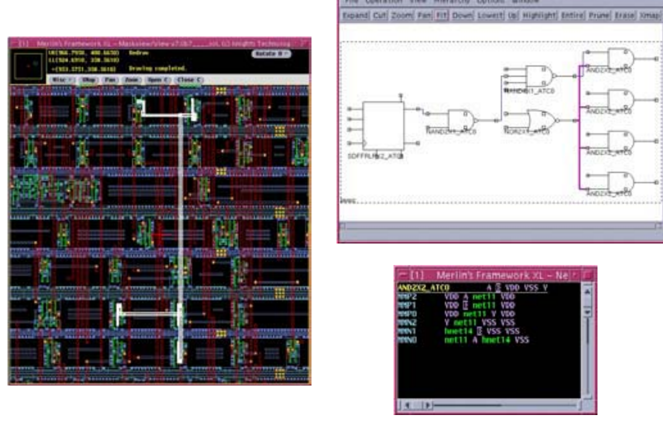


Figure 12: Merlin Framework interactive schematic tool. The correlated trace is highlighted in schematic and layout display.

Integration of simulation results

One challenge of probing complex device is the confidence level of the acquired signal. The question of whether the fault is isolated depends somewhat on familiarity with the circuit being probed or prior knowledge of the signals of interest. A common practice is to probe a golden unit for reference waveforms. This, however, does not solve the problem with non-functional first silicon. Moreover, acquiring golden waveforms consumes time, and is unfeasible under a time constraint situation. Alleviating some of these problems is the use of simulation output. This can be done with paper plots or from a remote session on a workstation.

Ng described an integrated environment with a simulation interface to work with CAD navigation[6], Integrated Diagnostic Assistance (IDA) [4]. The environment is shown in Figure 13. In addition to the indispensable navigation databases for complex designs, the environment provided the optional to compute optimal 'probe-able' locations based on area non-overlap of routing conductor layers. Utilization of this option allowed swift probe point placement during debugging.

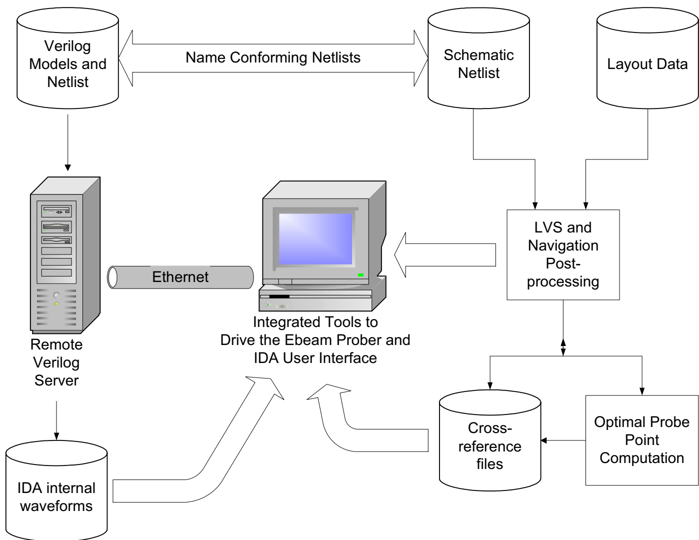


Figure 13: An integrated simulation environment with CAD navigation.

IDA was typically implemented with an ATE direct docking set up in which functional patterns driving the failing device were derived and verified on ATE. The associated simulation results containing logic transitions of all internal circuit nodes were captured and converted into IDA waveform format for retrieval within the integrated probing environment. There was also provision for on-the-fly block level simulation through a network connection when the storage of all levels logic simulation results became prohibitive. In this case, only simulation results of selective higher level logics were stored, the reference waveforms of the lower level logics were simulated and captured as needed. This provision also allowed local simulation to verify faulty circuitry behavior.

The tool also featured heuristics for backtracking faulty circuit logics from a node. The port direction inherent to Verilog syntax facilitated an interactive and cognitional schematic path browser to trace circuit paths backward or forward on demand.

The caveat of such implementation in the early days was the naming non-conformity between the simulation and LVS netlists, mainly due to the differences in design conventions and styles in backend and front-end tools. Considerable efforts

[6] Integrated Diagnostic Assistance, IDA, from formerly Schlumberger Technologies, ATE Division during mid-90s.

were needed to make sure the design hierarchy and net names were consistent for a successful debugging session.

Although the environment is no longer in used actively, this sophisticated implementation brought about the concept of an integrated environment for FA. It was a system to incorporate information from multiple disciplines. Good coordination among functional groups was crucial, but time saving at the debugging stage was considerable.

FA and ATPG

Scan insertion techniques in recent years have been widely used in Design for Testability (DFT). ATPG using scan DFT structures primarily addresses test issues for large sequential circuits, but can also perform IC diagnosis on a tester. Effective application of ATPG and ATE test results can deduce faults such as stuck-at and transition faults. Used in conjunction with CAD navigation, the technique automates fault isolation, which would greatly or totally eliminate electrical FA [5].

The application of ATPG tester diagnosis can be extended to separate wafer fabrication 'killer defects' from benign miniscule particles. It is a process of overlaying the results from ATPG on ATE with in-line scanned optical inspection results at different wafer processing steps. Faulty node candidates deduced from ATPG diagnosis are cross-referenced to the device layout using CAD navigation. Any physical overlapping to in-line scanned defects identifies culprits for further de-processing and inspection of the location [6].

Figure 14 illustrates such an ATPG flow for FA. Unlike bit-mapping on memory arrays from which electrical to physical translation is readily known from the memory architecture, random logics relies on CAD navigation to cross-reference fault candidates from ATPG diagnosis to their physical locations.

Figure 15 was an application in which a faulty node in netlist was deduced using ATPG diagnosis. The physical location of the faulty node was obtained from the CAD navigation system, and subsequent analysis uncovered the stuck-at fault due to a metal defect.

The ability of diagnosing ATE test datalog from ATPG for faulty node(s) is a clear benefit of adopting the methodology to FA. This is possible because of the heuristic from DFT. However, the accuracy of the diagnosis could come at the expense of other issues such as test coverage, silicon real estate and test patterns compression. In the case of multiple predicted faulty locations, other FA tools may be required to narrow down prospective fault candidates.

It is also worth mentioning that ATPG is typically done with gate level netlists and so is the fault diagnosis, meaning the root cause could be originated from node internal to a logic gate. As a result, access to schematic and layout information of the subject logic gates should aid fault localization as the failing device is undergoing layer deprocessing.

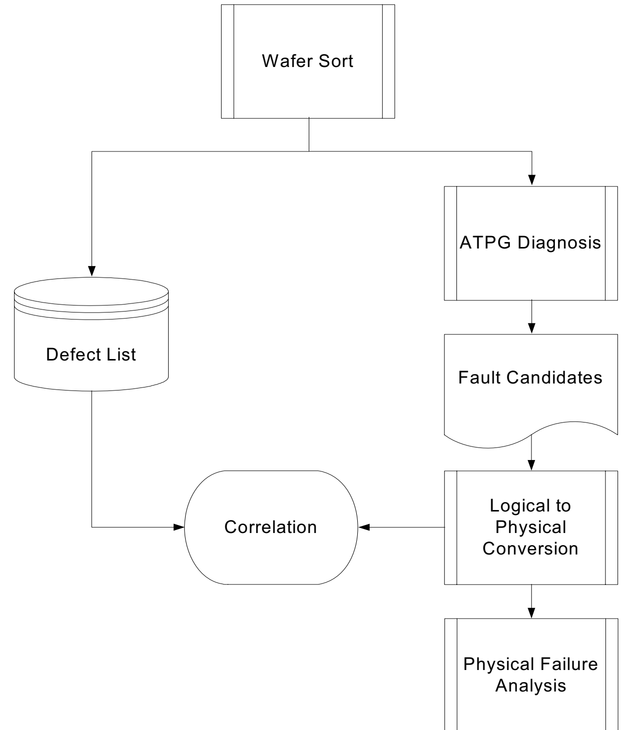


Figure 14: ATPG flow for FA. Only physical FA is needed as the fault sites are localized by ATPG diagnostics.

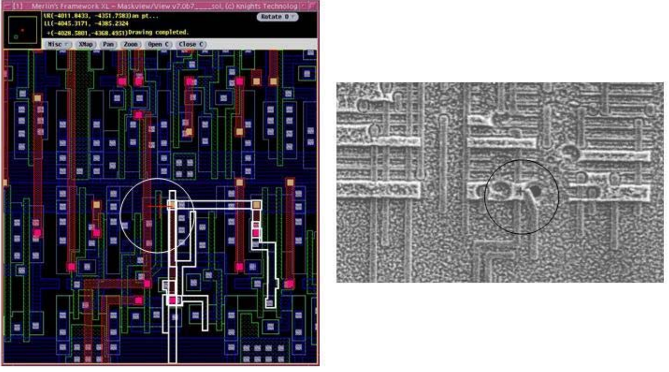


Figure 15: ATPG tester diagnostic produced a fault candidate and is highlighted on the layout (left). Subsequent material de-processing and SEM inspections of the failing die uncovered a metal defect (right). The SEM micrograph was a courtesy of NSC Yield Enhancement Team in South Portland, Maine.

Conclusion

This article described CAD navigation for FA, data preparation and possible incorporation of other design data. A general overview of an LVS process was described. Two commerical LVS packages were examined for data preparation. Although file and directory structures are different, similarity is not hard to realize.

Two approaches to import schematic symbols were presented, both of which utilized EDIF. A previously implemented integrated simulation environment was reiterated.

The applicability of ATPG from DFT designs for FA was briefly examined. The effectiveness of using the methodology was shown in a 'deprocess and look' type of analysis following an ATPG diagnosis.

The combination of these tools enhances the ability of swift FA on complex designs. Effective coupling among these tools is essential.

References

1. Concina, Stefano E. and Gerald S. Liu, *Integrating Design Information For IC Diagnosis*, 24thACM/IEEE Design Automation Conference, 251-257 (1987).
2. Henderson, Christopher L., CAD *Navigation Basics*, Microelectronic Failure Analysis Desk Reference, 4th Edition, 471-475 (2003).
3. Garyet, T. and D. Bodoh, *Schematic Synthesis of Die Navigation during Failure Analysis*, 21st International Symposium for Testing and Failure Analysis, 213-217 (1995).
4. Ng, William, *An Integrated Electron-Beam Probing Environment with a Simulation Interface and CAD Navigation*, Microelectronic Engineering 24, 287-294 (1994).
5. Balachandran, Hari, Jason Parker, Gordon Gammie, John Olson, Craig Force, Kenneth M. Butler and Sri Jandhyala, *Expediting Ramp-to-Volumn Production,* ITC International Test Conference, 103-112 (1999).
6. Balachandran, Hari, Jason Parker, Daniel Shupp, Stephanie Butler, Kenneth M. Butler, Craig Force and Jason Smith, *Correlation of Logical Failures to a Suspect Process Step,* ITC International Test Conference, 458-466 (1999).
7. Synopsis TetraMax User Manual, March 2003.
8. Cadence DRACULA User Manual.
9. Hercules Avant! User Manual.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Acoustic Microscopy of Semiconductor Packages

Cheryl D. Hartfield
Texas Instruments, Inc., Dallas, TX, USA
hartfield@ti.com

Thomas M. Moore
Omniprobe, Inc., Dallas, TX, USA
moore@omniprobe.com

Introduction

The Scanning Acoustic Microscope (SAM) has been adopted by assembly and test facilities, packaging researchers and integrated circuit (IC) failure analysis labs because it provides nondestructive imaging of defects and moisture/thermal-induced damage, such as package cracks and delaminations. The SAM is an important tool for development of improved molded and flip chip packages. It aids in the evaluation of assembly processes, die and package material selection, and detection of failed packaged parts during reliability testing. The reflected acoustic signal provides a wealth of information. The modern SAM provides automatic polarity analysis of the reflected pulse, which assists in the detection of delaminated interfaces. The reflected signal also enables production of images showing the three-dimensional internal structure of the package, which are often useful in recognizing package defects and in determining the mechanism for a package failure.[1] Also, the transmitted acoustic signal is useful when a sample is comprised of one or more layers that are thin relative to the wavelength of sound. Here are presented the historical development of SAM for IC package inspection, SAM theory, analysis considerations, and practical applications. Other non-destructive imaging tools are briefly discussed, as well as future SAM challenges.

History

The SAM used today in the IC industry is a hybrid instrument with characteristics of both the scanning acoustic microscope developed at Stanford in the early 1970s, and the C-scan which has been a part of the nondestructive test (NDT) industry since the 1950s.[2] The characteristics of each of these methods are briefly reviewed.

The C-scan

The term C-scan comes from early NDT nomenclature, which in turn was derived from early radar terminology. The C-scan image is an image of a planar area within the sample that is typically perpendicular to the incident sound beam. The C-scan image can be formed by mechanically scanning a piezoelectric transducer above the specimen and electronically gating the signal in time. The broad-band C-scan transducer has a lens designed for sub-surface imaging. C-scan imaging has played a major role in the macroscopic imaging of sub-surface flaws in industrial components (rails, pipe, welds, etc.) with center frequencies in the range of 1-10MHz. C-scan inspection takes advantage of its ability to penetrate optically opaque solids and detect thin cracks.[3,4]

The Stanford SAM

The SAM developed at Stanford was first demonstrated by Lemmons and Quate in 1973. It employs a large numerical aperture (NA) lens in order to excite longitudinal, shear and surface waves in the sample.[6,7] Instead of the large water baths of C-scan, a tiny water droplet acoustically couples the transducer and sample. Image contrast is formed by the combined interference of longitudinal, shear and leaky surface waves. Narrow-band RF pulses with frequencies in the range of 100 MHz-8 GHz are used. The upper frequency limit and spatial resolution limit are determined by frequency-dependent attenuation in the couplant. Cryogenic fluids and high pressure gases have been used as the couplant for frequencies above 2 GHz. Precision mechanical scanning is employed for sub-micron spatial resolution. Both the amplitude and phase of the reflected pulses are measured and used to produce images of the mechanical properties of the near-surface region.[5]

Development of the SAM at Stanford was strongly encouraged by biomedical researchers who anticipated improved contrast in acoustic images of tissue samples relative to the contrast possible with light imaging. Contrast in the SAM is determined by the large variation in elastic properties in these samples compared to a relatively small variation in dielectric properties, which determines optical contrast. Tissue samples normally required complicated chemical staining for optical microscopy. In the years following its initial implementation, the frequency of the Stanford SAM was continuously increased until the lateral resolution was comparable to that of optical microscopes.[8]

Today's SAM for IC Package Inspection

The application of reflection acoustic microscopy to the inspection of IC packages represents the convergence of the capabilities of focused C-scan and the Stanford SAM. It combines precision mechanical scanning for microscopic inspection of small samples, sophisticated RF signal analysis

and display, and a broad-band transducer with a small NA lens for sub-surface imaging.[9,10] Center frequencies commonly used are in the range of 15-300 MHz and are intermediate between the frequencies commonly used for C-scan and the Stanford SAM.[51-53]

Early SAM Applications in the IC Industry

SAMs operating at frequencies of 1 GHz and higher became commercially available in 1983. The initial application of SAM in the semiconductor industry was the high frequency inspection of device layers near the surface of the die. At these high frequencies (1 GHz and above) the depth of penetration was severely limited by frequency-dependent absorption. Some SAMs were modified to take advantage of available broad-band NDT transducers in the intermediate frequency range of 30-100 MHz for sub-surface studies such as die attach inspections. Die attach inspection through ceramic or metal packages had already been demonstrated using high resolution C-scan equipment. Prior to acoustic inspection, die attach inspection was performed primarily with x-ray radiography. Experience soon showed that reflected sound indicates the true percent of the total area of a die attach interface that is bonded while x-ray inspection reveals only large voids in the die attach material.[11-22]

Studies using pulse-echo acoustic microscopy for plastic package inspection began appearing in the open literature around 1985. These early applications were performed mostly by Japanese IC manufacturers. These instruments incorporated precision microscope-type scanning and improved data analysis and presentation features. However, the echo signals in the studies typically were rectified and delamination detection was based solely on amplitude imaging. These early studies were instrumental in correlating the amount of damage detected in plastic packages after reflow soldering to the moisture content in the package.[23-28]

After 1988, the limitations of the detection of delamination by amplitude alone in plastic-packages ICs were recognized. Reports using SAM instruments dedicated to IC package inspection began to appear.[1,29-32] Some of these instruments had the ability to detect phase inversion.[50]

An Important Tool for Popcorn Crack Detection

Acoustic microscopy rapidly rose to prominence as a key package development tool during the industry conversion in the 1980s from packaging small dies in conventional through-hole dual in-line packages (DIPs) to packaging large high-functionality dies in space-efficient surface mount packages. This conversion aggravated a basic materials problem with the molded IC package. The molded package is made up of materials with widely varying coefficients of thermal expansion (CTE) (Fig. 1) and is required to survive many large temperature excursions. The assembly of large-die surface mount molded packages can result in the development of moisture/thermal-induced stresses sufficient to exceed the mechanical strength of the materials and interfaces in the package, resulting in "popcorn crack" failures. Studies were published which correlated electrical testing and destructive physical analysis with the results of acoustic inspection in order to understand the moisture sensitivity of modern surface mount packages during board mounting.[33-37]

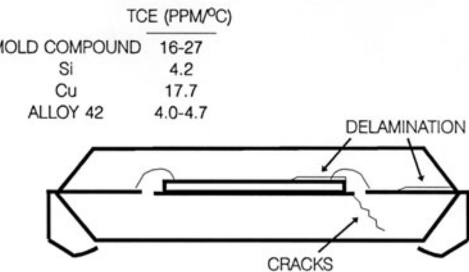


Figure 1: Coefficient of thermal expansion mismatches in plastic IC packages.

The internal stress (CTE) situation is further exacerbated by the fact that the plastic mold compound tends to absorb moisture from the air during shipping and storage. This is a problem when the device is soldered to the printed circuit board. Wave soldering of DIPs delivers a comparatively lesser thermal stress to the body of the package than that experienced by surface mount packages. When surface mount parts are reflowed, the entire package body is exposed to soldering temperatures. Absorbed moisture, especially at internal interfaces, expands during the high-temperature mounting operation. This greatly increases internal stresses and promotes delamination and package cracking. This "moisture sensitivity" is a problem primarily with surface mount package designs.

In the late 1980s, devices designated as being moisture sensitive began shipping in special dry bags. Limits were set for the maximum duration of exposure to air before assembly. These limits were based on a recommended moisture level threshold for the appearance of moisture/thermal-induced damage during mounting.[38] Dry packing did not significantly increase the manufacturing cost of the product. However, the possibility of mechanical damage and production delays associated with moisture control (such as in the baking of over-exposed devices) represented a risk for the assembly operation. IC manufacturers and mold compound producers worked to optimize package designs and mold compound characteristics in order to provide moisture insensitive packages. SAM played a critical role in the development of robust surface mount packages and more recently has proven itself essential in aiding the development of assembly processes for Cu/low-k dielectric devices. SAM is used by virtually every major semiconductor and packaging manufacturer, as well as a majority of failure analysis labs, and is required by international standards committees for certain package types.[58]

Theoretical Considerations

Overview: Acoustic Microscopy for IC Package Inspection

Typically, center frequencies in the range of 15-300 MHz are used for IC package inspection. The selection of a specific

frequency is dependant upon package construction and materials. Typically, a single acoustic transducer is mechanically scanned to form a C-scan image, and the sample and transducer are acoustically coupled by de-ionized water. Broad-band acoustic pulses are focused to a point within the IC package (Fig. 2). The pulse repetition rate is typically limited to 10 KHz due to decay of the reverberations that occur between the transducer and sample. The transducer is precisely scanned in a plane parallel to the plane of the package for microscopic imaging. At internal interfaces, a fraction of the incident acoustic energy is reflected and detected by the same piezoelectric transducer and converted back to an electrical signal. The amount of incident acoustic energy that is reflected depends on many factors including the materials in contact at the interface, the mechanical properties of the interface, absorption, and the size and orientation of the interface. The echo signal is analyzed and characteristics of the signal are used to form images of internal structures and defects. Sophisticated signal analysis techniques are used to extract characteristics from the echo signal such as amplitude, phase and depth. Because sound is a matter wave, the technique is sensitive even to cracks that are invisible to x-ray radiography.[1]

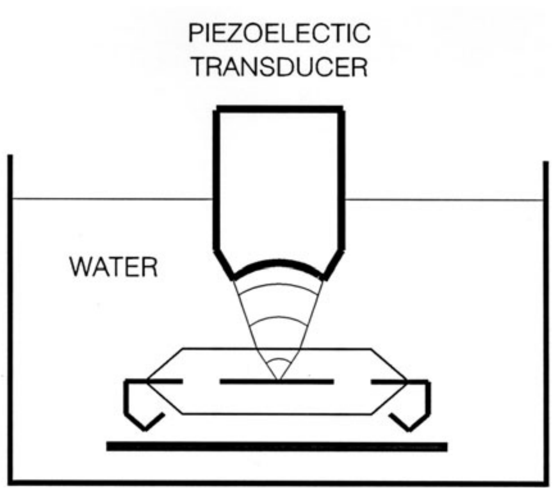


Figure 2: The inspection of IC packages with pulse-echo acoustic microscopy.

Use of reflected signals for acquisition of SAM images is commonly referred to as reflection mode or pulse-echo mode imaging. C-mode SAM and "C-scan" automatically infer the use of pulse-echo analysis. This analysis mode is the dominant method used for package inspection due to the ability to identify in many package types the exact interface at which a defect occurs, in addition to the wealth of information contained in the reflected signal as previously discussed. However, imaging by use of the transmitted acoustic signal is also frequently performed for IC package inspection. Use of the transmitted signal is referred to as through-transmission imaging. Pulse-echo mode and through-transmission mode methods of scanning acoustic microscopy can be abbreviated as PE-SAM and TT-SAM, respectively.

The use of TT-SAM has gained popularity due to the emergence of packages with multiple thin layers, including Ball Grid Array (BGA) packages and thin chip-scale packages (CSPs), which often are 1 mm or less in thickness. The echoes from thin layers can overlap and constructively or destructively interfere with each other. An "interference" image made using such echoes can provide extremely misleading information, although the image appears convincing.[54,55] A pre-analysis assessment of whether interfering echoes are a concern is not always possible due to lack of information regarding the package materials and construction. For these package types, TT-SAM is often employed as the primary means of acoustic inspection for suspected defects (rapid screening), and PE-SAM is used as an optional follow-up to verify defects and pinpoint their location. The use of TT-SAM as a screening tool is discussed further later in this chapter.

Types of Reflected Signal Analysis Modes:
For time-domain analysis of the reflected acoustic signal, there exist 3 main types of analysis modes for PE-SAM analysis, and they are designated simply as A-scans, B-scans, and C-scans. Each one provides a different perspective of the sample and accomplishes specific functions. These analysis modes are discussed here in the typical order in which they are applied.

The A-scan is the simplest time-domain analysis of an echo signal and provides critical information that allows one to properly set the transducer focus, measure thicknesses, and provides the first glimpse into the acoustic characteristics of the sample. To obtain an A-scan, the transducer is placed over the sample at a specific x-y location, and the reflected signal is captured at that x-y location as a function of the time-of-flight of the signal (x-axis) vs. the signal amplitude (y-axis). This is illustrated in Fig. 3. Since the A-scan is used to establish proper focus, it is essential that the transducer be positioned over the sample at the x-y location of interest. For example, the transducer would be placed towards the edge of the package if trying to focus on the lead frame, or over the center of the package if trying to focus on the die surface.

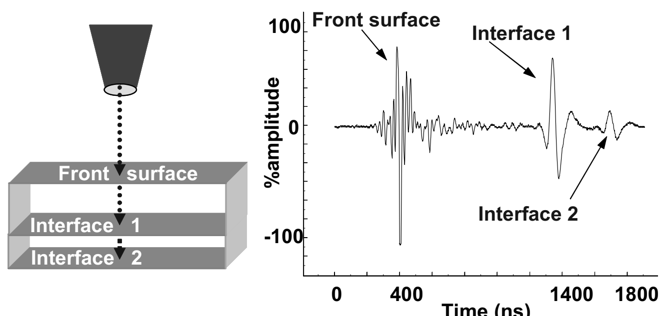


Figure 3: An A-scan displays echo amplitude information at a single x-y location of the sample as a function of the echo return time.

The A-scan is an essential component to understanding and interpreting SAM images. If any question arises as to the nature of a feature in a SAM image, A-scans obtained over the

feature in question, as well as from over a "normal" area, can provide extremely useful information. It enables assessments such as whether the sample was imaged in focus, whether the data gate was positioned properly, whether the transducer had sufficient resolution to separate echoes of interest, whether a phase inversion truly occurs, and at what depth a feature occurs.

A C-scan is typically the type of analysis that is performed to produce the PE-SAM images that often frequent failure analysis reports. As shown in Fig. 4, a C-scan image is obtained by gating an echo at a specific depth from within the A-scan, scanning the transducer across the sample in both x and y axes while acquiring the signal from within the gate, and converting the amplitude of the gated signals to grayscale values, which then are plotted in relation to their x-y location. Ideally, this produces an image of the sample representing a single plane within the sample that is perpendicular to the direction of the incident acoustic pulse.

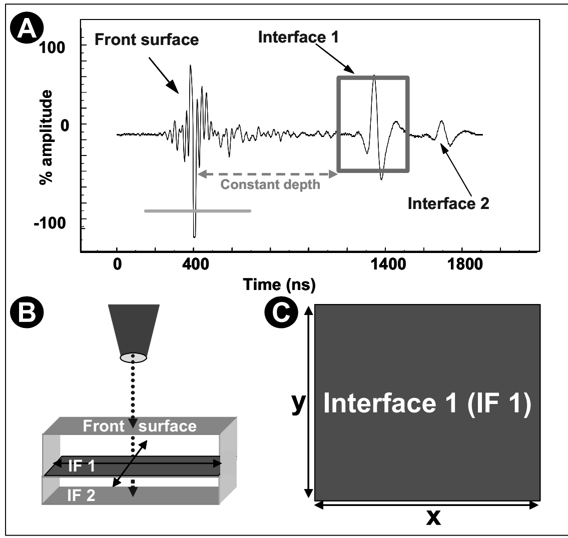


Figure 4: By gating a specific A-scan echo in the time domain (A) and scanning the transducer over the entire X-Y sample area of interest (B), a C-scan image is obtained (C).

The B-scan is often referred to as a "virtual cross-section". Figure 5 shows that the B-scan is simply a grayscale representation of a series of A-scans plotted one after another (Fig. 5(A)), with depth as the vertical axis and position along the single line scan as the horizontal axis in the image (Fig. 5(B)). In the B-scan image, the plane of the image is parallel to the direction of the incident acoustic pulses. The typical B-scan image described here is made with a single line scan, but can also be made with a series of repetitive scans along the same line in the x-y plane with different z positions of the transducer to optimize the focus as a function of depth within the sample. B-scans are frequently used for visualization of depth information within a sample and are effective for revealing package defects such as die tilt. They are also a good way to evaluate whether features from the top surface of the sample are interfering with signals deeper inside (example: a scratch or air bubble on the surface will cause disturbances to all subsequent echo signals at the same x-y location). Most acoustic microscope software is written so that a C-scan must be acquired before a B-scan can be performed. Once a C-scan is obtained, an area from the C-scan is selected to indicate the area of the sample to be analyzed in B-scan mode.

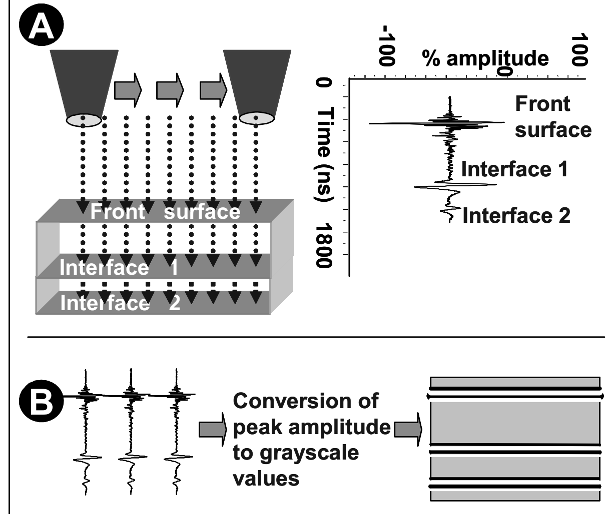


Figure 5: A B-scan is a series of A-scans from a line scan (A) converted to grayscale values and plotted along a single x or y axis. The end result is a "virtual cross-section" of the sample (B).

Combined and properly used, the A-scan, B-scan and C-scan offer powerful ways to understand the nature of a sample's interior, rapidly and non-destructively.

Polarity Analysis of Reflected Signals

Information about delamination at internal interfaces in plastic-packaged ICs is assisted by applying polarity analysis to the acoustic echo signals. Polarity analysis is helpful, because not all delaminations result in an amplitude change. Thus, delaminations can be missed if relying solely on amplitude information. This is especially true, for example, when evaluating the interface between package lead frames and mold compound (see Fig. 6).

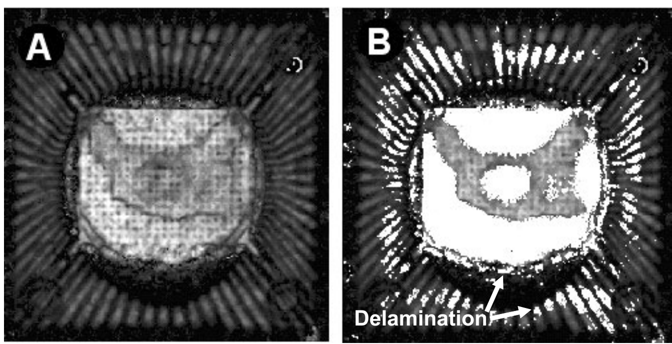


Figure 6: C-scan images of the mold compound to die paddle and lead frame interface of a 68PLCC, obtained with a 50 MHz transducer. (A) The peak amplitude image shows a range of grayscale variation, but delamination locations are not obvious. (B) The phase inversion image clearly reveals the delaminations.

Figure 7 shows typical examples of acoustic echo signals from an area with good adhesion, and a delaminated area, on the die of a 68-pin plastic leaded chip carrier (68PLCC). In each of the two echo signals in Fig. 7, a reflection from the top surface of the package and a later sub-surface reflection from the die surface can be seen. Note the 180° phase inversion of the reflection at the delamination, and the deeper partially resolved reflections in the signal from the die surface area with good adhesion to the package.

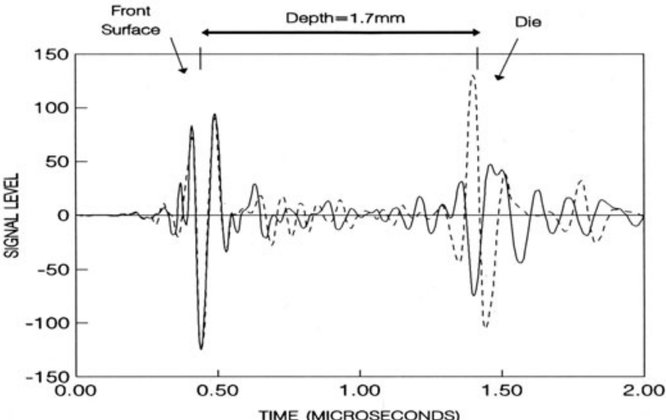


Figure 7: Typical acoustic echo signals (15MHz) from an area of good adhesion (solid) and a delaminated area (dashed) at the mold compound/die interface in a 68PLCC.

For an explanation of the phase inversion phenomenon, consider the simplified example of plane wave reflection, at normal incidence, at an ideal planar interface (Fig. 8).[39] The incident plane wave has the sinusoidal acoustic pressure amplitude P_I and reflected and transmitted pressures amplitudes P_R and P_T, respectively. As a result of the boundary conditions that the acoustic pressure and particle velocity in both materials must be equal at the interface, the frequency remains unchanged across the interface, and the reflected and transmitted pressure amplitudes can be described as functions of the acoustic impedances, Z_i, of the two materials (Eq. 1 and 2).

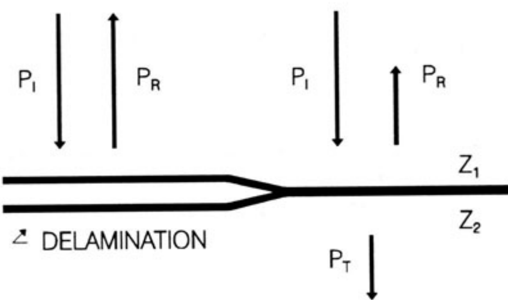


Figure 8: Reflection at normal incidence of a plane wave at a delamination (left) and at a bonded interface (right).

$$P_R = (Z_2 - Z_1)/(Z_2 + Z_1) \quad [1]$$

$$P_T = 2\, Z_2/(Z_2 + Z_1) \quad [2]$$

The acoustic impedance is the ratio of the acoustic pressure to the particle velocity per unit area and can be estimated in this model as the product of the density (ρ_i) and the speed of sound (v_i) in layer i.

$$Z_i = \rho_i\, v_i \quad [3]$$

Equation 1 is plotted in Fig. 9 for two values of Z_1. Curve 1 is calculated for Z_1 equal to the impedance of mold compound (plastic package), and Curve 2 for Z_1 equal to the impedance of Al_2O_3 (ceramic package). Note that each curve passes through the horizontal axis at the value of Z_2 for the appropriate package material. This indicates that there should be no reflection at an ideal interface between identical materials, as expected. In Fig. 9, reflectivities less than zero refer to reflected pulses with inverted phase relative to the incident pulse.

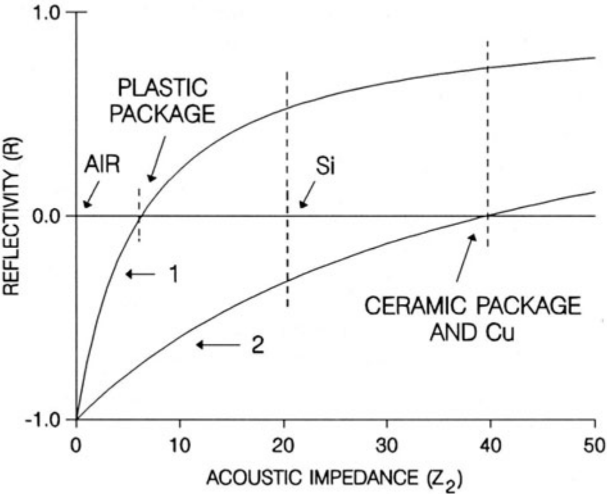


Figure 9: Ideal acoustic reflectivity (R) versus acoustic impedance of the second layer (Z2) for plastic packages (Curve 1) and for ceramic packages (Curve 2). (Z units: 105g/cm2sec)

Curve 1 in Figure 9 shows that at bonded interfaces between the plastic mold compound and the die (MC/die), and between the mold compound and a Cu lead (MC/Cu), the transition is from lower to higher acoustic impedance. Therefore, P_R is positive at these interfaces and there is no phase inversion. However, at a delamination or a package crack, which is represented by an interface between mold compound and air (MC/air), ideally 100% of the energy is reflected, and the phase of the reflected pulse is inverted relative to the incident pulse. This model does not include the effects of attenuation losses. Attenuation losses in plastic packages can often obscure the increase in the amplitude of signals reflected at delaminations that are deeper in the package than another interface that is not delaminated. Phase inversion is very important for reliable detection of delamination and cracks in plastic packages.

The simplified plane wave model (Eq.1) is useful in describing phase inversion of reflected acoustic pulses at delaminations and cracks. In practice, apparent phase shifts at similar interfaces due to multi-layer interference effects, frequency dependent attenuation and spatial resolution limitations can also be encountered. In addition, interfaces at a

constant depth can produce reflections with a continuous variation in phase shift between bonded and delaminated areas that are not explained by the simplified model. These intermediate phase shifts may provide additional information on the condition of the interface.[48,49] However, in spite of these practical limitations, the detection of phase inversion in reflected acoustic pulses has proven to be extremely useful in the detection of delaminations in plastic IC packages, and is a distinct advantage of pulse-echo inspection.

Curve 2 in Fig. 9 describes the ideal reflectivities in ceramic (Al_2O_3) packages. In typical ceramic package applications, SAM is used to inspect the die attach layer. Acoustic inspection offers an advantage over x-ray inspection of die attach quality in that acoustic images show the actual area of good adhesion while x-ray images indicate only voids in the die attach layer.

Table 1: Acoustic Parameters

Material	v, (m/sec)	ρ, (g/cc)	Z, 10^5g/cm^{2}sec
Al_2O_3	10400	3.8	40
Cu	4400	8.9	39
Si	8430	2.4	20
Mold Comp.	~3500	1.8	6.3
Water	1480	1.0	1.5
Air	343	0.0012	0.00041

Table 1 shows the acoustic impedances of Al_2O_3 and Cu are very similar. The acoustic impedance of the ceramic package is so high that phase inversion detection is not applicable in ceramic package inspection. However, this is compensated for by the fact that the amplitude contrast in an image is typically very high. In a ceramic package with Cu leads, for example, the ceramic/Cu reflections are weak compared to a reflection from a crack or a disbonded lead. These defects provide almost 100% contrast. Inspection of eutectic die attach quality also typically shows dramatic contrast between bonds and disbonds. Polymeric die attach adhesives provide lower, but sufficient, contrast.

Time of Flight Imaging Using Reflected Signals

In addition to polarity analysis, PE-SAM allows images to be displayed based on acoustic velocity and depth information. The speed of sound in materials is typically less than 13 km/sec. This is roughly four orders of magnitude less than the speed of light. The time delay between returning echoes can be easily measured electronically and images with three-dimensional information can be displayed. This is a unique advantage of reflection mode imaging and has been useful in determining the mechanism for package crack formation, for example. Figure 10 shows the acoustic time-of-flight image of a cavity-down ball grid array (BGA) package (die-side view). The contrast at the die surface indicates the variation in depth of the die surface relative to package surface (darker means deeper).

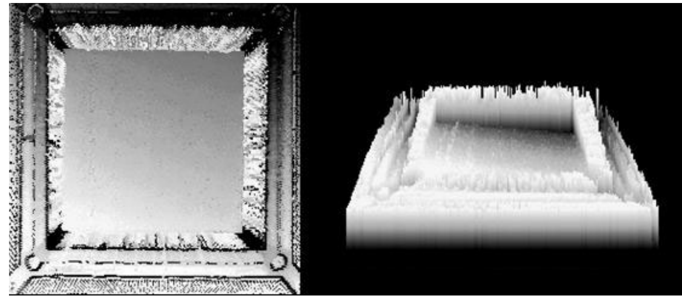


Figure 10: Time-of-flight image of a cavity-down BGA package. Darker areas are deeper in the package. A 3-D view is shown to assist visualization.

Figure 11 is a B-scan image of the same device. The B-scan clearly shows the tilt of the die (caused by uneven die attach thickness and substrate warpage). However, it does not convey the depth of all 4 die corners at the same time, as does the time-of-flight image. Hence, the 3-D visualization is useful for obtaining qualitative information over a larger area. The B-scan is superior, however, if one wishes to make the depth information quantitative (if acoustic velocities for a material are known, the time domain can easily be converted to specific depth values).

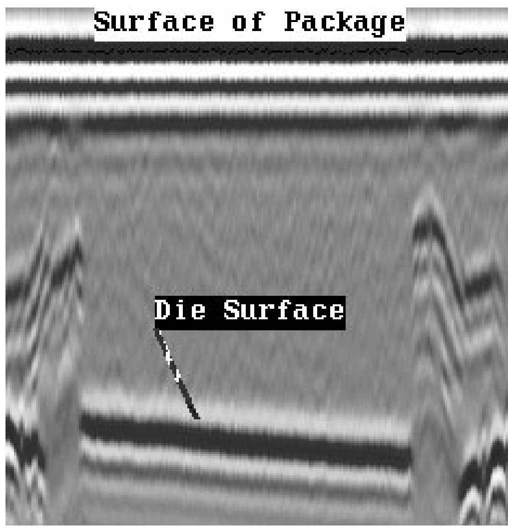


Figure 11: B-scan image showing die tilt, acquired with a 75 MHz transducer focused at the die surface.

Resolution and Sensitivity of SAM Inspection

The resolution and sensitivity of SAM is dependant upon many factors, some of which are sample dependant, and some of which are hardware dependant (including transducers, amplifiers, cables, and other electronic components). Thus it is not possible to clearly state resolution capability based on the center frequency of the transducer alone, as every sample type and every acoustic microscope may produce slightly different results.

Transducer properties play a key role in resolution. The spot size obtainable with a spherical lens is limited by diffraction effects. If the resolution is defined by the first zero in the Airy

disk (Rayleigh criterion), the lateral resolution (d) is given by [40]:

$$d = 1.22 \lambda F / D \quad [4]$$

Here, λ is the acoustic wavelength, F is the focal length, and D is the diameter of the lens. The F/D term is the inverse of the numerical aperture (N) of the lens. Perhaps a more representative resolution criterion with today's electronic signal processing technology is the detection of a significant local minimum in the signal between two points in the image (Sparrow criterion). Also, unlike the case with a telescope, the acoustic transducer lens both transmits and receives the signal pulse, so the point transfer function is squared at each point in the image. This means that the constant in Eq. 4 may be as small as 1.02.

The value of F/D ranges from 2 to 4 for typical transducers used for subsurface inspection in IC packages. So, practically speaking, the best resolution obtainable is roughly twice the wavelength. At a center frequency of 75 MHz the wavelength in water is approximately 20 μm and the best expected lateral resolution with F/D=2 is roughly 40 μm. This does not account for frequency-dependent attenuation. Attenuation in a typical mold compound has been reported to be 40 dB/cm at 15 MHz and to increase rapidly with increasing frequency.[41] This attenuation acts as a low-pass filter and shifts the center of the pulse frequency distribution to a lower frequency. In a very highly attenuating mold compound (large irregular filler particles) the observed spatial resolution can be as large as 400 μm at a depth of 1.6 mm (or 3.2 mm round trip). Mold compound attenuation varies considerably from one formulation to the next. Both the penetration and resolution are noticeably degraded by temperature shock damage to the mold compound.

Depth (or axial) resolution is important for distinguishing reflections from closely spaced layers within the package. Depth resolution in the time domain is determined by pulse duration as well as frequency. The inherent decay time for the transducer, focusing properties of the lens and frequency dependent attenuation all contribute to pulse duration. A typical pulse duration for a broad band transducer is 2 periods at the pulse center frequency. This effect creates what has been termed the "dead zone" below an interface, in which a reflection from this interface may interfere with the reflection from a deeper interface. This makes detection of the deeper interface difficult, especially since the signal amplitude incident on the deeper interface is diminished by losses at the first interface. Real-time frequency domain analysis techniques may become useful for reducing this effect.

The sensitivity of reflection acoustic imaging is superior to the lateral resolution of the technique. For example, 25 μm diameter bond wires are often seen in a 20 MHz pulse echo SAM image at a depth of 1.6 mm when the spot size is significantly larger than 25 μm. An object with lateral dimensions smaller than the spot size that is easily detected in reflection (compared to the noise background) may produce only a negligibly small contrast effect in transmission (compared to a large transmitted signal). However, the apparent size of a small reflector in the image will be determined by the spot size.

Sound is a matter wave and depends on molecular vibrations for propagation. This is why acoustic inspection is much more sensitive to thin cracks than x-ray radiography. Theory predicts that crack openings greater than the particle displacement amplitude produced by the interrogating sound wave should be detectable. Experiments with steels, for example, indicate that air-filled cracks on the order of 10 nm are detectable.[4,42]

Transducer Selection

Since transducer properties play a large role in determining the "effective" resolution (both lateral and axial) that can be achieved, selection of the proper transducer for an application is paramount. Key transducer properties include the frequency distribution of the pulse, lens diameter, and focal length. Typically, best resolution can be obtained at higher frequencies, larger lens diameters, and shorter focal lengths (more critical for high frequency transducers >100 MHz than for transducers <100 MHz). Although higher frequency transducers allow better resolution, they also suffer from higher frequency-dependent absorption in the water path and less penetration into the sample. To illustrate the effect of transducer properties on resolution, consider the 3 images in Figs. 12, 13, and 14. These images were obtained using the identical sample, but 3 different transducers with slightly different pulse center frequencies (f_C) and/or focal lengths (lens diameters were equivalent) as outlined in Table 2.

Table 2. Transducer parameters for the examples shown in Figs. 12 – 14.

	Listed f_C	Measured f_C	Focal Length
T1	110 MHz	75 MHz	10 mm
T2	110 MHz	120 MHz	5 mm
T3	150 MHz	180 MHz	5 mm

The series of Figs. 12-14 shows that as the received center frequency is increased from 120 MHz to 180 MHz, better resolution is achieved and new details become apparent.

Comparisons of Figs. 12 and 13 reveal that transducer T1, having longer focal length than transducer T2, has had significant frequency-dependant absorption of the return signal, resulting in degraded resolution. Although the listed transducer frequency of 110 MHz is identical for both, the "effective" resolution varies significantly, resulting in much better resolution for the transducer having the shorter focal length. Practically speaking, it is not always possible to use the shortest focal length. A 5.9 mm focal length transducer, when focused on a 16 mil thick flip chip, is almost touching the sample surface. Such an analysis requires extraordinary sample alignment for maintaining clearance during scanning. In addition, this working distance is too short for use on

boards or on samples having components at multiple heights (such as a flip chip with capacitors embedded in the substrate). The proper choice of transducer focal length takes into account not only the sample thickness and desired resolution, but also the physical restraints imposed by the scanning set-up. It should be noted that for frequencies below 100 MHz, the water-dependant frequency absorption is less severe. Thus, concerns of focal length effects on resolution are not as critical, and the focal length required can be chosen using sample thickness and set-up constraints as the sole criteria.

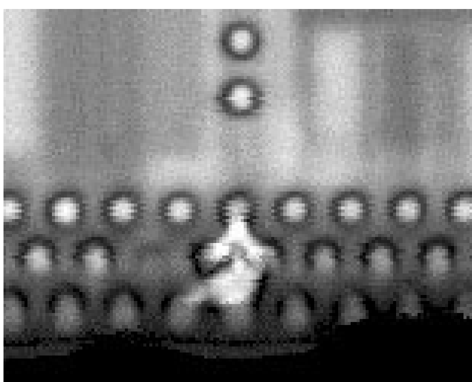


Figure 12: image obtained with transducer T1. Received transducer frequency was measured at 75 MHz.

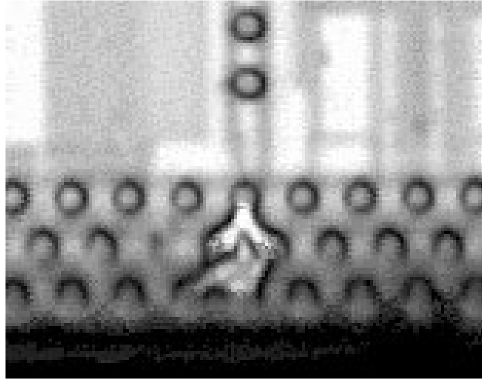


Figure 13: image obtained with transducer T2.

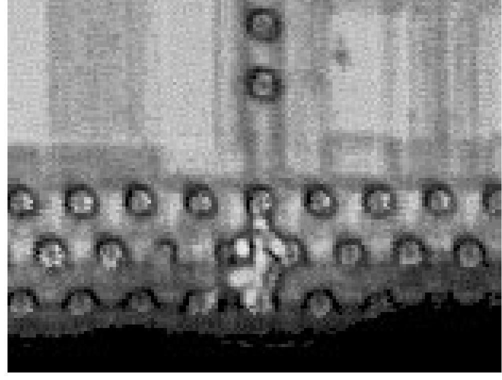


Figure 14: image obtained with transducer T3.

Practical Applications

Plastic Package Inspection

Figure 15 shows two acoustic micrographs of the same 68PLCC at different times. The initial pulse-echo image of this device appears in Fig. 15(a). This device was subsequently saturated with moisture (0.32 wt.%) during 168 hours of exposure to 85°C/85%RH (relative humidity), subjected to vapor phase reflow (VPR) mounting, and imaged again. This post-VPR image is shown in Fig. 15(b). The micrographs in Fig. 15 include both amplitude and phase information. The image of the amplitude of the sub-surface reflection is displayed as a gray scale image on a white background. Delaminated interfaces are designated by total black superimposed over the amplitude image. The delaminated areas were identified by phase analysis of the reflected acoustic pulse.

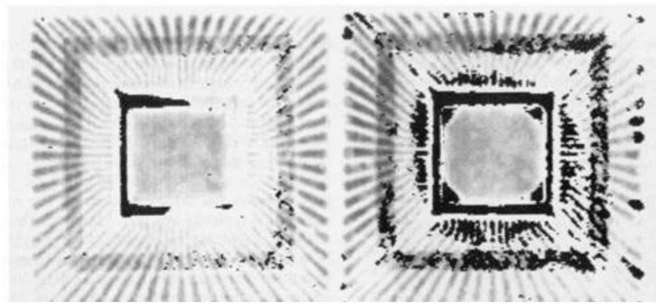


Figure 15: Delamination images of the same 68PLCC taken initially (A) and then after VPR exposure (B). 68PLCCs are 24mm square.

In the initial pulse-echo image in Fig. 15, the only significant delamination appears on the die pad periphery surrounding the die (Ag spot). After VPR, delamination has appeared on the entire die pad periphery, the corners of the die, on the leads (predominantly at the internal terminations), and at scattered locations on the lead tape. Studies have indicated that the primary reliability threat during temperature cycling for this type of package is wire bond degradation due to delamination at the die surface. It is likely that this delamination will spread from the shear stress maxima at the die corners toward the die center during subsequent temperature cycling. Typically, as the delamination spreads, stress-induced damage will occur at the die surface within the shrinking boundary of good adhesion. In the delaminated corners, shear displacement between package and die will damage wire bonds. If delamination at the die surface is initiated during board mounting, it spreads during subsequent temperature cycling and leads to early device failure due to stress-induced damage to the device and wire bond degradation. Also, package cracks and delamination at the leads increase the risk of contamination-related failure of the device.[33,36]

Inspection of Packages Having High Velocity Substrates

Since the acoustic impedances of Cu and Al_2O_3 are so similar, Curve 2 in Fig. 9 can be used to predict reflectivities for die attach inspections in ceramic packages, power packages with

metal heat sinks, or other high velocity substrates such as metal assemblies for microwave components. For example, Fig. 16 is a sketch of a standard TO-220 molded power IC package. The package is characterized by its thick (1.2 mm) Cu heat sink, which is necessary to manage the heat produced by high power output transistors. The die is attached directly to the inside surface of the heat sink with an experimental Pb/Sn die attach system. Figure 17 shows two 50 MHz pulse-echo images of the die attach in the same device at different times. Figure 17(a) is an initial image, while Fig. 17(b) was recorded after 200 cycles of temperature cycle reliability testing. The dark areas denote good adhesion and are areas where most of the energy in the acoustic pulse was transmitted into the die. The bright areas in the die attach indicate almost total reflection due to disbonding. Note the reduction of the area of good adhesion after temperature cycling. A real-time x-ray image was taken after temperature cycling and is shown in Fig. 18. Due to x-ray attenuation in the thick Cu heat sink, the x-ray image required a significant amount of image processing to reveal contrast produced by the die attach layer. The die attach voids seen in the x-ray image taken after temperature cycling agree well with the features in the initial pulse-echo acoustic image. The acoustic image at 200 cycles indicates a reduction in the total area of adhesion that was not detectable by x-ray radiography because no significant increase in x-ray absorption was produced by the thin air gap in the delaminated areas.

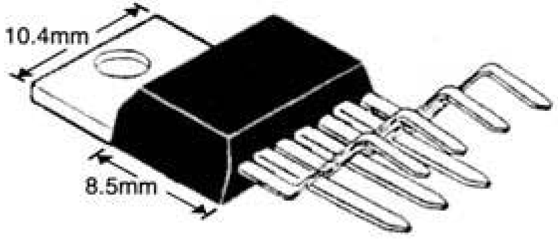


Figure 16: The TO-220 power IC package with Cu heat sink.

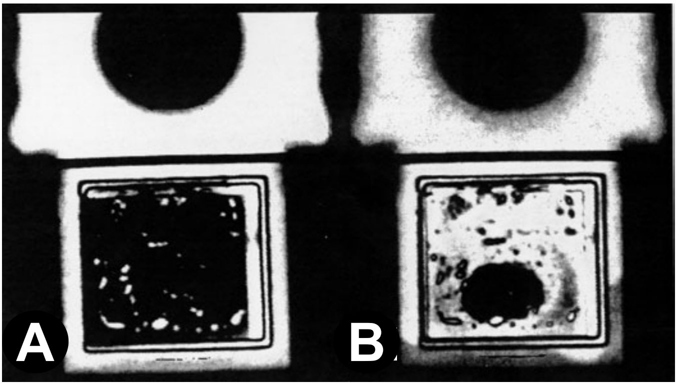


Figure 17: Die attach images in a TO-220 package before testing (A) and after 200 temperature cycles (B). Bright areas indicate a high reflected intensity. Dark areas in the die attach region indicate good adhesion. The heat sink is 7 mm wide at the die attach region.

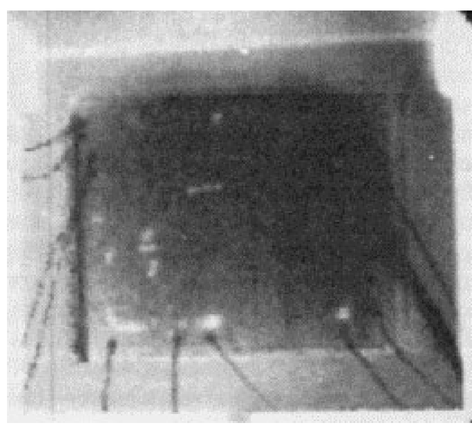


Figure 18: A real-time x-ray image of the TO-220 package shown in Fig.16, taken after temperature cycling.

Strategies for Inspection of Packages Having Thin Layers

The introduction of technologies such as ball grid array (BGA) and flip chip packages has driven improvements in acoustic inspection techniques. BGA packages often have laminated substrates composed of several layers. These packages can not always be reliably inspected by PE-SAM due to echo interference problems from the many thin layers in the package. Polarity analysis, an important tool which assists delamination detection in molded surface mount packages, is not always useful for BGA packages due to very low acoustic impedances of some substrate materials. Additionally, the size and height of features has decreased, requiring increases in frequency to improve depth and spatial resolution. Alternative approaches such as through-transmission screening of BGAs and high frequency (>200 MHz) pulse-echo inspection of flip chip bumps are addressing these new issues.

Initially, the Stanford SAM operated in through-transmission mode, but was later converted to pulse-echo for easier alignment and improved sample flexibility. The advantages offered by pulse-echo acoustic microscopy make it one of the first choices used in beginning an analysis. However, packages such as BGAs and stacked die CSPs have geometries that can result in an inability to temporally resolve echoes, inhibiting the identification of the specific plane where a defect occurs. Additionally, the pulse-echo mode advantage of phase inversion to locate delaminations is often negated in BGA packages by the use of substrates having lower acoustic impedance than the mold compound. These facts combine to make PE-SAM less advantageous for these package types, as well as for certain interfaces that come from thin layers contained within traditional packages. In these instances, TT-SAM is mandatory for accurate defect detection. This important fact is captured in specification IPC/JEDEC JSTD-020, requiring TT-SAM for BGA package inspection. Using the wrong approach can be misleading and result in the identification of actual package fails as passing during package qualification.

TT-SAM Set-up: Whereas in pulse-echo imaging the same transducer is used to both send and receive signals,

transmission analysis requires the placement of a second transducer for receiving signal on the opposite side of the sample from the input signal. Typically, the focused transducer used in pulse-echo analysis is used as the sending transducer in transmission analysis. The receiving transducer most commonly used for IC package transmission analysis is a low frequency, unfocused transducer.

With good set-up, it is frequently possible to obtain TT-SAM images that are nearly as sharp as the PE-SAM image. Consider the PE-SAM image in Fig. 19(a) and the corresponding optimized TT-SAM image in Fig. 19(d). The most critical factor in set-up is establishing the correct focus for the sending transducer. To illustrate this, consider the image in Fig. 19(c), acquired with the sending transducer moved 10 μs out of focus (for reference, 10 μs in water is equivalent to a distance of 7.5 mm). Ideally, the focus should be set at the interface of interest. If this is not possible, then a good starting place is setting the focus at the sample surface.

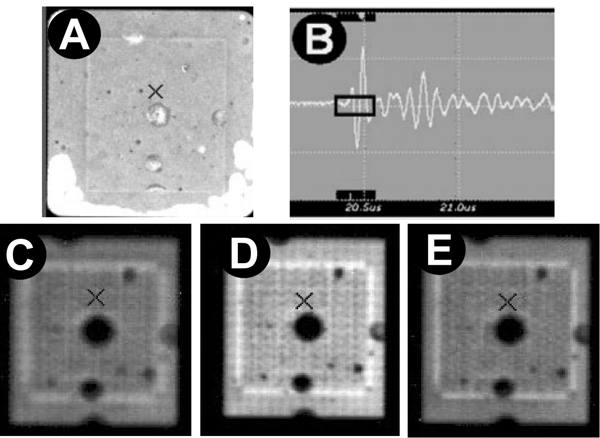


Figure 19: Set-up considerations for TT-SAM. A good pulse-echo focus of the sending transducer must be obtained at the interface of interest, a narrow gate placed only around the first transmitted echo, and the receiving transducer should be as close to the sample as possible. (A) Pulse-echo image with proper focus at lid attach interface. (B) The data gate should be placed only around the first transmitted echo. (C) The TT-SAM image resolution is degraded when the sending transducer is moved 10 μs out of focus). (D) A TT-SAM image obtained with proper set-up. (E) A TT-SAM image made with the receiver is moved 10 μs away from its optimized placement below the sample.

After the focus is established with the sending transducer, the receiving transducer needs to be positioned appropriately. In addition to ensuring the receiver is aligned with the transmitter, the distance of the receiver transducer from the sample is important, even though the receiver transducer does not focus. The best imaging is usually obtained when the receiving transducer is placed as closely beneath the sample as possible. Compare the effect of moving the receiving transducer 10 μs away from its optimized placement by comparing Fig. 19 TT-SAM images (d) and (e). The signal is dimmer and slightly more blurry when the receiver is further away. As expected, moving a focused sending transducer 10 μs out of focus (Fig. 19(c)) has a much larger impact on TT-SAM image sharpness than does moving the unfocused receiver 10 μs away from the sample (Fig. 19(e)).

Proper positioning and set-up of the data gate is also important in transmission analysis. It is best to use a narrow data gate that is placed around the first transmitted pulse to emerge from the package (Fig. 19(b)). Use of a wide gate results in artifacts caused by mixtures of signal that include not only the first transmitted pulse, but also transmitted echoes that have been reverberating inside the sample before finally emerging.

Case 1 – Die Attach Analysis in a BGA Package: Figure 20 shows a pulse-echo image produced during an attempted die attach analysis of a BGA package. The star-shaped pattern of the die pad is evident. Since delaminations appear bright in pulse-echo mode, this image would be incorrectly interpreted as having a large area of delamination between the copper die pad and die attach material.

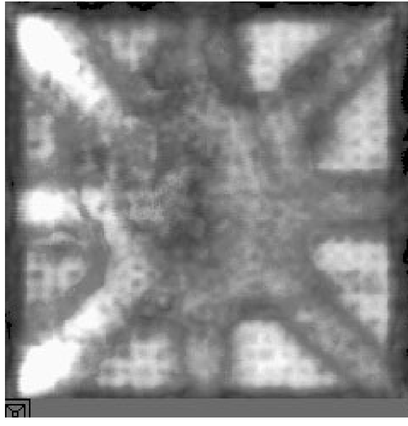


Figure 20: An "interference image" produced by attempted pulse-echo analysis of the die attach in an overmolded BGA package.

The through-transmission image in Fig. 21 shows reliable data indicating only small areas of delaminations are present in the die attach area. However, this image alone does not reveal whether the delamination occurs between the die and die attach, die attach and solder mask, or solder mask and substrate. This is but one example of how real delaminations can be missed in pulse-echo inspection, and similarly bonded areas can appear delaminated, due to echo interferences. This is not necessarily a rare occurrence [49].

Unlike PE-SAM, interface echoes are not analyzed in TT-SAM. Rather, the identification of delaminations is based upon the fact that air gaps (cracks, voids, delaminations) block sound transmission. Thus, a lack of transmission of sound through a package identifies the presence of a delamination. The application of through-transmission acoustic inspection to BGA packages as the initial inspection method both bypasses the echo interference problem and precludes the need for use of phase inversion to detect delaminations. This method has

been applied successfully on virtually every type of BGA in production, including multi-layer packages. However, due to a loss of resolution with this technique caused by scattering from the substrate fibers and BGA balls, it is possible that smaller delaminations may be missed.

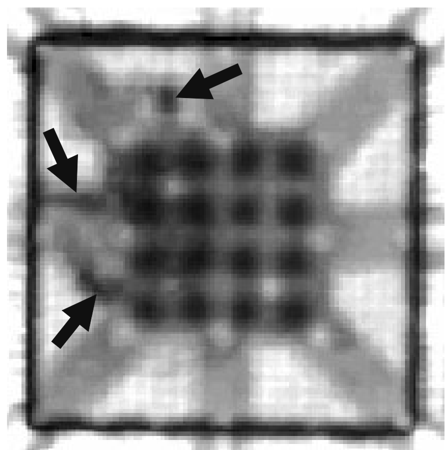


Figure 21: The corresponding through-transmission image for the package in Fig. 14. The 16 solder balls beneath the die are evident, as is the weave pattern in the substrate. Three small delaminations are shown in this image (arrows).

Conventional molded packages, as well as lidded flip chip packages, may also have thin layers present and therefore are subject to the same concerns as outlined for the BGAs. Specifically, the inspection of die attach and lid attach layers of these packages often results in erroneous conclusions when using PE-SAM as the sole inspection criteria.

Case 2 – Die Attach Analysis in a Molded Package: The die attach layer is typically quite thin (50 µm thick or less), making it difficult to cleanly isolate each side of the die attach interface using PE-SAM. Figure 22 shows that when imaging through the heat sink, a 50 MHz transducer captured information at both interfaces of a die attach layer within a molded package. The resulting image is an "interference image". One sign that interference is present is the presence of more than 2 grayscale values in the image. For this particular sample, by applying higher frequency, the two interfaces of the die attach layer can indeed by cleanly resolved, as indicated by the images in Fig. 23 acquired with a 150 MHz transducer.

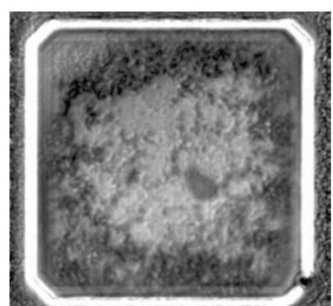


Figure 22: An interference image produced when analyzing the die attach layer using a 50 MHz transducer.

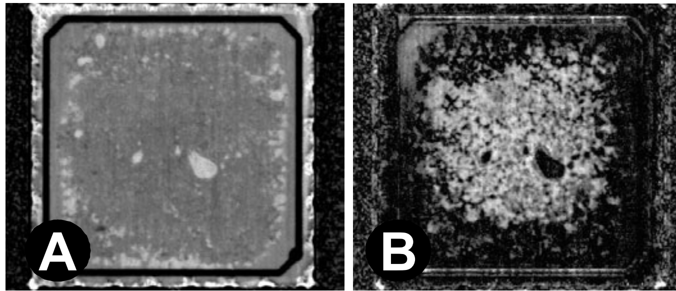


Figure 23: Images acquired with a 150 MHz transducer reveal bright areas indicating voiding and delamination occurring a) in a small amount at the interface between the heat sink and die attach material, and b) in a large amount at the interface between die attach material and the die.

Although the higher frequency could isolate both interfaces of the die attach, in practice, this application of PE-SAM is difficult on this device and requires an experienced analyst to properly interpret the A-scan. Fig. 24 shows why this is so. An extremely large discrepancy in amplitude exists between the two echoes of the die attach. In fact, the die attach/die interface has extremely low amplitude at bonded locations and is so small as to appear as though it is part of the heat sink/die attach interface echo decay. When moving the transducer over a delaminated region, the echo in question has much higher amplitude and is much more easily identified.

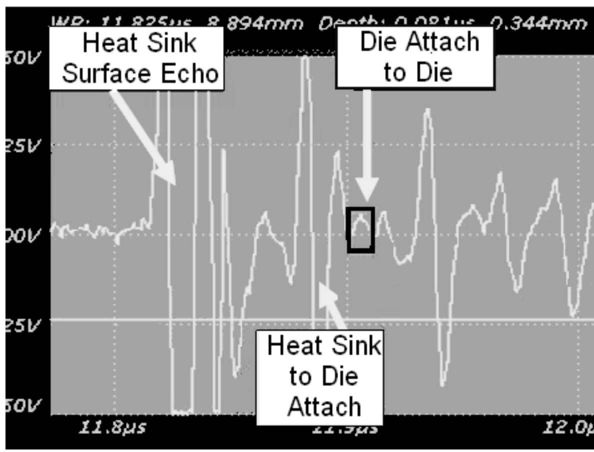


Figure 24: The A-scan from a 150 MHz transducer positioned over a bonded die attach region. The echo from the die attach to die interface is resolved, but extremely attenuated.

If the transducer position over a bonded or delaminated region affects the ability to correctly identify and place the data gate for acquisition of a PE-SAM image, it can be very difficult to properly set up, since usually the presence and location of delaminations are not known in advance. However, by obtaining a TT-SAM image, the precise x-y location of delaminations can be confirmed and can contribute greatly to further analysis by PE-SAM.

Figure 25 is the 50 MHz transmission image obtained from the same part shown in Figs. 23(a) and 23(b). The large area of non-bonded die attach from over the center area of the die

as identified in Fig. 23(b) is corroborated. However, the transmission image cannot corroborate the interfacial failure identified in Fig. 23(b), since it contains no time of flight information. To further confirm a proper application and interpretation of PE-SAM, a cross-section of the device was performed. An image obtained from the center of the cross-sectioned package is shown in Fig. 26. The cross-section validates the application of 150 MHz PE-SAM, showing that indeed, the delamination and voiding occurs primarily at the die attach to die interface.

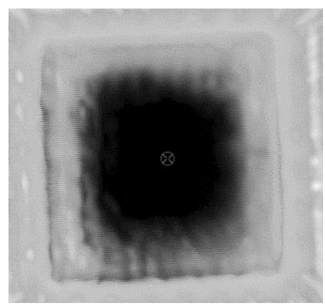


Figure 25: A TT-SAM image shows that the majority of the package area involving the die is unable to transmit sound and is therefore not bonded. This confirms the information obtained from the high frequency PE-SAM image gated at the die attach to die interface.

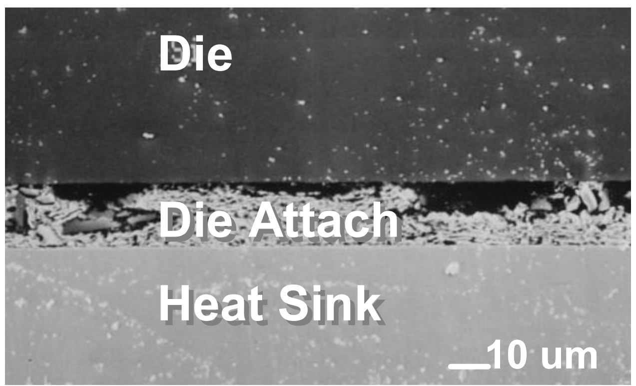


Figure 26: A cross-section confirms the 150 MHz PE-SAM data. The majority of the non-bonded die attach surface is on the die side, rather than the heat sink side.

Case 3 – Lid Attach Analysis in a Flip Chip Package: Just as PE-SAM analysis of die attach can be problematic, the analysis of lid attach bonding on flip chips can be difficult for similar reasons. The lid attach layer is quite thin (~ 50 – 100 µm) and beneath a rather thick lid (1 mm or more). The lid causes attenuation and frequency-dependent absorption of the signal, negatively impacting the ability to clearly isolate the lid attach echoes.

The combination of high frequency and low frequency PE-SAM and TT-SAM is sometimes required to non-destructively obtain a complete picture of the lid attach condition. This is shown in Fig. 27. Use of higher frequency (75 MHz) shows only one side of the lid attach interface. Low frequency (15 MHz) captures information from both sides of the lid attach material and produces variable shading that is difficult to interpret and not suitable for automated calculations based on grayscale values. TT-SAM provides clear information regarding the lid attach integrity at both lid attach interfaces, and produces an image that is suitable for automated void calculations. Evaluated in context with the two PE-SAM images, we can non-destructively deduce that 2 small voids were present between the lid and lid attach material (arrows in figure), while the majority of non-bonded region occurs between the lid attach material and the die.

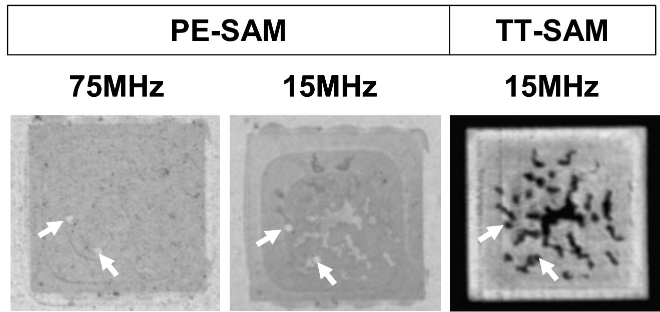


Figure 27: Images from a "thick" package with a thin lid attach interface acquired by both high frequency (75 MHz) PE-SAM, low frequency (15 MHz) PE-SAM, and TT-SAM. The lid is 1 mm thick and the lid attach layer is ~50 µm thick.

Case 4 – Packages with AlSiC Lids: For some lidded flip chip packages, PE-SAM is not an option. This can be the case for packages that have AlSiC lids. Depending on processing, the various components of the AlSiC lid can form a heterogeneous composite that produces multiple internal reflections within the lid itself, due to non-uniform segregation of the various elements. Thus, despite an experienced analyst's best efforts, the PE-SAM analysis will result in confusing, distorted interference images.

Take for example the images from the 3 packages that are shown in Fig. 28. The PE-SAM images from the top row appear different for each of the three packages. Bright areas as well as dark areas are present. This might lead one to believe that each package has some amount of delamination occurring in different patterns. However, an experienced analyst will notice the presence of more than 2 grayscale values in the image, which is often an indicator of interference. Further, if an analyst were to compare the A-scans from a bright and dark region of each image, it would become clear that the PE-SAM images are comprised from very complicated time of flight data that does not match expectations based on the known lid thickness and package construction.

TT-SAM analysis easily works around the PE-SAM problems and provides the desired lid attach information. Generally speaking, the TT-SAM images in the bottom row of Fig. 28 appear similar for all 3 packages and indicate good bonding. Note, however, the small areas of darker regions that occur in the 3 images. It can not be determined based on these images alone whether these features represent defects in the lid attach, or whether defects from the underfill or some other level are affecting the image.

circuitry, and the latest echoes contain the information from the underfill level.

In Fig. 29, SAM images are shown of a flip chip gated at different locations, with the A-scan presented beneath to indicate the gate positions. A narrow gate placed early in the echo packet shows interconnect defects caused by a bump that broke off, damaging the underlying interconnect layers in the process. The image from this gate position shows no other defect. However, a narrow gate placed late in the echo packet shows the absence of a solder at the location the bump broke off, and reveals a large area of excessive solder bridging several bumps. If an analyst attempted to use a wide gate encompassing the entire echo packet, as is often done for lower frequency imaging of molded packages, the resulting image shows indications of both defects, but the image clarity is degraded and the nature of the defect is not clear.

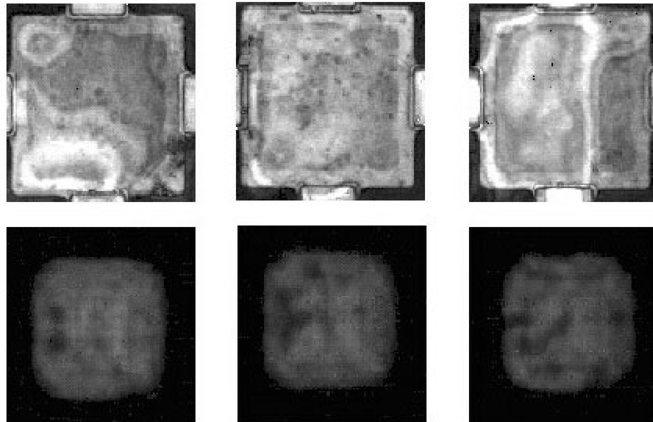


Figure 28: Images acquired by PE-SAM (top row) and TT-SAM (bottom row) of 3 packages with AlSiC lids. Lid material inhomogeneity prevents a valid PE-SAM analysis. The TT-SAM images in the bottom row show generally good bonding.

Imaging Flip Chip Devices

The development of flip chip technology introduced the need for high frequency non-destructive imaging of flip chip bump interconnects and underfill defects. Transducers with improved lens designs and center frequencies above 200 MHz have been developed specifically for flip chip applications. These transducers allow excellent imaging of flip chip bumps and have detected defects such as non-wets of solder, missing bumps, solder bridging, and damaged die-level interconnect cracks.

Data Gate Set-up: For ultra high frequency PE-SAM of flip chips, choosing the appropriate data gate length and placement is critical. Even when the gate encompasses the echo of interest, if the gate is not set correctly, the resulting images may lack the physical defects of interest within the sample.

Figure 29 illustrates the effects that can occur in an image as the gate position is slightly changed or the gate is lengthened. Typically, the interface between the die and the underfill does not produce a single echo, but produces multiple echoes which are overlapping, even with frequencies >200 MHz. The echoes are produced at each interconnect layer within the die, which typically all blend together to form interference patterns of the die circuitry. There is another echo produced at the interface of the die's protective overcoat (PO) and the typical polyimide coatings applied to flip chips at relatively thick dimensions (on the order of 3 μm thick or more). Lastly, an echo is produced by the interface of the polyimide to the underfill layer that is used to help neutralize stresses produced by CTE mismatches between the die and the package substrate.

In flip chip inspection, since the die is mounted upside down with the circuitry facing the substrate, the earliest signals in the packet of echoes that first emerges after the signal passes through the bulk Si of the die contain information from the die

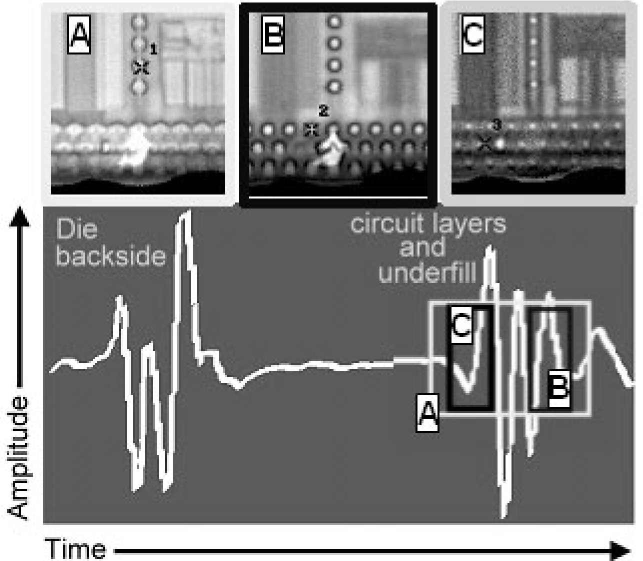


Figure 29: The effect of gate size and placement on the resulting image is demonstrated: A) image acquired with wide gate B) image with "late" narrow gate C) image with "early" narrow gate.

If the echoes from these interfaces were perfectly separated and isolated from each other, gate placement could easily be determined based on calculations that incorporate the structure of the die and package, and the velocity of sound in the various layers. However, since the echoes are overlapping, gate placement is qualitatively determined. Computer and software improvements have enabled multiple gate scanning capability, which greatly aids the determination of proper gating for flip chips.

This concept operates on the principle that an initial data gate can be sized and placed at the beginning of the echo packet. Then, the gate is automatically duplicated and placed at set intervals across the entire echo packet. When the sample is scanned, multiple images are simultaneously produced from the many gates. The gate position used for a particular image can easily be visualized so there is no guessing as to which

gating conditions produce which type of image. The only drawback to this application is that typically, all images produced by a multiple gate scan are saved in a single file, which results in extremely large file sizes (40 Mb or higher for a single package scan).

Thus, in practice, a multiple gate scan is extremely effective to establish the proper single gate set-up condition for each new flip chip device, which is then applied to all subsequent samples of the same type. This is easily done for example, if one is working on process development for underfill, where the only information desired is how well the underfill flowed, whether filler particle settling occurs, and whether capture voids are present. In some instances, however, the use of multiple gate scanning must be performed on every sample. This is typically the case if one is trying to discriminate between defects in the interconnect (especially a concern with Cu/low-k dielectric devices), failures at the PO to polyimide interface, and failures between the polyimide and underfill. For example, Fig. 30 shows images from a multiple gate scan that differentiate defects between the PO and polyimide, and the polyimide and underfill. With the proper imaging conditions, many types of defects in flip chips can be detected. Unless specifically stated otherwise, all high resolution flip chip imaging discussed throughout this section was performed using an ultra high frequency (UHF) transducer having a 150 MHz center frequency.

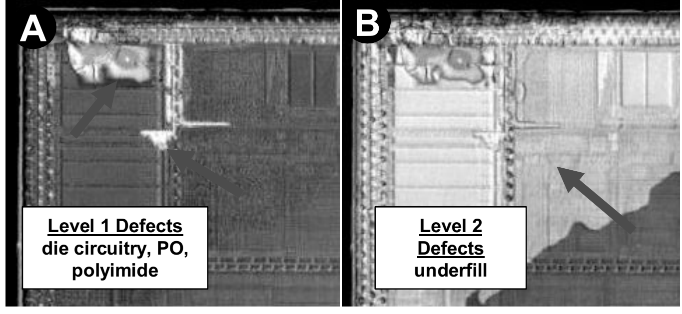


Figure 30: Use of a multiple gate scan aids in diagnosing two types of fail modes. (A) device level defect are found at an early gate placement (B) underfill to die surface delamination at a later gate placement.

Setting the Focus: Normally, the pulse-echo focus is set by watching the echo of interest, and varying the distance between the transducer and the sample such that the maximum amplitude of the echo is reached. With the UHF analysis used for flip chips, this focusing mechanism does not work. This is because water strongly attenuates the high frequency signal. As the transducer is moved closer to the sample, the water path shrinks, so there is less absorption and the signal increases in amplitude. One can move the transducer past the focal point, and there will be no indication of having moved past focus, because the signal only continues to increase.

This is illustrated in Fig. 31. Here are shown 2 A-scans captured from the same x-y location on a device, but with the 180 MHz transducer placed either at the focal point, or moved closer towards the sample and past the focal point. The signal strength from the out of focus placement is significantly increased compared with the in-focus setting, at both the front surface and sub-surface echoes.

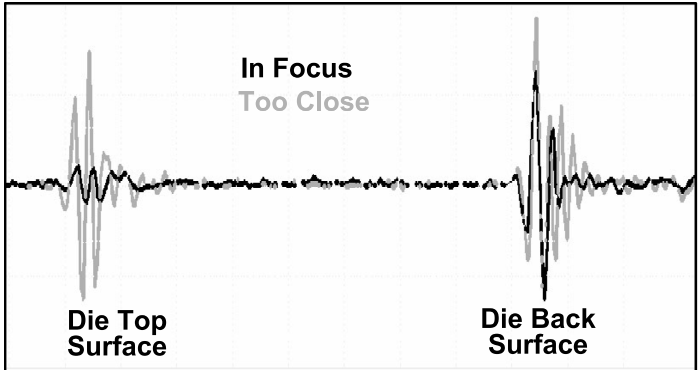


Figure 31: Using signal amplitude for focus determination of UHF transducers can be misleading. A 180 MHz transducer was positioned over a flip chip at the proper pulse-echo focus position. It was then moved through the focus, closer to the die. The out of focus position produces signals of the highest amplitude.

Usually, the UHF transducers are designed to target a water path length of minimum working distance from the sample, in order to minimize the water absorption effects. The typical die thickness is taken into account during this calculation. So for a given transducer and given die thickness, once the focus setting is known, it can be applied uniformly to all flip chip die of the same thickness. Sometimes, however, the focus must be determined empirically. This may be necessary for example, if an analyst receives a flip chip for SAM analysis that has been thinned to facilitate other failure analysis procedures such as photoemission microscopy. The thinner the die, the further away the transducer will be from the sample at focus, and the weaker the signal amplitude. To identify the proper focus setting, it is best to run a series of high digital resolution focus scans, where the transducer position is varied by set intervals. An example of images from such a series is shown in Fig. 32. In this case, focus increments of 0.5 μs work very well. A focus series should always be obtained by starting with the transducer in the lowest position (closest to the sample), and then moving it away from the sample.

Case 1 – Solder Bridging: Figure 33 shows acoustic detection of solder bridging in a delidded flip chip. While the bridging is detected acoustically, it is difficult to conclusively say bridging is the cause of the light and dark stripes connecting the bumps. This is because other defects can cause a similar appearance, such as corroded metal in the die-level interconnect. The bridging is confirmed by X-ray analysis. The contrast variation in the SAM-detected solder bridging is caused by variation in the solder thickness. X-ray is not as sensitive to the thickness variation, and all bridging bumps have a uniform appearance. X-ray is a powerful non-destructive imaging tool that is complementary to SAM.

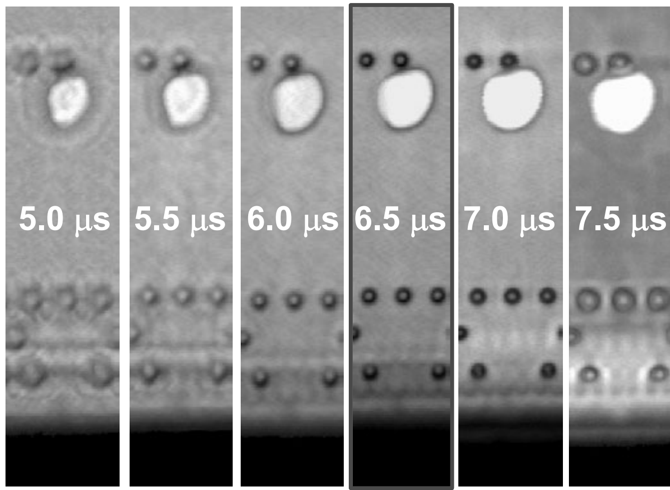


Figure 32: Using signal amplitude for focus determination of UHF transducers can be misleading. One workaround is to perform a high digital resolution focus series such as shown here. The numbers refer to the front surface reflection water path time of flight. For this die thinned to 480 μm, best focus occurs when the 180 MHz transducer is positioned so the front surface echo occurs at 6.5 μs.

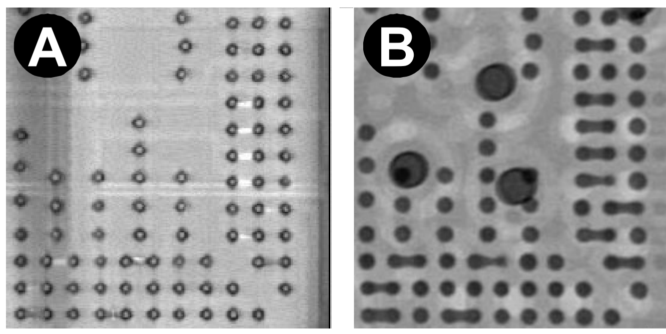
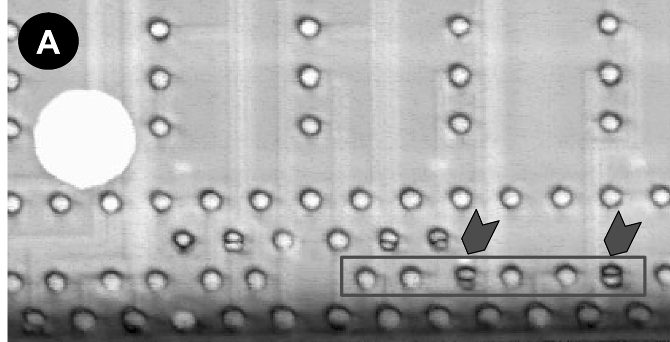


Figure 33: SAM imaging shows shorted flip chip bumps (A). The contrast variation of white or dark in the bridging area in the SAM image is caused by solder thickness variations. X-ray confirms the shorting defect (B).

Case 2 – Non-wet Bumps: In many cases, it is possible to detect flip chip bumps that have not properly wet to the die surface. Figure 34 is one such example. The high frequency PE-SAM inspection produces an image where some of the bumps take on a "snowman" appearance. The cause of this appearance is determined by a physical cross-section of the bumps in question. Incomplete wetting of solder to the bump pads on the die has occurred. In this specific instance, underfill flowed into the space that should have been occupied by solder. If air were present instead, a large amplitude reflection would have been received at these locations, similar in brightness to the large circular underfill delamination visible in the SAM image.

Case 3 – Defects in Die Interconnect: In addition to revealing bump defects, acoustic analysis of flip chips is able to detect large-scale damage that has occurred within the die circuitry layers. This is demonstrated within Fig. 35. A device failing for shorts at a specific bump was submitted for analysis. One of the images produced by a multi-gate SAM scan shows a bright reflection associated with a specific bump. This indicates a likelihood of delamination and/or cracking beneath the bump. One of the other images produced by the multi-gate scan clearly shows a crack surrounding a portion of the bump periphery. A cross-section into this periphery region of the bump shows cracking in the dielectric layers, which has now been filled in with Cu, causing the short. The subsequent failure of the same bump on two other devices allowed the identification of a bad socket on a burn-in board as being the root cause of the shorts failure.



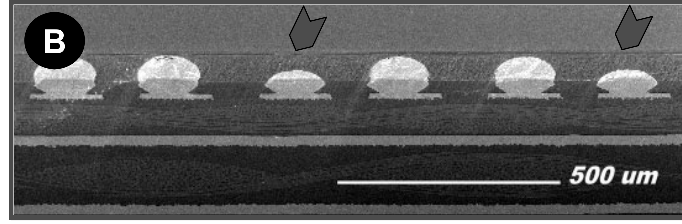


Figure 34: Flip chip images showing bump defects. (A) UHF PE-SAM image. Defective bumps have a "snowman" appearance (arrows). (B) SEM image of cross-sectioned bumps corresponding to the bumps outlined by the box in "A". Arrows point to the same bumps in both "A" and "B". The cross-section identifies non-wetting of the solder to the die as the cause of the "snowman" bumps in the SAM image.

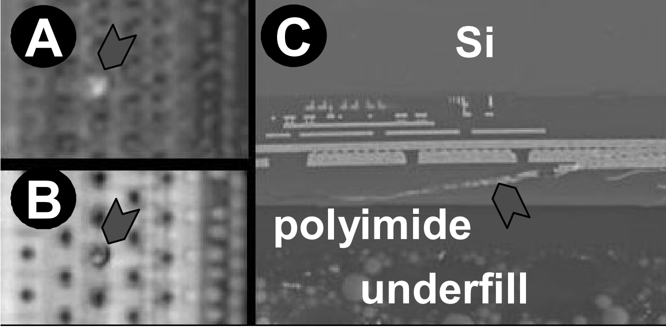


Figure 35: A bump anomaly detected by UHF PE-SAM (A: arrow). A slight shift in the data gate produces a SAM image that reveals cracking just outside the bump perimeter (B: arrow), confirmed in an SEM cross-section (C: arrow).

Case 4 – Delamination of Low-k Dielectric Layers: Beginning since the late 1990s, the semiconductor industry has been working on converting material sets, replacing Al and silicon dioxides with Cu metal and either polymer or C-doped oxide (CDO) low-k dielectrics. The initial introduction of low-k dielectrics posed new challenges for developing robust devices. Typically, the adhesion of these dielectrics to other materials of the die was suboptimal during early learning cycles. The industry has learned that with the appropriate process improvements, strong adhesion could be obtained. This is critical to enable packaging and good reliability for Cu/low-k products.

One assembly step in particular that highlighted weak adhesion was the saw process. Mechanical sawing initiates cracks within the scribe street of all devices manufactured today. With good adhesion, the cracks do not pose a problem and crack growth is sufficiently arrested by optimally designed scribe seals. When suboptimal adhesion exists, the cracks grow into delaminations that propagate through the scribe seal when subjected to conditions that stress CTE mismatches, such as temperature cycles.

Figure 36 shows the PE-SAM result when this type of defect is present in a flip chip. Depending on gate settings, whether water enters the crack, and other factors, the delaminated area may or may not be bright. It is typically located at the die edge and propagates into several rows of bumps. One noticeable feature that frequently occurs and helps distinguish the low-k dielectric delamination from underfill adhesion issues is that typically, if underfill delamination occurs, the bumps are still clearly observable. On the other hand, as Fig. 36 shows, low-k delamination often results in the disappearance of the bumps from the image.

PE-SAM analysis has proven an extremely valuable tool during Cu/low-k device development, enabling optimization of assembly processes for low-k devices, and enabling the development of process improvements for better adhesion in low-k die.

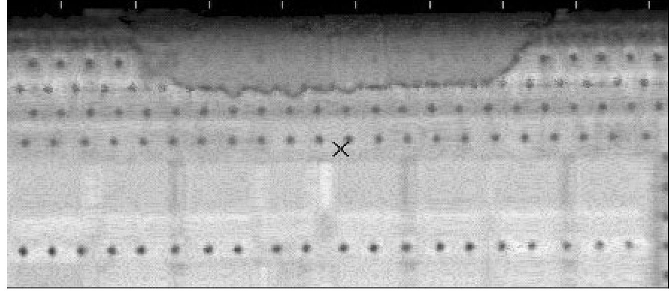


Figure 36: SAM image of a flip chip with cracking in the low-k dielectric layers. Often, a defect in the low-k layers results in the loss of bumps from the image.

Case 5 – Bump Size Variations: Although SAM is excellent for detecting non-wet bumps, solder bridging, and cracking beneath the bump, some types of bump defects are not readily observable with SAM. This fact is highlighted by Fig. 36. SAM easily detects broken bumps and massive solder bridging, as discussed previously and again shown in this figure. However, Fig. 36 also highlights a weakness of SAM as applied to flip chip inspection. The SAM image does not produce an image of the whole bump, but mostly just the bump to die contact area. This can be partly explained by the spherical shape of the bumps, which reflects sound away from the transducer. The end result is that bumps with inadequate solder volumes appear the same size in a SAM image as perfectly normal bumps. So SAM does not work so well for showing bump size defects.

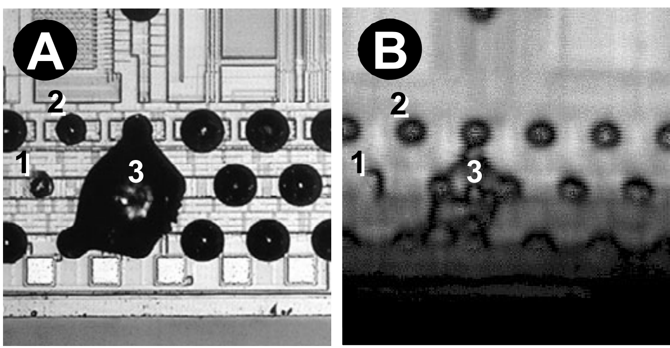
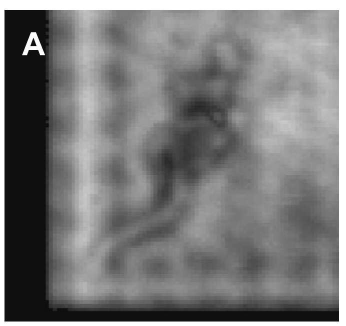


Figure 37: Optical (A) and UHF PE-SAM (B) image comparison shows SAM Capability for 3 types of defects. Broken bump (1) and massive solder bridging (3) are clearly evident in the SAM image. Solder bumps of reduced size (2) are not detected by SAM. The "edge effect" phenomenon, which interferes with imaging the outermost row of bumps, is observed in this SAM image.

Figure 37 is also a good example of the "edge effect" phenomenon, which limits the ability to image features that are along the edge of a sample. This is caused by the reflection of a portion of the conical signal away from the focus near the edge. The thicker the die, or the larger the numerical aperture of the lens, the worse the edge effect. For flip chips, bumps within 200 μm of the edge are difficult to image using most high frequency transducers available commercially. The development of higher power transducers that maintain high resolution at longer working distances may enable better edge imaging.

Case 6 –Lidded Flip Chip Packages: To allow acoustic inspection of flip chip bumps and underfill, the backside of the die must be exposed. This reduces the absorption of the high frequency sound waves in the lid, which would degrade the resolution that is needed for bump inspection. For over-molded flip chips, the mold compound above the die must be removed, usually by grinding and polishing. For lidded flip chips, there are various methods available for lid removal. All methods, however, pose the risk of inducing damage during handling. This is especially a concern for devices containing low-k dielectric materials, which require careful handling to avoid artifacts. By performing TT-SAM analysis prior to any sample preparation, one can ensure detected defects are real and not a result of handling. Figure 38 shows a portion of a flip chip package imaged with TT-SAM prior to delidding, and with ultra high frequency PE-SAM after lid removal.

Three points can be made with this figure. First, one must remember TT-SAM provides information from all interfaces. If performing lid attach analysis, be aware that an observed defect in a TT-SAM image may in fact come from another layer. Second, the TT-SAM image provides convincing evidence that the underfill defect is not an artifact of the delidding process. This tactic is especially valuable when evaluating packages containing Cu/low-k die. Third, the TT-SAM resolution is much worse than the resolution obtained in the UHF PE-SAM image. Thus, very small defects will not be confirmed by TT-SAM. Note that although a defect is identified from TT-SAM alone, its location is not truly revealed until the PE-SAM is applied.



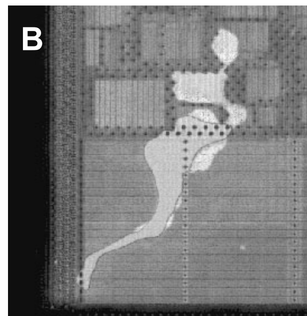


Figure 38: A flip chip was imaged by TT-SAM analysis prior to removing the lid (A) and then imaged post-lid removal using ultra high frequency PE-SAM inspection (B).

Figure 39 again highlights the first point made in the above paragraph. TT-SAM images contain information from the entire sample. This includes features that exist on the outside of the package, as well as from anywhere in the interior. In this example, residue that has collected during autoclave testing on the outside of the package is affecting the TT-SAM image, causing some regions to transmit less strongly than they should. It is easily confirmed that the residue on the outside of the package is the cause of the reduced transmission. By taking an eraser and scrubbing the package substrate, the residue is removed. A new TT-SAM image shows much more even transmission across the entire package.

Current Challenges

Stacked Die Packages

Form factor considerations and the need for increasing functionality have driven the development of stacked die packages for certain markets. In particular, stacked die packages are increasingly used in hand-held wireless communications, such as cell phones. These packages stack 2 or more functional die on top of each other. Package interconnects may be mixed, using both wirebond and flip chip die in the same package. The packages are required to be thin, so the die are regularly background to very thin dimensions (<150 μm thick each). Some of the more integrated packages currently being developed are reported to use 6 or more stacked functional die, each 60 μm thick. At least one major semiconductor manufacturer has announced the ability to create 7 to 8 die stacks, each only 50 μm thick.

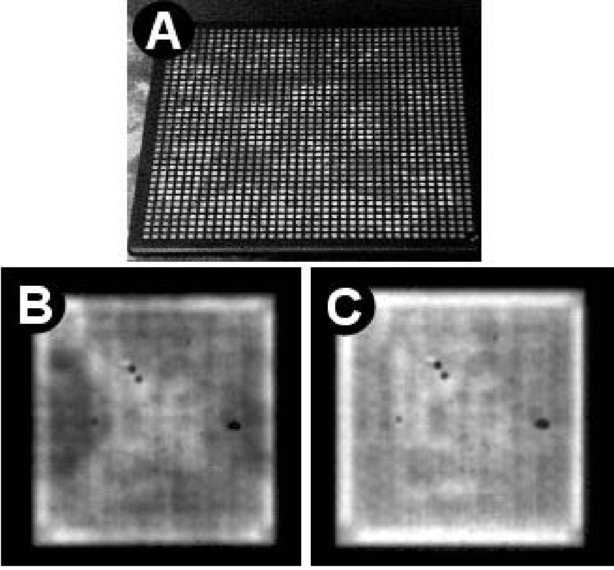


Figure 39: A JEDEC tray pattern is outlined by residue that accumulated during autoclave testing and observable optically when the part is tilted at an angle (A). The pattern affects the TT-SAM image (B). After removing accumulated residue with an eraser, the TT-SAM image in (C) is obtained.

Stacked die packages are difficult to image using SAM if one is interested in obtaining information beyond the top die surface. In addition to the large number of thin layers, the Si die itself is thin, and the velocity of sound is quite high in Si (~8500 m/s). The higher the acoustic velocity, or the thinner the layer, the closer the echoes will be. When both happen together, the echoes between the die can not be resolved. Given the acoustic velocity of Si above, the first sub-surface echo from a 100 μm thick die should appear 20 ns after the front surface echo. At this close spacing, typically it is difficult to resolve both sides of a single thinned die, because the die backside echo partially overlaps the front-side echo.

To better understand the imaging challenges of stacked die packages, consider the simplest structure available. Figure 40 shows the general structure of a two die stacked package, with die adhered by epoxy and no silicon spacers present. In order to non-destructively evaluate adhesion between the two die, SAM needs to be able to resolve the echo from the surface of the lower die.

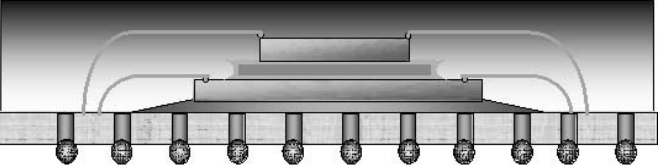


Figure 40: A general diagram of the most simple stacked die package in production today, having only 2 stacked die. Total package thickness is <1 mm. Each die is <150 μm thick.

A stacked die package similar to the cartoon in Fig. 40 was analyzed by SAM. Since the mold compound strongly attenuates the signal, ultra high frequency inspection could not be applied. A high numerical aperture 50 MHz transducer was used to form the image. To obtain as much information as possible, 3 types of PE-SAM images were obtained (peak amplitude, phase inversion, and time of flight). In addition, a TT-SAM image was acquired. These four images are shown in Fig. 41.

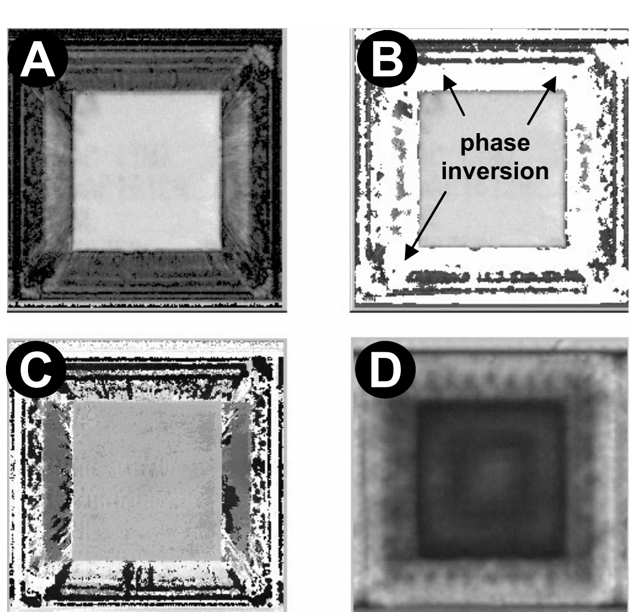


Figure 41: SAM analysis showing 4 types of images acquired from a 2-die stacked package and focused at the top die surface. (A) Peak amplitude image (B) Phase inversion image (C) Time of flight image (D) Transmission image.

The peak amplitude image indicates the die sizes. The top die is almost square and the bottom die is rectangular. The top surface of the bottom die is somewhat obscured by the wires bonded to the edges of the top die. The phase inversion image indicates the possibility of delamination (red areas), both over the substrate as well as the surface of the bottom die. However, the known low impedance of substrates, combined with the known artifacts that occur over non-planar features such as wires, requires caution to be used. The only conclusive finding from the phase inversion image is that the upper die top surface is 100% bonded to the mold compound. The time of flight image confirms the location of features above the upper die surface, such as wires (brighter), and below the upper die (darker), such as the lower die. The transmission image confirms the fallacy of the phase inversion image, and clearly shows good transmission throughout the entire package.

To obtain a better idea of the behavior of the acoustic signals as they pass through the package, a B-scan was acquired across the line indicated by the short arrows in Fig.42. The B-scan in this figure shows strong signals from the top surface of the package, and surfaces of the upper die, the lower die (but only in areas not covered by the upper die), and the substrate. Signals from the bond wires are also visible.

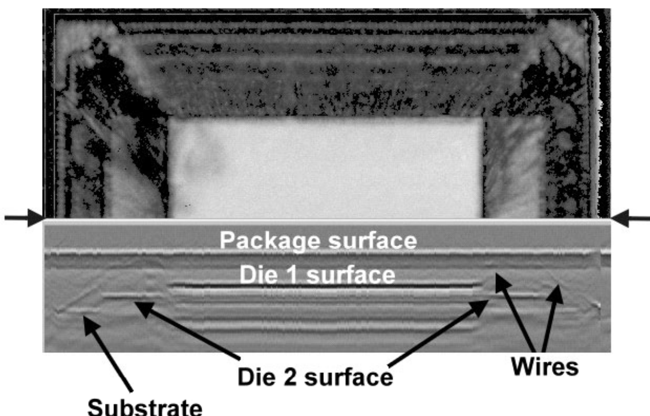


Figure 42: A B-scan of a stacked die package reveals the acoustic signal pattern when only 2 die are present. Significant ringing occurs beneath die 1, which prevents identification of the die 1 to die 2 interface.

However, there is no obvious signal that corresponds to the interface between die 1 and die 2. The die 2 signal visible in the B-scan occurs away from the center, where the signal goes straight through mold compound to die 2. However, in the center of the package, the sound must travel through mold compound and then through die 1 before it reaches die 2. Because the acoustic velocity of Si is much higher than mold compound, this means that over the center of the package, the echo corresponding to die 2 should occur sooner than it occurs in the regions where only mold compound is above die 2. The absence of a distinct signal in this expected location indicates that either the echo from die 2 is overlapping with the die 1 echo, or it has been severely attenuated. The subsequent echoes below the die 1 surface are due to ringing of the signal between the package surface and die 1, and also due to mode-converted shear waves from the die 1 surface. It would be easy to view this B-scan and assume many die were in this package, had it not been known previously that only 2 die are present.

An x-ray image from this die (Fig. 43) shows that the lower die is wire bonded only on the left and right edges, while the upper die is wire bonded on all 4 sides. The Si die are transparent in x-ray. From the SAM and x-ray images combined, it would be possible to deduce that only 2 die are in this package.

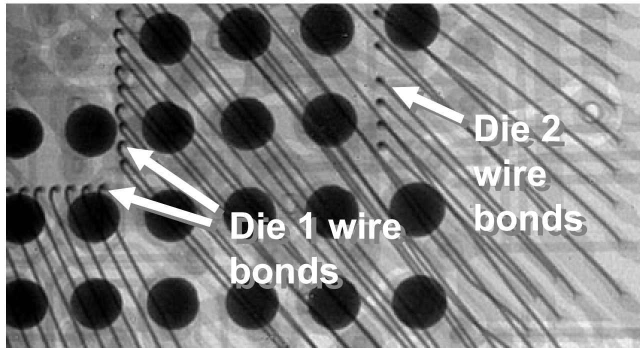


Figure 43: An x-ray image of a stacked die confirms the lower die is not wire bonded on all edges.

It is worthy to note that newer features recently available, such as frequency domain imaging, help identify defects when imaging through the substrate of a stacked die package.[59] Defects isolated by this approach are likely associated with the bottom-most die.

Complementary Nondestructive Techniques

SAM is one of three complementary techniques used in the nondestructive inspection of packaged ICs. A transmission acoustic technique called scanning laser acoustic microscopy (SLAM), and x-ray radiography are also used for this application (Figure 45). These two techniques can each produce images of packaged ICs, and therefore have partially overlapping capabilities. However, unique capabilities of each technique make them complementary for the nondestructive inspection of IC packages.

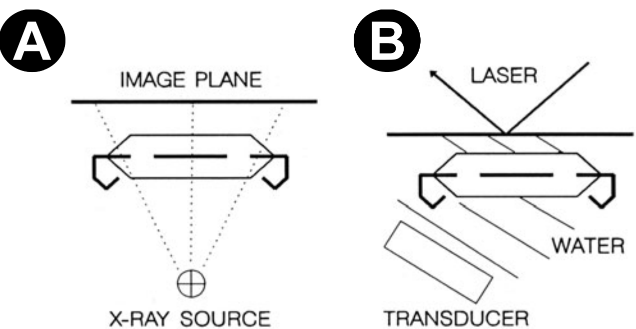


Figure 44: X-ray radiographic inspection (A) and scanning laser acoustic microscopy (SLAM) (B) of IC packages.

Real-time x-ray inspection offers the best lateral resolution and is unsurpassed for nondestructive wire sweep inspection. Images are produced at TV rates and the sample is easily manipulated. Although commonly used for die attach inspection, x-ray imaging shows only the location of voids in the die attach layer, and not the total area of attachment. X-ray inspection often exhibits limited contrast for low mass thickness die attach layers on metal heat sinks (see Fig. 18). Similarly, detection of package voids can be limited by lead frame shadowing.[32] X-ray is an excellent technique for detecting voids in flip chip bumps, but is not useful for detecting underfill voiding or delamination in flip chips.

X-ray laminography is a tomographic technique that produces an image of x-ray attenuation at a specific plane within the sample. X-ray laminography has seen limited use for component level inspection, but has been applied to the inspection of solder joint quality for surface mount process control.[43,44]

SLAM is a transmission acoustic technique developed in the early 1970's at Zenith and later transferred to Sonoscan in 1974.[45] SLAM incorporates concepts similar to those presented by Sokolov in 1936 for his proposed acoustic microscope.[46] In SLAM, a planar sound wave is transmitted through the sample onto a mirrored coverslip. The disturbances produced in the coverslip by the transmitted wave are imaged with a scanned laser beam. The technique produces either an acoustic shadow image or an interference pattern by comparison of the signal to a phase reference. SLAM offers the advantage of real time imaging which is more suitable for 100% inspection than a technique involving mechanical scanning. SLAM has been successfully applied to situations where real time imaging is critical, and in applications involving samples with irregular shapes and very thin multi-layer construction that cannot be imaged well by reflected sound due to the "dead zone" and reverberation effects described earlier. SLAM applications have included the inspection of tape automated bonding (TAB) packages, and thin film ceramic chip capacitors.[9] In practical application the spatial resolution of SAM images of IC packages is typically superior to that of SLAM images at the same frequency.[47] SLAM cannot be applied to die attach inspection in packages with air cavities, such as ceramic DIPs and pin grid arrays, unless the lid is removed. Similar information as provided by SLAM can be obtained from a SAM operating in through-transmission mode.

New Developments

For features smaller than 20 μm, another method to temporally resolve echoes is required. With the emerging availability of streamlined algorithms, fast and inexpensive PCs, and dedicated digital signal processor chips (DSPs), it becomes feasible to apply frequency-domain signal analysis to routine inspection. In addition to enabling inspection of future packages having much smaller features than exist today, this type of analysis will also assist inspection of current packages where overlapping echoes from finely spaced layers is today a problem. For example, often it is difficult to discriminate between the signal received from a bonded layer, or the signal received from a water-filled delamination. Also, at present bondline thicknesses beneath lids on flip chip packages can not be measured acoustically due to absorption of the high frequencies required for resolution by the thick copper lid. Development of real-time frequency-domain signal analysis may offer a non-destructive way to obtain these measurements.

The availability of high speed processors, inexpensive DRAM, and large capacity hard drives is also enabling analysis methods whereby after acquiring an image with one scan, the size and position of data gates can be re-set and through software analysis, the image is adjusted accordingly. This "virtual scanning" will enable re-examination of devices after they are no longer available (i.e. time zero data can be revisited after devices have been put through environmental testing), will allow engineers to verify the proper set-up was used for data acquisition, and will be a remarkable teaching tool to demonstrate to novice acoustic microscope users how positioning the data gates influences the resulting image.

Summary

The SAM used today for IC inspection is a hybrid instrument with characteristics of both the Stanford SAM and the C-scan

recorder.[9,10] For inspection in layered IC packages both amplitude and phase are measured and used to produce images of the internal structure and locations of interfacial delamination. Precise scanning and high frequencies are used for optimum lateral resolution. In addition, the depth of the primary sub-surface reflection is recorded for pseudo-three-dimensional representations of the package.

Acoustic microscopy is well suited to plastic package inspection. Although the plastic mold compound strongly attenuates sound and thereby limits resolution, the planar, layered design of IC packages is ideal for acoustic inspection. The typically featureless and planar front surface facilitates sub-surface imaging.

The acoustic microscope has been a key factor in the understanding of the moisture sensitivity problem in surface mount plastic IC packages. Polarity analysis of reflected acoustic pulses is extremely useful in the detection of delaminations in plastic IC packages, and is an important advantage of inspection with reflected sound. By quantifying limits on the amount of delamination that is acceptable from a reliability standpoint, SAM inspection can be incorporated into standard test methods and is preferred over destructive cross sectioning and electrical test methods.

The nondestructive nature of acoustic inspection means that in reliability evaluations involving a dense matrix of variables, fewer devices are required. Previously, a fraction of the remaining devices would have to be sacrificed for destructive analysis at each inspection interval. Also, the data of reliability evaluations using SAM are more valuable because the initiation and propagation of damage can be tracked on individual ICs throughout the test.

As packages have continued to evolve with time, adjustments to the application of acoustic microscopy and improvements in resolution have successfully overcome analysis challenges and ensured that SAM continues to be a non-destructive technique of significant importance for the development and evaluation of IC packages. Specifically, analysis of BGA packages in through-transmission mode bypasses two potential problems that exist for pulse-echo mode: echo interferences from the multi-layered substrates, and reduced acoustic impedances of the substrates that make phase inversion identification of delamination invalid. Transducers with center frequencies over 150MHz deliver the resolution necessary for flip chip bump inspection.

Due to frequency-dependant attenuation, the desired feature size of interest has become smaller than SAM can resolve with available transducer technology. Frequency-domain analysis may partially overcome this limitation and is currently being developed.[48,56] A one-dimensional or two-dimensional scanned array technology for acoustic imaging of IC packages would have a dramatic impact on the effectiveness of SAM for real-time applications such as process control by eliminating the need for mechanical scanning. Such an “acoustic camera” would not have to offer the same analytical capabilities of a modern SAM to be effective as an on-line screening tool.[57]

References

1. T. Moore, Proc. Int. Symp. Testing and Failure Analysis, 1989, pp. 61-67.
2. Y. Bar-Cohen and A.K. Mal, in Metals Handbook, 9th edn., ASM International, 17, 1989, pp. 231-277.
3. R.C. McMaster, Nondestructive Testing Handbook, Vol. 2, Ronald Press, 1959, p. 43.
4. J. Szilard, in Ultrasonic Testing, J. Szilard (ed.), John Wiley and Sons, 1982, pp. 1-23.
5. C.F. Quate, A. Atalar, H.K. Wickramasinghe, Proc. IEEE, 67, 1979, pp. 1092-1114.
6. R.A. Lemmons and C.F. Quate, App. Phys. Lett., 25, 1974, pp. 251-253.
7. R.A. Lemmons and C.F. Quate, Proc. 1973 IEEE Ultrasonic Symp., 1973, pp. 18-21.
8. C.F. Quate, IEEE Trans. Sonics and Untrason., SU-32 (2), 1985, pp. 132-135.
9. L.W. Kessler and S.R. Martell, Proc. Int. Soc. for Testing and Failure Anal., 1991, pp. 491-504.
10. B.T. Khuri-Yakub, in New Technology in Electronic Packaging, B.R. Livesay and M.D. Nagarkar, ASM International, 1990, pp- 311-315.
11. J.L. Rose and P.A. Meyer, Mater. Evaluation, 31 (6), 1973, p. 109.
12. G.J. Curtis, in Ultrasonic Testing, J. Szilard (ed.), John Wiley and Sons, 1982, pp. 495-555.
13. N.J. Burton and D.M. Thacker, Proc. Int. Soc. for Testing and Failure Analysis, 1985, pp. 187-192.
14. K. Shirai, K. Kobayashi, T. Noguchi and T. Goka, Proc. Int. Soc. for Testing and Failure Analysis, 1988, pp. 47-52.
15. M.J. Mirasole, Proc. Int. Soc. For Testing and Failure Analysis, 1988, pp. 77-88.
16. R.A. Lemmons and C.F. Quate, Appl. Phys. Lett., 25 (5), 1974, pp. 251-253.
17. R.G. Wilson, R.D. Weglein and D.M. Bonnell, Semiconductor Silicon 1977, 1977, pp. 431-435.
18. C.F. Quate, Semiconductor Silicon 1977, 1977, pp. 422-430.
19. A.J. Miller, Inst. Phys. Conf. Ser. No. 67: Section 8, 1983, pp. 393-398.
20. A.J. Miller, Acoust. Imaging, 12, 1982, pp. 67-78.
21. H.K. Wikramasinghe, J. Micros., 129, 1983, pp. 63-67.
22. H.R. Vetters, et al., Scanning Electron Microwscopy/III, 1985, pp. 981-989.
23. M. Sakimoto, et al., Proc. Int. Symp. for Testing and Failure Anal., 1985, pp. 173-177.
24. A. Kitayama, H. Tabata and H. Suziki, Proc. IMC, 1986, pp. 462-469.
25. S. Ito, et al., Proc. Elect. Comp. Conf., 1986, pp. 360-365.
26. S. Okikawa, et al., Proc. Int. Symp. for Testing and Failure Anal., 1987, pp. 75-81.
27. S. Kuroki and K. Oota, Proc. Elect. Comp. Conf., 1989, pp. 885-890.

28. A. Nishimura, S. Kawai and G. Murakami, Proc. Elect. Comp. Conf., 1989, pp. 524-530.
29. T.M. Moore, Texas Instruments Technical Report TR-088778, Feb. 1988.
30. L.W. Kessler and S.R. Martell, Proc. Int. Symp. for Testing and Failure Anal., 1990, pp. 491-504.
31. R. Birudavolu, Proc. Surface Mount 1989, pp. 751-766.
32. A. van der Wijk, K. van Doorselaer, Proc. Int. Symp for Testing and Failure Anal., 1989, pp. 69-74.
33. K. van Doorselaer and K. de Zeeuw, Proc. Elect. Comp. Conf., 1990, pp. B49-B53.
34. T. Moore, R. McKenna, S.J. Kelsall, Proc. Int. Symp. Testing and Failure Analysis, 1990, pp. 251-258.
35. T. Moore, R. McKenna, S.J. Kelsall, J. Surface Mount Tech., 4 (3), 1990, pp. 31-38.
36. T. Moore, R. McKenna, S.J. Kelsall, IEEE Int. Reliability Physics Symp., 1991, pp. 160-166.
37. K.R. Kinsman, J. Metals, 40 (6), 1988, pp. 8-13.
38. IPC-SM-786, "Impact of Moisture on Plastic Package Cracking," and IPC-Test Method 650-2.6.20, "Plastic Surface Mount Component Cracking," Institute for Interconnecting and Packaging Electronic Circuits (IPC), Lincolnwood, IL, 1991.
39. L.A. Kinsler, et al., Fundamentals of Acoustics, John Wiley and Sons, 1982, p. 125.
40. R.S. Gilmore, R.A. Hewes, L.J. Thomas, and J.D. Young, in Acoustical Imaging, 17, H. Shimizu, N. Chubachi, and J. Kushibiki (eds.), Plenum Press, 1989, pp. 97-109.
41. B.T. Khuri-Yakub, in New Technology in Electronic Packaging, B.R. Livesay and M.D. Nagarkar (eds.), ASM International, 1990, pp. 311-315.
42. J. Krautkramer and H. Krautkramer, Ultrasonic Testing of Materials, Springer-Verglag, 3rd edn., 1983, p. 28.
43. B. Baker, Electronic Manufacturing, Feb. 1989, pp. 20-22.
44. C. McBee, Circuits Manufacturing, Jan. 1989, pp-67-69.
45. L.W. Kessler, Proc. IEEE, 67, 1979, pp. 526-536.
46. S. Sokolov, USSR Patent No. 49, Aug. 31, 1936.
47. R.K. Mueller and R.L. Rylander, IEEE Spectrum, 1982, pp. 28-33.
48. T.M. Moore, Reliable Delamination Detection by Polarity Analysis of Reflected Acoustic Pulses, Proc. Int. Symp. For Testing and Failure Anal., 1991, pp. 49-54.
49. T.M. Moore and C.D. Hartfield, Proc. Characterization and Metrology for ULSI Technology, NIST, 1998.
50. U.S. Patents 5641906, 5641906, “Apparatus and method for automated non-destructive inspection of integrated circuit packages”
51. Moore, T.M., and Hartfield, C.D., “Package Analysis SAM and X-Ray”, in Failure Analysis of Integrated Circuits: Tools and Techniques, p. 43-54, L C. Wagner (ed.), (Kluwer Academic Publishers, MA) Chapter 3 (1999)
52. Moore, T.M., and McKenna, R. Characterization of Integrated Circuit Packaging Materials, Butterworth-Heinemann, London, 1993.
53. Moore, T.M., and Hartfield, C.D. (1997) Through - Transmission Acoustic Inspection Of Ball Grid Array (BGA) Packages. Proceedings 23rd International Symposium for Testing and Failure Analysis. 197- 204
54. Plikat, B. (2000) Contrast Inversions in Scanning Acoustic Microscopy (C-SAM) of Glue Die Attach. Proceedings 26th International Symposium for Testing and Failure Analysis. 293 – 302
55. Canumalla, S. (2002) A Broadband Model for Ultrasonic Pulses in the Presence of Thin Layers in Microelectronics. Proceedings 28th International Symposium for Testing and Failure Analysis. 235 – 244
56. J.E. Semmons and L.W. Kessler, Proc. Int. Soc. for Testing and Failure Anal., 2002, pp. 55-59.
57. Lasser, R. et al., On-Line Large Area Ultrasonic Imaging Using Ultrasound Camera Technology, presented at Advancements in Ultrasonic Analysis Symposium 2001, Sonix, Inc. Springfield, Va.
58. IPC/JEDEC J-STD-020B: Moisture/reflow sensitivity classification for non-hermetic solid state surface mount devices.
59. F.B. Aspera and C. Flores, Frequency Domain Imaging: Acoustic Microscopy Technique for Die Stacking Application (to be published, ISTFA 2004).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor
Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Electronic Package Fault Isolation Using TDR

D. Smolyansky
TDA Systems, Portland, OR, USA

Introduction

Time Domain Reflectometry (TDR) measurement methodology is increasing in importance as a non-destructive method for fault location in electronic packages [1-4]. The visual nature of TDR makes it a very natural technology that can assist with fault location in BGA packages, which typically have complex interweaving layouts that make standard failure analysis techniques, such as acoustic imaging and X-ray, less effective and more difficult to utilize [5].

In this paper, we will discuss the use of TDR for non-desctructive package failure analysis and fault isolation work. We will analyze in detail the TDR impedance deconvolution algorithm as applicable to electronic packaging fault location work, focusing on the opportunities that impedance deconvolution and the resulting true impedance profile opens up for such work. We will discuss the place of TDR in the overall failure analysis process, and present examples of proper fault isolation techniques.

TDR Fundamentals

TDR was initially developed for fault location of long electrical systems such as cables. Currently, high-performance TDR instruments, coupled with add-on analysis tools, are commonly used as the tool of choice for failure analysis and signal integrity characterization of board, package, socket, connector and cable interconnects. Such high-end TDR equipment is currently available from two manufacturers – Agilent and Tektronix.

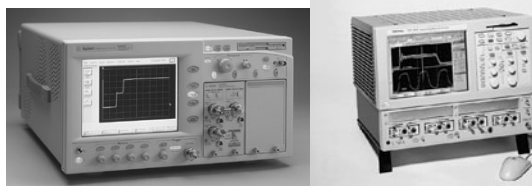


Figure 1. *TDR instruments from Agilent and Tektronix*

Based on the TDR impedance measurements, the designer can perform signal integrity analysis of the system interconnect, and the digital system performance can be predicted accurately. A failure analyst can use TDR impedance measurements to locate an interconnect fault more accurately and quickly, allowing the analyst to focus on understanding the physics of the failure at this failure location. A typical system required for such work will consist of a TDR oscilloscope, a probing or fixturing setup (for example, from Cascade Microtech), and analysis software, such as IConnect® from TDA Systems.

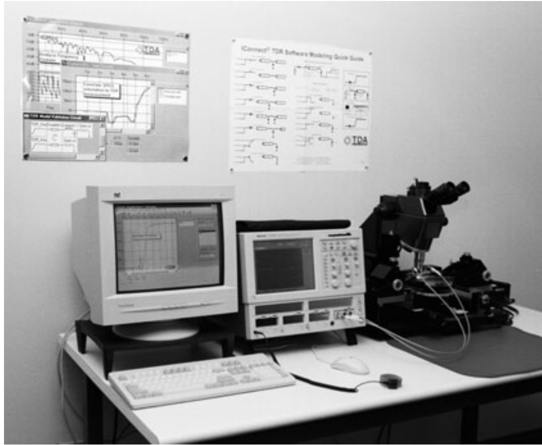


Figure 2. *A typical failure analysis setup includes a TDR instrument, a probing or fixturing setup, and analysis software*

The TDR instrument is a very wide bandwidth equivalent sampling oscilloscope (18-20 Ghz or even more) with an internal step generator. It is connected to the Device Under Test (DUT) via cables, probes and fixtures. It delivers a fast rise time to the DUT, and based on the reflection from the DUT the failure analyst can perform the fault isolation analysis of the DUT.

TDR is very similar to X-ray and acoustic imaging techniques in that it sends the signal to the DUT and looks at the reflection to obtain the information about the DUT. The difference between X-ray or acoustic imaging and TDR is in the type of signal and the type of propagation media for the signal. X-ray and acoustic imaging use X-ray and acoustic stimuli correspondingly, propagating through the free space to the DUT, whereas TDR uses fast-electrical-step stimulus, delivered to each trace in the DUT via electrical cables, probes, and fixtures.

Table 1: *Comparison between TDR, SAM and X-Ray failure analysis techniques*

	TDR	SAM	X-ray
Stimulus type	Electric	Acoustic	X-ray
Stimulus delivery medium	Electrical wires	Water	Air
Direct contact required?	Yes, signal and ground	No	No
Output presented for analysis	Package trace reflection profile	Optical image	Optical image
Ability to locate failures between package or board layers	Good	Poor	Poor

A direct electrical contact between the TDR instrument and the DUT is required to perform the measurement. In addition, not only the signal, but also the ground contact must be provided in order for the TDR signal to provide meaningful information about the DUT. Without a good ground contact, the TDR signal will not have a good current ground return path, and the TDR picture will be extremely hard to interpret.

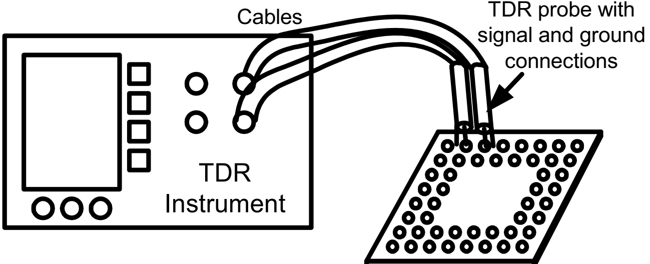


Figure 3: *TDR is connected to the DUT via cables, probes and fixtures. A direct electrical contact to the DUT is required for both signal and ground pins of the TDR probe.*

Because of the wide bandwidth of the oscilloscope, and to ensure that this bandwidth and fast rise time can be delivered to the DUT, one must use high-quality cables, probes, and fixtures, since these cables, probes and fixtures can significantly degrade the rise time of the instrument, reduce the resolution, and decrease the impedance measurement accuracy.

A typical TDR oscilloscope block diagram is shown in Figure 4 below. The fast-step-stimulus waveform is delivered to the DUT via electrical cable, probe, and fixture interconnects. One way to think of the incident TDR step is as a wave front propagating through the interconnect and reflecting back from the discontinuities. The superposition of all the wave fronts, reflected from all discontinuities, is what is displayed on a TDR oscilloscope.

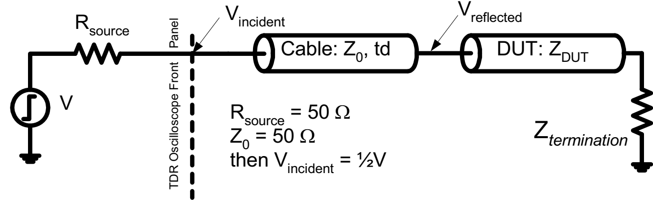


Figure 4: *TDR oscilloscope equivalent circuit*

The waveform reflected from the DUT is delayed by two electrical lengths of the interconnect between the DUT to the TDR oscilloscope, and superimposed with the incident waveform at the TDR sampling head (*Figure 5*). The incident waveform amplitude at the DUT is typically half the original stimulus amplitude (V) at the TDR source. The smaller DUT incident waveform amplitude is due to the resistive divider effect between the 50 Ohm resistance of the source and 50 Ohm impedance of the coaxial cables connecting the TDR sampling head and the DUT.

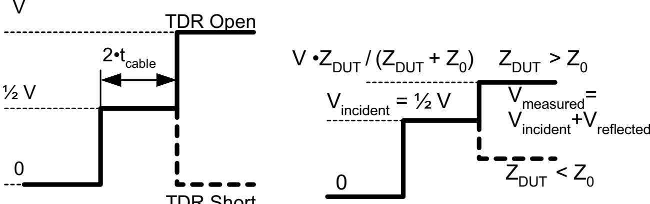


Figure 5: *The incident waveform is delayed by twice the length of the interconnect between the DUT and the TDR oscilloscope, and is divided in half by the resistor divider effect between resistance of the TDR source and resistance of the interconnect to the DUT*

Equivalent resistance of the TDR source R_{source} defines the characteristic impedance of the measurement system. Since R_{source} is 50 Ohm for high-performance TDR instruments available today, using non-50 Ohm cables and probes can produce confusing results. Unlike with a regular oscilloscope, no active probes or resistor divider probes are allowed for use with TDR.

TDR does not provide an optical image of the package, but rather an electrical signature of the trace in the package. Because of the nature of the information that TDR provides, it is important to be aware of typical TDR signatures that correspond to simple package failures, such as a short or an open connection (*Figure 6*).

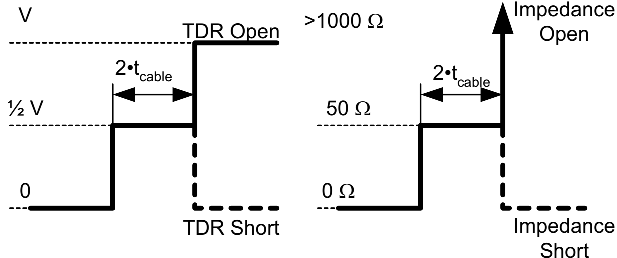


Figure 6: *Open and short connection TDR and impedance profile signatures. V is the full voltage amplitude of the TDR step source;* t_{cable} *is the electrical length of the cable and probe interconnecting the TDR oscilloscope and the DUT*

Note that everything in TDR is round trip delay. This applies not only to the cable, probe and fixture interconnecting the TDR oscilloscope to the DUT, but also to all delay measurements on the DUT itself. In order to obtain an accurate delay readout, the designer has to divide the measured delay by 2.

After the round trip delay of the cable, the voltage reflected from the DUT arrives back to the oscilloscope and is added to the incident voltage on the oscilloscope to produce the measured voltage values. The oscilloscope then converts these voltage values into the values for reflection coefficient and impedance. It is the impedance and delay that the failure analyst is most interested in, and the accuracy and resolution of the impedance and delay measurements is what determines the accuracy of fault isolation.

The faster the rise time that the TDR interconnect can deliver to the package under test, the smaller the size of the discontinuities that can be resolved with a TDR oscilloscope. Available TDR instrumentation provides very fast rise times; reflected signal rise times of the order of 25-35 ps can be observed at the TDR oscilloscope. However, poor quality cabling and fixturing can quickly degrade the TDR instrumentation rise time and decrease the instrument resolution.

It is important to wear good personal ESD protection when working with high-performance TDR oscilloscopes. TDR instruments are ESD-sensitive, high-precision and high frequency instruments. Personal ESD protection (e.g., anti-static strap connected to the instrument) will protect the instrument while maintaining its performance. Add-on ESD modules in the signal path will degrade the rise time of the instrument and degrade its resolution. It is also important to discharge possible charge accumulated on your probe or cable before making the connection to the DUT.

In addition to measuring impedance, the TDR oscilloscope is capable of providing L, C and R signatures for the DUT. For example, an experienced TDR user can, without difficulty, recognize a “dip” in a TDR waveform as a shunt capacitance, and a “spike” as a series inductance. Any L and C combination can also be represented as shown in *Figure 7*.

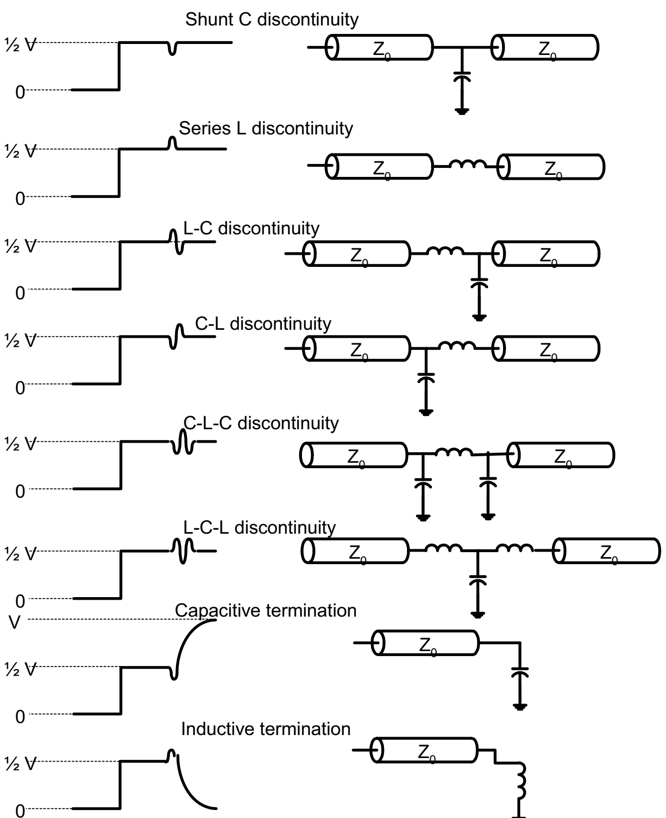


Figure 7: *Visual lumped (LCR) interconnect analysis using TDR*

A series C or a shunt L, however, will represent a high-pass filter for the TDR signal, and the resulting reflection from the elements beyond such series C or shunt L can not be interpreted without prior knowledge of the DUT topology.

Additional information about TDR measurement technology and TDR oscilloscopes can be found in references [6] through [8].

TDR Multiple Reflection Effects

One of the limitations of TDR is the effect of multiple reflections, which is present in multi-segment interconnect structures, such as an electrical package. The accuracy of the DUT signature observed at the TDR oscilloscope is dependent on the assumption that at each point in the DUT, the incident signal amplitude equals the original signal amplitude at the probe-to-DUT interface. In reality, however, at each impedance discontinuity, a portion of the TDR incident signal propagating through the DUT is reflected back, and only a portion of this signal is transmitted to the next discontinuity in the DUT. In addition, the signal reflected back to the scope may re-reflect and again arrive at the next discontinuity at the DUT. These so called “ghost” reflections are illustrated on the lattice diagram in *Figure 8*.

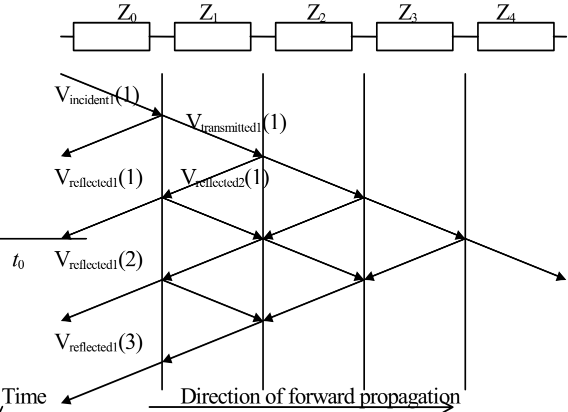


Figure 8: *Lattice diagram of TDR waveform propagating through a DUT with multiple impedance discontinuities*

As a result of these re-reflections, the signature of the DUT becomes less clear, and additional processing is required using the impedance deconvolution algorithm ([9] and [10]), which is currently not available in TDR oscilloscopes. The impedance deconvolution algorithm deconvolves the multiple reflections from the TDR waveform and provides the *true-impedance-profile* for the DUT, significantly improving the clarity of the DUT signature and simplifying further analysis of the TDR data.

For example, if a failure analysis technician were looking for an open failure in an electrical package, TDR data by itself would not have been sufficient to locate the position of the failure (*Figure 9*) as there appears to be multiple potential fault locations. The true-impedance-profile provides an exact location of the open in the DUT whereas the TDR waveform

by itself provides confusing information about the location of this open. In addition, the impedance profile, being an exact signature of the DUT, is relatively easy to correlate to different layers in a BGA package. Such correlation is practically impossible with a TDR waveform alone.

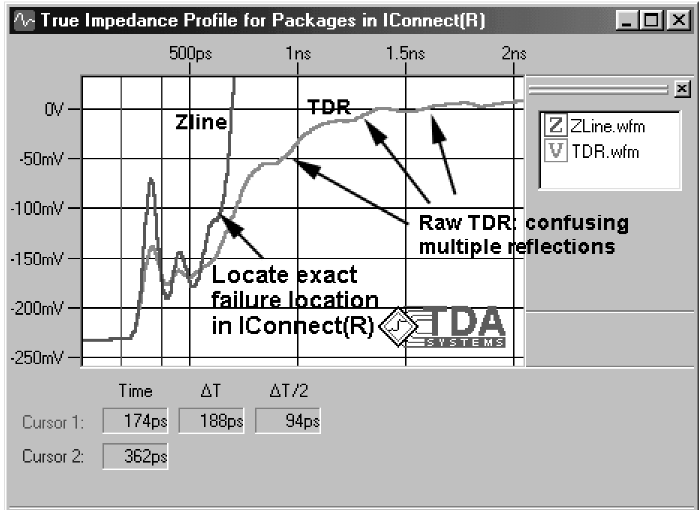


Figure 9: *True-impedance-profile vs. the raw TDR waveform for a BGA package. The true-impedance-profile provides much more accurate information about the failure location.*

An additional advantage that the true-impedance-profile provides is that it is very easy to evaluate capacitance or inductance of an impedance profile segment using the following equations:

$$C = \frac{1}{2} \cdot \int_{t_1}^{t_2} \frac{1}{Z(t)} dt \quad L = \frac{1}{2} \cdot \int_{t_1}^{t_2} Z(t) dt \qquad (1)$$

The type of discontinuity (inductive or capacitive) that we observe in the impedance profile, can also be easily identified — “dips” in the impedance profile correspond to the capacitive discontinuity, and “peaks” correspond to inductive discontinuity. Being able to estimate the value of capacitance or inductance for any given segment can be a significant help in understanding which package segment is being analyzed and in locating the failure more accurately.

Before discussing package failure analysis techniques using TDR in further detail, it is imperative to note the importance of obtaining a good quality TDR measurement and a clean impedance profile. Without a good TDR measurement for the DUT and the reference, the true-impedance-profile is likely to be computed incorrectly, and both TDR data for the DUT and the true-impedance-profile will provide a confusing picture.

TDR Resolution and Rise time

The issues of TDR resolution are often misunderstood or misrepresented, because the TDR resolution is believed to be completely governed by the following rule of thumb. Two small discontinuities, such as two vias in a PCB, can still be resolved as two separate ones, as long as they are separated by at least ½ the TDR rise time:

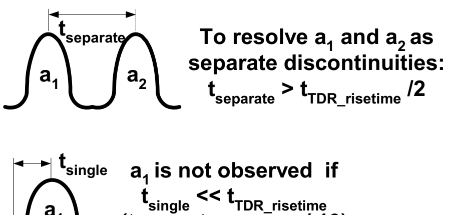


Figure 10: *TDR resolution rules of thumb*

If these two vias are not separated by half the TDR rise time as it reaches the vias, they will be shown by TDR as a single discontinuity. Assuming we use good cables, probes, fixtures, and we can deliver the full 30-40ps rise time of the instrument to the discontinuities in question, the minimal physical separation between these vias will be 15-20ps. For FR4 board material with dielectric constant E_r=4, this results in 2.5-3mm (0.1”) resolution. Often this number (or other similar calculation) is quoted as the TDR resolution limit.

However, in real-life situation, the designer typically is looking to ***observe or characterize a single discontinuity***, such as a single via, or a single bondwire in a package, rather than several of such vias or bondwires! In this case, the above rule is totally irrelevant, and TDR can allow the designer to observe discontinuities of 1/10 to 1/5 of the TDR rise time, bringing the numbers above to 5ps or less than 1mm (25milliinches) range (Figure 10b).

Furthermore, there are well-developed ***relative TDR procedures*** for observing and characterizing even smaller discontinuities, such as the golden device comparisons for failure analysis [1-4].

In addition, faster TDR modules are available from Picosecond Pulse Labs, which makes it easier to resolve some of the finer discontinuities and faults.

Figure 11. *Picosecond Pulse Labs fast TDR add-on module*

TDR Measurements of "Splits" and "Stubs"

If the package trace under test splits into two or more directions, the TDR instrument shows the sum of all reflection from all the N legs in the split, but cannot separate which reflection came from which leg in the split. This means the failure location in case of the traces with splits or stubs can be extremely challenging.

If the splits are of the same impedance and delay (as sometimes is the case in a "star" interconnect topology), they can be simply represented by transmission lines running in parallel, and the impedance measured by the TDR oscilloscope equals Z_1 / N, with the delay of each trace being equal to the delay measured by TDR (Figure 12).

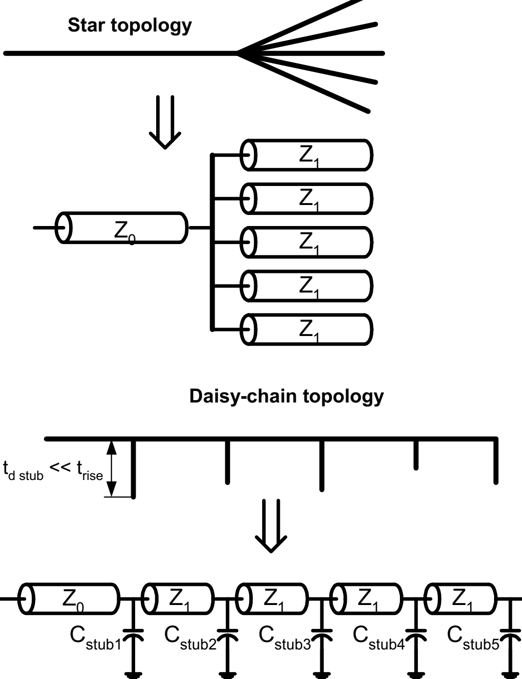


Figure 12. *Taking TDR measurements of splits in a star topology and stubs in a daisy-chain topology*

In case of a stub (which often takes place in a daisy chain configuration), if the length of each stub on the main bus is much shorter than the rise time of the signal propagating through the bus, the stub can be treated as lumped capacitances loading the main bus, thus simplifying the measurement problem.

TDR Probing and Fixturing

TDR is delivered to the DUT via electrical cable, probe, and fixture interconnects. The quality of these interconnects is the key to obtaining a good measurement. As noted before, poor quality cabling and probes can degrade the TDR rise time and decrease the resolution of the instrument. In addition, when computing the impedance profile, it is necessary to have a clean reference short or open waveform; without a good reference, we are not likely to get a clear signature of the DUT. Because of these factors, good quality microwave probes and cables are required to obtain a good quality TDR measurement.

Fixtures, probes, and probing stations for package failure analysis work are available from various manufacturers. A full-featured failure analysis probing station can provide easy viewing and access to the package with a probe, and enable a failure analyst to perform at-temperature analysis of the package failures.

TDR cables and probes will degrade the rise time of the signal measured on the TDR oscilloscope approximately as follows:

$$t_{measured} = \sqrt{t_{TDR}^2 + 2 \cdot \left(\frac{0.35}{f_{3dB}}\right)^2} \qquad (2)$$

where t_{TDR} is the rise time measured on the TDR scope with no cable connected, and f_{3dB} is the 3dB bandwidth of the cable and probe. The factor of 2 in this equation is due to the fact that the signal has to take a roundtrip through the cable before it is observed and measured on the oscilloscope. Specifying a cable with a 3dB bandwidth (f_{3dB}) of about 10 Ghz for the scope with its own rise time of 30ps, will result in the rise time at the cable end of about 58ps. Specifying 3dB bandwidth of 17.5 Ghz will give the rise time end of the cable of about 40ps.

In an application where the TDR cable length can be limited to less than 2 ft, requesting a "lowest-loss" flexible cable from your favorite high quality low cost coaxial cable manufacturer would be sufficient. If you are working with a 3-4 ft cable, however, or require full resolution and rise time that the oscilloscope can offer, you will have to work with a high-end microwave cable manufacturer. Semi rigid cables can provide better performance than flexible ones, but are more difficult to use. SMA connector is commonly used in TDR cables, since it provides acceptable performance, and can be mated directly to the 3.5mm connector found on 20Ghz TDR sampling modules. For even faster rise time, a higher bandwidth microwave connector, such as 2.92 or even 2.4 may be required. [1]

When using a probe for taking a TDR measurement on a package, the designer has to define a ground location *near* the signal location. If such ground location is not available, or if the spacing from signal to ground varies widely across the PCB, the designer may have to use a probe which has a long ground wire, or a variable length wire. For a probe with a long ground wire, the parasitic inductance will be very large, and will not allow the designer to obtain a good quality TDR measurement. Variable length ground wires, and variable pitch (signal-to-ground spacing) probes do not provide sufficient measurement repeatability, and will not provide accurate impedance measurement results or signal integrity interconnect models.

---

[1] 3.5mm connector is specified to operate to a 26.5 Ghz, whereas a typical SMA is rated to 12.5 or 18 Ghz. 2.92 mm connector is specified to 40 Ghz, and 2.4mm to 50 Ghz.

In many cases, a simple, inexpensive and convenient TDR probe can be obtained by using a 3 inch length of semi-rigid coaxial cable with an SMA connector, exposing the center conductor of such cable, and either using the sleeve of the semi-rigid coax as the ground contact, or attaching a ground wire. Using different diameter coax will result in different probe pitch, and making the center and ground conductors shorter or longer can provide the right trade-off between convenience of use and performance.[2] Such probes are also available commercially, *Figure 13*.

Figure 13. *TDA Systems QuickTDR™ probe supported by Cascade Microtech EZProbe™ positioner. The spacing between signal and ground in the tip of the probe is extremely small to ensure best performance*

A package fixture may be used to conveniently provide a connection to the balls or contact pads of the package. Such fixture must also ensure contact for the signal connection, and it also must provide a ground plane to allow accurate impedance measurement. (*Figure 14*). An automated version of such fixture may be even more beneficial for high-throughput fault isolation work.

A homemade version of the same fixture can be also made. However, the designer of such a fixture needs to ensure both good low-inductance connection to the signal pins and a good ground connection, such as provided in a commercial fixture shown on Figure 14.

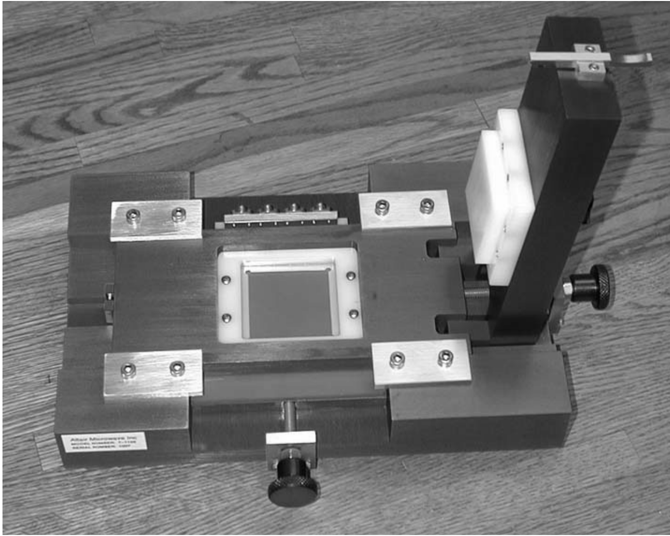


Figure 14. *Altair Microwave BGA package probing fixture provides a high-frequency connection for four pins of the package simultaneously. The body of the fixture serves as the ground plane*

Using Good Measurement Practices

To obtain good quality impedance, signal integrity modeling, and failure analysis data, it is important to follow general good measurement practices when using a TDR oscilloscope. The instrument should be turned on and its internal temperature should be allowed to stabilize for 20-30 minutes before performing any measurements. Calibration, compensation and normalization for the instrument must be performed regularly, as specified by the instrument manufacturer. The internal instrument temperature must be within the specified range from the calibration points for the given instrument.

To maximize the resolution of the scope, particularly in the time axis, it is important to zoom in on the DUT – but at the same time to allow a window that is sufficiently long to include all the reflections related to the DUT. The window that is too short may prevent the designer from obtaining complete and accurate information about the DUT.

It is critical to use a torque wrench when mating any two connectors in your probing, cabling and fixturing setup. Such torque wrenches ensure a repeatable connection between the connectors, thus providing better measurement repeatability. These torque wrenches are available from most TDR manufacturers and many microwave component suppliers. It is also important to clean the RF connectors used in the probing / fixturing setup with isopropyl alcohol and lint free swab.

[2] However, a ground lead that is 10mm long will probably produce a 10nH parasitic inductance and pretty much destroy the measurement accuracy.

Failure Analysis Goals and Methods

The goal and the task of the failure analyst is to determine *whether* there is a possible connection failure in the given package trace, and what the *exact position* was when the failure occurred. Once the position of the failure is determined, further analysis can be performed to determine the physical cause and the nature of the failure, possibly with destructive analysis methods. Thus, in this scenario TDR is a fault isolation or a fault locator technique, allowing the failure analyst to quickly find the failure, and analyze it using other, possibly destructive, analysis techniques.

For example, the following picture illustrates how TDR was used to locate a short in a package.

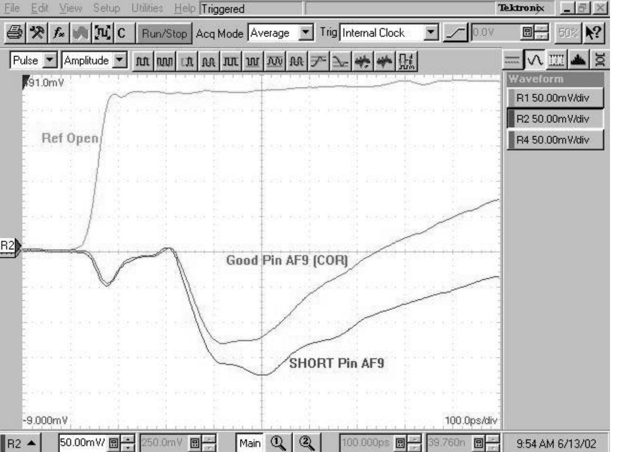


Figure 15. TDR result showing I/O short (AF9 and AE8) in a package

And the picture that follows shows an X-ray image of the same short failure.

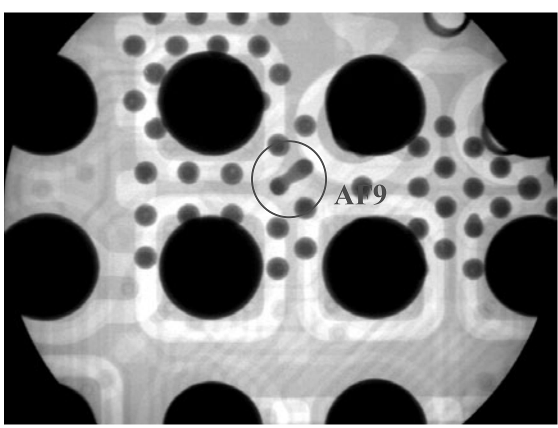


***Figure 16**. X-ray image of two solder bump shorts*

The following four pictures are additional examples of TDR open signatures in a package, a short signature in a die, and corresponding optical images.

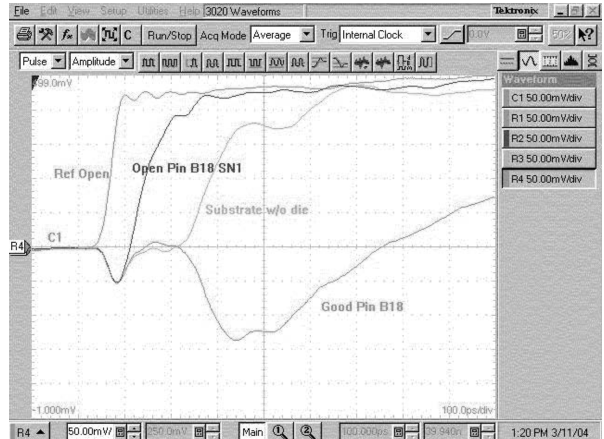


***Figure 17.** TDR result showing an open in the package substrate*

***Figure 18.** Optical image showing a broken trace in the package substrate*

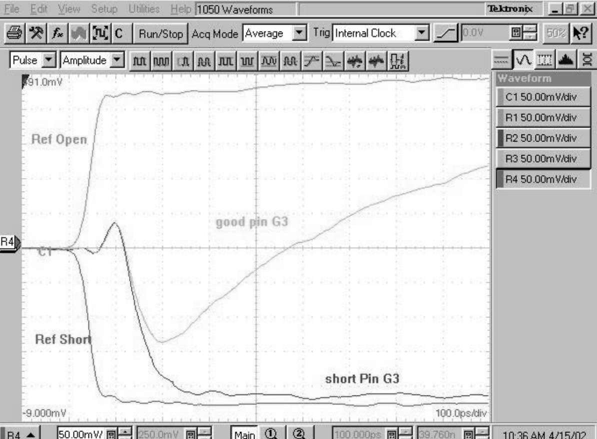


***Figure 19.** TDR result showing I/O short in the die*

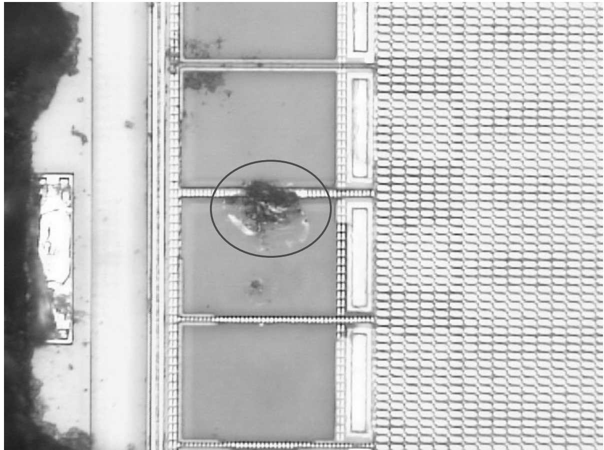


Figure 20. Optical photo of an EOS short in the die on pin G3

Typical approaches that can be used to determine whether there is a failure present are *signature analysis*, where the package trace true-impedance-profile data is analyzed for known failure signatures, and *comparative analysis*, where the package trace data is compared to the data of a trace in a known good package. Both approaches will be applied to the true-impedance-profile data obtained from the TDR using the impedance deconvolution algorithm as it is implemented in TDA Systems' IConnect® software.

As *Figure 9* indicates, the true-impedance-profile provides a much clearer picture of the failure type, and also enables the user to easily determine the exact position of the failure in an electrical sense, i.e., in terms of electrical length of the interconnect in picoseconds. Additional analysis must be performed to determine the *physical* location (in millimeters or milli-inches) of the failure with the goal of locating the package element that is failing. The true-impedance-profile provides the user with a way to correlate the TDR data to the specific layers in the package, as well as provide an estimate of a constant that would allow the user to convert the electrical length in picoseconds into physical lengths in mils.

Signature Analysis

In the true-impedance-profile, open and short failures can be easily identified as 0 Ohm or very low impedance readout for the short and very high (1000 Ohms or more) impedance readout for the open (*Figure 6*). The exact electrical position of a short or an open can be easily identified in the true-impedance-profile, even in the presence of multiple reflections, as previously described.

In the following example (*Figure 21*), the known good BGA package (ZlineGood.wfm) was analyzed alongside a suspect package (ZlineBad.wfm). The fixture-impedance-profile (ZlineFixture.wfm) is shown for reference. The known good package impedance profile ends with a large capacitive dip, corresponding to the input package capacitance. An open failure is clearly observed in the BGA package at about 80 ps inside the package (160 ps roundtrip delay).

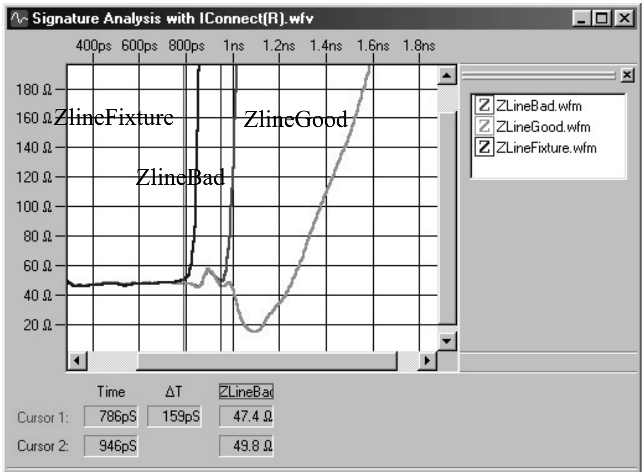


Figure 21. Signature analysis of a BGA package failure using the true-impedance-profile in IConnect

So called "soft" failures, i.e., partly shorted or partly open leads, can also be identified using the signature analysis, but their impedance profile and TDR signatures must be identified beforehand. The only alternative to knowing the soft failure signature beforehand is to observe the changes in capacitance of the known good device compared to the failing device.

TDR has specific signatures for the open and short connections, as shown in *Figure 6*, and can also be used for identifying the failures. However, in multi-segment structures, such as BGA packages, the exact location of the failure can be difficult to determine because of the multiple reflection effects. This is why we want to emphasize the importance of using the true impedance profile when performing fault isolation on electronic packages.

Comparative Analysis

Comparative package failure analysis, as the name implies, relies on comparison of the known good waveform to the suspect waveform. Even though some discrepancy between different measurements may still be observed due to measurement repeatability, comparative analysis utilizing the true-impedance-profile waveforms, computed using IConnect, yields very quick and intuitive results.
Consider the following example. In *Figure 21*, the package failure is identified as an open failure. In *Figure 22*, the analysis is continued by comparing the failed waveform to the package substrate waveform only, without connection to the die. The challenge is to determine what package component is failing based on this comparative analysis. Because the failed impedance profile waveform overlays directly over the substrate waveform, it is easy to deduce that the likely failure source is the broken connection between the package and the die. Again, the large capacitive dip is due to the input capacitance of the die.

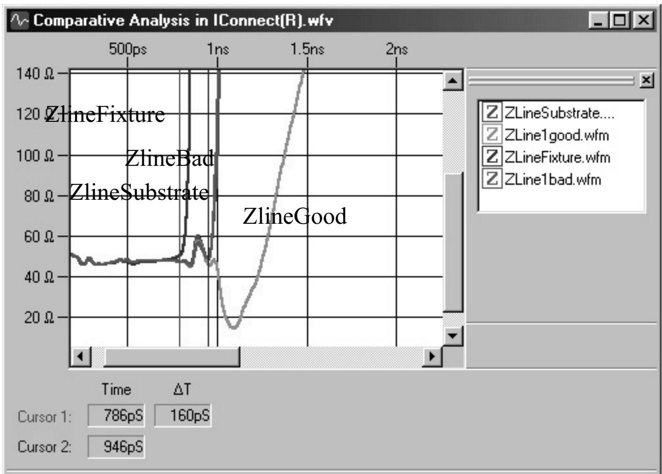


__Figure 22.__ Comparative analysis for a BGA package. The bad impedance profile waveform clearly indicates an open failure signature. Comparing it to the package substrate waveform only without connection to the die, allows pinpointing the likely failure source – a broken connection to the die

Based on this analysis, a failure analyst can focus on the connection to the die area, and use additional failure analysis techniques to determine the physics of the failure.

Thus, in comparative analysis at the very least we need to have a known good device. For a typical package, one of the key impedance profile features differentiating an open fault in the package or bondwire from an open fault in the die itself is a dip in the impedance profile corresponding to an ***input die capacitance*** and presenting itself shortly before the open impedance signature. The presence of this characteristic dip in the impedance profile indicates a good connection to the die, whereas its absence indicates a problem in the package structure. Such a capacitive signature can be observed much more readily on the impedance profile waveform than on a raw TDR waveform. An additional known good bare package substrate can also be useful in identifying the exact location of the failure.

It should be noted, however, that even though the appropriate signatures are much more readily observable on a true impedance profile waveform, some basic failure analysis work can be done with a TDR instrument alone, coupled with a simple probe or fixture. For example, the following two pictures illustrate a failure near the connection to the die in a chip-scale package, as well as near the BGA package ball.

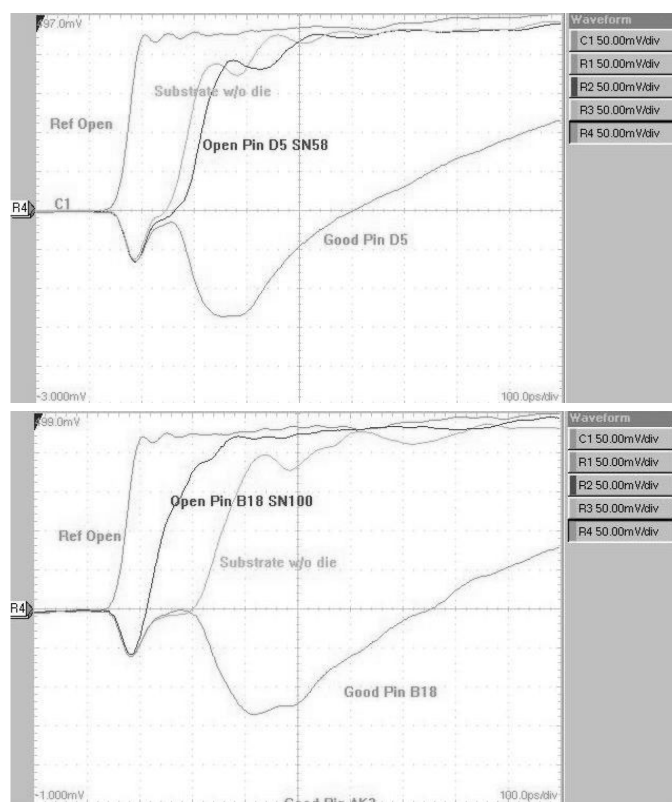


__Figure 23.__ Failures in a chip-scale package near the connection to the die, as well as near the BGA package ball

An important issue when performing comparative analysis is measurement repeatability. Following good measurement practices will enable the analyst to minimize any non-repeatability errors, These good measurement practices include:

- maintaining TDR instrument calibration
- keeping the instrument well-warmed in a lab with constant ambient temperature
- maintaining the probe or cable position and spacing between the probe signal and ground during the measurement

However, a failure analyst must be aware that small differences between different impedance profiles may actually result from measurement non-repeatability, rather than failures in the package under test.

For example, because of the differences between the good package impedance profile (ZlineGood.wfm) and bad package impedance profile (ZlineBad.wfm) in the outlined region of *Figure 24*, a failure analyst may view the differences between the good and bad waveforms in the selected region as the cause for the failure observed in the later portion of the impedance profile. However, because we are working with the impedance profile and not the TDR waveform, any effect of the reflections in the selected region on the rest of the impedance profile waveform is minimal. With that in mind, the differences between the two impedance profiles are too small to be viewed as the cause of the failure. And, one can comfortably conclude that the failure occurred in the later

portion of the package (in this case, again, it is a failure of the package-to-die connection.)

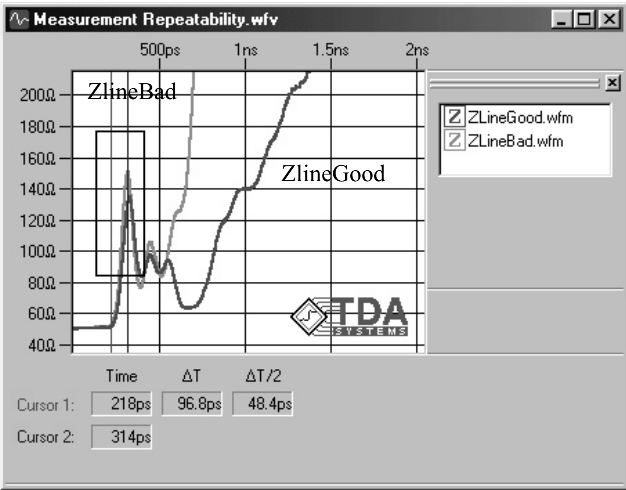
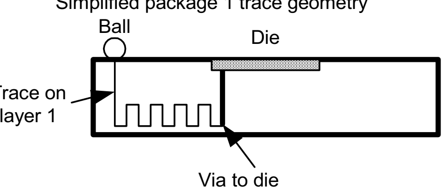


Figure 24. *Measurement repeatability considerations*

Additional Considerations for Package Open Failure Analysis

The true-impedance-profile is very powerful because it opens up other interesting venues for FA on electronic packages. For example, because the true-impedance-profile represents an exact signature of the DUT, one can now analyze the package impedance profile and quite easily correlate it to the physical layers in the BGA package, which can be observed in the package layout or drawing.

Consider the following two package samples with the following simplified trace layouts in *Figure 25*. The two packages are quite similar, except that the trace leading to the via connecting the package trace to the die is significantly longer for package 1. Both of these packages were analyzed with a TDR instrument and an impedance profile in IConnect. In both cases a good package sample and a sample with a failure of the connection between the package trace and the die has been analyzed.



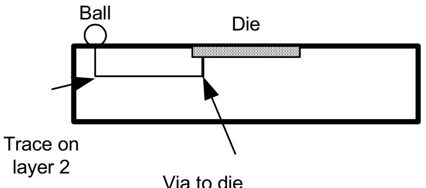


Figure 25. S*ample package trace geometries used for correlation to the impedance profiles*

The impedance profile enables a simple correlation to the package geometry (*Figure 26*). In package 1, the known good waveform (Zline1good.wfm) shows a segment with inductive behavior (estimated to be about 2nH in inductance), correlating to the long package trace, then a short segment correlating to the via, and then a segment correlating to the input capacitance of the die. When the connection to the die is broken, the corresponding waveform (Zline1bad.wfm) still shows the long trace in the package, but does not go into the capacitance of the die (estimated to be 800fF). Finally, the shorter second package trace correlates to the shorter section in the impedance profile waveform (Zline2good.wfm), whereas for the failed trace in package 2, the impedance profile goes up to high impedance at a much earlier point. The estimates for the inductance of the trace and input capacitance of the die match the expected numbers well, which provides further confirmation for the accuracy of the analysis of the failure type and location.

Once the correlation from the physical package structure to the impedance profile waveform has been determined, the location of the fault in the package can be found easily.

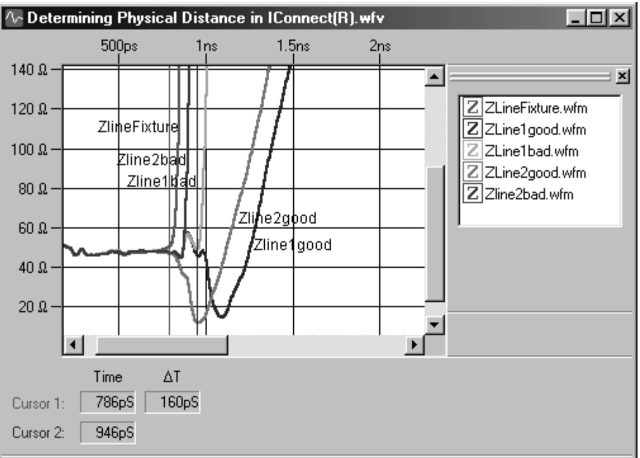


Figure 26. *Layer correlation and distance analysis in IConnect based on the impedance profiles of two packages with similar layouts*

In addition, since the overall physical length of the package trace can be quickly found from the package layout, and the impedance profile provides exact information about the electrical length of the package trace, this correspondence can provide a reasonably good estimate of the physical location of the failure. For example, if the package layout software gives a reading for the overall package trace length of l_{total} meters, and the true-impedance-profile shows that the package length is $t_{d\ total}$ seconds, then the average relative velocity of propagation through the package can be estimated as

$$V_{prop\ average} = \frac{l_{total}}{t_{d\ total}} \cdot \frac{1}{V_C} \qquad (3)$$

where V_C is the speed of light. For example, the difference between the length of the traces in package 1 and package 2 is 45 ps (90 ps roundtrip). Based on the layout file data, the corresponding physical length is 10 mm, which provides an estimated relative velocity of propagation of 4.5 ps/mm, or 0.74 the speed of light.

In addition, if a correlation between an electrical position in seconds to the physical position in meters needs to be estimated, it can be done using the following equation

$$l = t_d \cdot \frac{l_{total}}{t_{d\ total}} \qquad (4)$$

Using equation (4), one can estimate the relative position of the failure within a layer, if it is suspected that the failure actually occurred *within* a layer.

One can extend the computation above to determine the average dielectric constant E_r through the package using the following equation:

$$V_{prop\ average} = \frac{V_c}{\sqrt{\varepsilon_{r\ average}}} \qquad (5)$$

where $V_{prop\ average}$ is the average signal propagation velocity through the package, V_c is the speed of light in vacuum. We can rewrite this equation as

$$\varepsilon_{r\ average} = \left(\frac{V_c}{V_{prop\ average}}\right)^2 \qquad (6)$$

Now, the computed ε_r value can be entered into the software, and the results can be displayed as impedance vs. physical length (millimeters, inches, or feet).

Clearly, equations (3) through (6) are only estimates. The propagation velocity will vary through the different layers in the package. To get a more accurate value for the propagation velocity one needs to do extensive characterization of the package substrate material, as well as other characteristics. Such characterization is very time consuming and requires that special test structures be laid out on the material under test ([11, 12]). Because of such complexity, the exact data about the velocity of propagation through the separate package layers is rarely available to a failure analyst. A much easier approach is to correlate the layers in the package to the segments in the true-impedance-profile and use equation (3) to estimate the propagation velocity in each layer. However, sufficient resolution of the TDR instrument is required to resolve the layers, which can be on the order of 10 ps or less.

An attractive approach for a failure analyst could be to model the package under test, and then attempt to predict the TDR waveform of the package trace via SPICE or full-wave circuit simulations. The problem with this approach is, again, that the properties of the package material must be known with a reasonably high level of accuracy in order to ensure that the simulation predicts the TDR waveform correctly, unless the package model has been directly extracted from TDR measurement.

Signal-to-Ground Short Failures

Experiments have demonstrated that a signal to ground plane short can be located relatively easily using the same techniques as those used for locating an open fault. Since the trace has a certain amount of characteristic impedance, a short failure will clearly exhibit itself on the true impedance profile, as a rapid decrease in impedance until this impedance reaches zero Ohm.

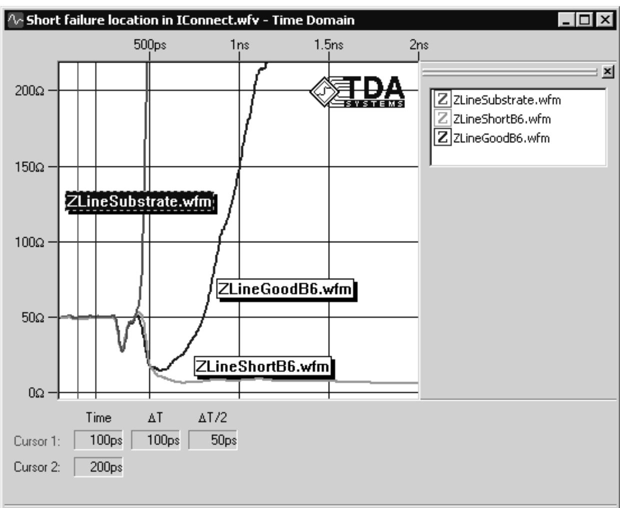


Figure 27. *Locating a signal to plane short failure using the impedance profile signature. By comparing the short waveform (ZlineShortB6.wfm) to the substrate waveform (ZlineSubstrate.wfm), one easily concludes that the short is located near the connection to the die*

Since in *Figure 27* we can observe that the short waveform goes towards zero Ohm right where the bare package substrate waveforms goes into an open (high impedance), and right before the good device waveform goes to the die capacitance, we conclude that the short is located at or near the connection to the die.

In this example, the electrical length is 63ps, and the physical length is 8.7mm, which gives an average propagation velocity of 7.24ps/mm or 0.138mm/ps. Using this number, and knowing that the speed of light in vacuum is 0.305 mm/ps, we obtain the average E_r of 4.9. Then, one can display the data in the software as impedance versus physical time.

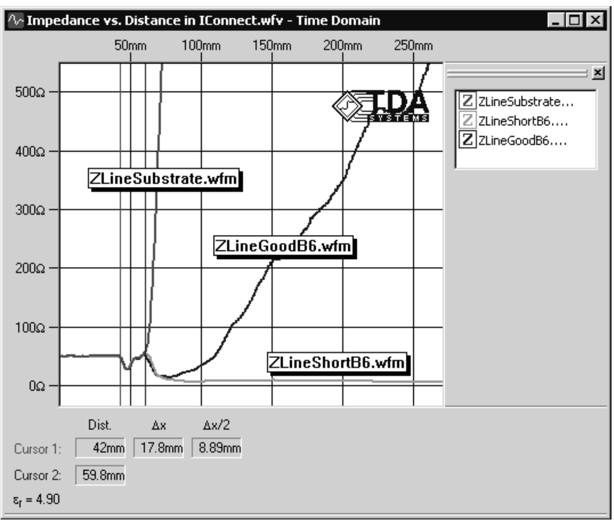


Figure 28. *Impedance vs. distance in IConnect TDR software. The fault occurred about 63ps or 14mm inside the package*

The actual dielectric constant (E_r) for the substrate has been measured to be 4.2 at 1 MHz, which comes from the material property data provided by the substrate vendor. Using E_r=4.2 to recalculate the physical length, the measured total length of the trace is 9.4 mm, which is very close the actual CAD layout length of 8.7 mm (~8% error). It is a very good correlation, considering that the E_r would be higher at higher frequency for the TDR measurement, and that we are looking for the average E_r through the whole package. The localized E_r variations, non-homogeneity, effects of different conductor shape, all contribute to the difference between the vendor supplied E_r and the measured value. However, the overall correlation we observe is considered to be quite good.

Plane-to-Plane Short

Locating a plane-to-plane short, such as ground plane to power plane short, is a more difficult task. In the case of a signal trace, the characteristic impedance of such a trace is normally in the range of 30 to 80 Ohms, and observing the change from this range to zero Ohm in the impedance profile waveform is relatively easy. The power plane, however, presents much lower impedance to the TDR signal, typically on the order of less than 0.5-2 Ohm, and the change from that impedance to zero does not always allow the failure analyst to use the impedance profile effectively to find the exact location of the short between the planes using either time or distance. In this paper, however, we propose two comparative techniques for plane-to-plane short location, both based on secondary information in the TDR data. One technique looks for the difference in the secondary reflections in the TDR waveform, and can be performed with the raw TDR data or using the true impedance profile. The second technique looks at the inductance of the current return path, which can be computed using IConnect software, based on the JEDEC standard described in [13, 14]. Smaller inductance indicates a shorter distance to the short, and by comparing the failing device measurement to that of the good device and a shorted package substrate, one can determine the relative position of the short failure. For both techniques, repeating the measurements multiple times to ensure good repeatability is key to finding a fault.

Consider the following simple test board example (*Figure 29*)

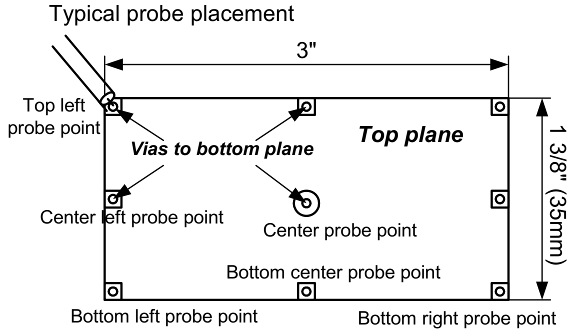


Figure 29. *Test board for plane-to-plane short failure location. The board consists of two planes that can be probed with TDR at multiple locations at one side of the board and shorted at another side*

This test board consists of two planes, which have via probe points at multiple locations on the board. Using these vias, the failure analyst can perform TDR measurements of the two planes on one side of the board, while at the same time shorting the planes at the other side and attempting to locate the position of the short. The board is about twice the size of a typical BGA package, which makes it an easier test case. The following table summarizes the expected closeness of the shorted probe point to the top left probe point, where the TDR signal is applied, based on visual analysis of the board.

Table 2. *Physical closeness of short point to the top left probe point (closest, 1 to farthest, 8), based on visual analysis*

	Left	Center	Right
Top	X	2	6
Center	1	4	7
Bottom	3	5	8

We applied a signal every time at the same location (top left probe point), while creating a short between the planes by connecting the via from the bottom plane to the top plane contact. We shorted the two planes together at each of the remaining probe point locations and acquired the corresponding TDR waveform (Figure 30).

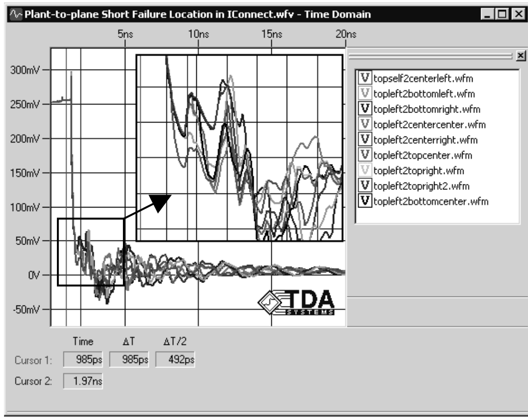


Figure 30. *Locating plane-to-plane short on the test board. The fault location is based on the secondary TDR measurements, such as secondary reflection delay and inductance measurement*

As one can see from Figure 30, there is little delay between the different waveforms. The inset, however, demonstrates, that the waveforms do exhibit a difference in position and waveshape. If a failure analyst tried to analyze this difference and determine which short is closer to the point where the TDR signal is applied, the following order could be established as shown in Table 3 (from closest short to the probe point to the farthest).

Table 3. *Electrical closeness of short point to the top left probe point (closest, 1 to farthest, 8), based on TDR measurement*

	Left	Center	Right
Top	X	4	5
Center	1	3	7
Bottom	2	6	8

By comparing these results with the expected results from Table 2, a failure analyst can observe a good correlation. The only questionable result is that the center top probe point comes after the center point. The difference between the two waveforms corresponding to those probe points, however, is small, and can be attributed to measurement repeatability issues.

To confirm our conclusions, we compute the total inductance of the plane for each measurement in IConnect software. This measurement is a good figure of merit for the current return path through the plane, which, in turn, is a good indicator of where the short may have occurred.

The following table summarizes the inductance data, listing the shorted probe point inductance measurements from the smallest to the largest. All the inductance measurements have been in 2-4 nH range.

Table 4. *Electrical closeness of short probe point to the top left probe point (from smallest inductance, 1 to largest, 8), based on inductance measurements in IConnect TDR software*

	Left	Center	Right
Top	X	3	4
Center	1	5	7
Bottom	6	2	8

Again, the center point and the bottom left point are outliers, but otherwise this table correlates well to the Table 2 of expected physical closeness.

These “outliers,” or incorrect data points, may come from some minor details and changes in the current return path between the planes. Vias and other plane openings, such as those on the test board in question can affect this return path. All these issues indicate that the failure analyst must apply these techniques to plane-to-plane short analysis with the good understanding of the TDR measurement technique, and must study the layout and structure of the package carefully. Repeated measurements may be necessary in order to prove the actual location of the short.

As a final note, additional analysis using triangulation (making TDR measurements from three different points on the plane) would enable even easier location of the position of the short between the planes.

Summary, Conclusions and Future Work

In this paper, we discussed TDR measurement technology as it applies to the failure analysis of electronic packaging. We analyzed the impedance deconvolution algorithm, and demonstrated the advantages that the true-impedance-profile (resulting from applying this algorithm to the TDR data), provides for a package failure analyst over a simple TDR data set, for both signature and comparative package failure analysis. Additional analyses were presented, which can be performed on the true-impedance-profile, and that can further simplify the location of the failures in electronic packaging.

We have analyzed the application of the TDR measurement techniques and the true impedance profile to finding the location of signal-to-ground and plane-to-plane shorts in electronic packages. Locating a signal-to-ground short has been shown to present little difficulty over a comparable open fault locating task. Plane-to-plane shorts, however, present additional challenges, which require more attention to the repeatability and accuracy of the measurements. However, with the true impedance profile and plane inductance analyses, the claim of impossibility of locating a plane-to-plane short is effectively challenged in this paper.

For future work, more automated fixturing is required, which would enable the failure analyst to increase the throughput and locate more failures with less time invested. Faster rise time TDR modules can produce better resolution results and allow the analyst to locate the failure even more easily.

References

1. C. Odegard and C. Lambert, *Comparative TDR Analysis as a Packaging FA Tool, — Proceedings from the 25th International Symposium for Testing and Failure Analysis (ISTFA)*, 1999
2. D. Bethke, W. Seifert, *TDR Analysis of Advanced Microprocessors, – Proceedings from the 26th International Symposium for Testing and Failure Analysis*, 2000
3. D.A. Smolyansky, *Electronic Package Failure Analysis Using TDR, – Proceedings from the 26th International Symposium for Testing and Failure Analysis*, 2000
4. D.A. Smolyansky, D. Staab, P. Tan, *Signal Trace and Power Plane Shorts Fault Isolation Using TDR, - Proceedings from the 27th International Symposium for Testing and Failure Analysis*, 2001
5. P. Viswanadham, P. Singh, *Failure Modes and Mechanisms in electronic packages*, — Chapman and Hall (1998)
6. *TDR Impedance Measurements: A Foundation for Signal Integrity*, — Tektronix Application Note 55W-14601-0
7. *Agilent High Precision Time Domain Reflectometry,* – Agilent Application Note 1304-7
8. D.A. Smolyansky, *TDR Testing Primer, – Printed Circuit Design Magazine*, March 2002
9. L.A. Hayden, V.K. Tripathi, *Characterization and modeling of multiple line interconnections from TDR measurements,—IEEE Transactions on Microwave Theory and Techniques, Vol. 42*, September 1994, pp.1737-1743
10. C.-W. Hsue, T.-W. Pan, *Reconstruction of Nonuniform Transmission Lines from Time-Domain Reflectometry,—IEEE Transactions on Microwave Theory and Techniques, Vol 45*, No. 1, January 1997, pp. 32-38
11. D.A. Rudy, J.P. Mendelsohn, P.J. Muniz, *Measurement of RF Dielectric Properties with Series Resonant Microstrip Elements, — Microwave Journal,* March 1998, pp. 22-39
12. D. I. Amey, S.J. Horowitz, *Tests Characterize High-Frequency Material Properties"—Microwaves and RF*, August 1997
13. *Guideline for Measurement of Electronic Package Inductance and Capacitance Model Parameters*, – JEDEC Publication #123, JC-15 Committee, October 1995
14. D.A. Smolyansky, *TDR Techniques for Characterization and Modeling of Electronic Packaging, – High Density Interconnect Magazine,* March 2001, Part 1 of 2.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

DELAYERING TECHNIQUES: DRY PROCESSES WET CHEMICAL PROCESSING AND PARALLEL LAPPING

Kendall Scott Wills, Srikanth Perungulam

Texas Instruments, Stafford, Texas USA

Purpose

The purpose of this paper is to present the techniques required to delayer a semiconductor device built with current state of the art silicon processing technologies. Many of the techniques with appropriate modifications will be transferable to other technologies. The user will be left to decide when and how to use them.

Before a discussion on delayering begins, a discussion of "Standardization" is appropriate. To illustrate the main concept of standardization I have employed the help of Standard Sam drawn by Estong Peralta of Texas Instruments Philippines, retired.

In the first illustration, Sam has just been given a job. This is his first day to go to work and obviously has never made his way to this job before. The grass is tall so Sam must make his way to work finding the best path he can. See Figure 1.

Figure 1. Sam goes to his first day of work.

Sam perseveres, he pushes onward to build a path to work. He now has a "standard" path that gets him to work on time every day, no problem. Sam can be seen in Figure 2 walking happily to work.

Figure 2. Sam has his first "standard" path.

Unfortunately, in the Philippines rain is common. When a Typhoon comes through the rain is very heavy. In Figure 3 Sam tries to go to work in a Typhoon. The rain is heavy. Try as he might he cannot find his way to work. Sam gets lost and eventually looses a day of work.

Figure 3. Sam finds problems with his "standard". He gets lost going to work in the rain.

The lessons learned from the Typhoon prompted Sam to make major changes in his standard. Sam learned that cutting corners to have a less expensive standard is not always cost

effective. Sam improved his standard by considering all known potential problems. Sam built a roof system over the path. The path needed to have grass to make it comfortable but grass becomes worn out when it is walked on a lot. So Sam put cement blocks with holes in them for the grass. This way Sam can walk on the blocks without hurting the grass.

Sam did a lot of work on his new standard as can be seen in Figure 4. The new standard is defiantly better than the old one but he did not for see all the problems. He still did not account for light or water for the grass. You see his new roof changed the effect he had on his process of going to work.

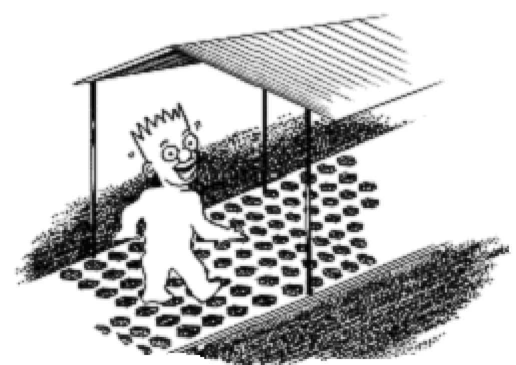


Figure 4. Sam improves upon his standard. Yet, his standard is not good enough. Do you know what is wrong?

Why is it important to understand standardization before we discuss delayering? The answer is simple. Figure 5 lists the reasons for standardization.

Item 3 in the list of reasons for standardization seems odd. After all change is not considered to be permanent, "Standard." Sam developed a standard path, then due to the Typhoon changed his standard. Just as an analyst might need to change their "standard" due to a change in the materials used to fabricate the device being deprocessed.

In Sam's case one could say Sam did not study his environment well enough to make a good standard in the first place. In this case that is most likely true. Sam should have taken into account the rains of the Philippines but he chose a cheaper "standard" hoping it would be sufficient.

In real life all the data for any given problem is seldom available. Decisions must be made on the information available at the time. Cost of implementation of the standard is always a concern. As situations change or information that is more complete becomes available, the "Standard" may need to change.

A standard therefore is like a stable platform. From a stable position change can be made in small increments without upsetting the balance. The Standard must always keep in mind why the original standard was developed and why it needs to be changed to meet future requirements.

List of Reasons for Standardization

1. Allows the individual using the data obtained by the analysis a way to know how the data was gathered.
2. Consistent results when the standard is used for the approved application.
3. Permits the transfer of the standard to others.
4. Permits organized change.
5. There is an understanding of why the previous standard was changed.
6. There is an understanding of where the new standard should be used.
7. When the standard fails to provide the required results a fixed procedure can be followed to find out why the standard failed.

Figure 5. List of reasons for standardization.

DELAYERING

The primary goal of delayering is to successfully remove layers of material in order to locate and identify the physical defect causing a failure. Critically important is the selection of the appropriate methodology and analytical tool set. The correct delayering standard permits the successful identification of the fail mechanism. Figure 6 lists the considerations for a successful delayering standard.

Considerations for a Delayering Standard

- SELECTIVITY
- ETCH RATE
- (AN)ISOTROPY
- MINIMAL DAMAGE TO UNDERLYING LAYERS
- EFFECTIVE REMOVAL OF REACTANTS AND PRODUCTS…..

Figure 6. Considerations for the "standard" process of Delayering.

Selectivity as listed in Figure 6 is the ratio of the rate of etching of the material being removed to the rate of etching of the surrounding material. The following formula describes the selectivity.

SAB = etch rate A/etch rate B

Where SAB is the selectivity of A to B.

Anisotropy as listed in Figure 6 describes the ability of the etch to remove material preferentially in the horizontal direction over the vertical direction. The following formula describes the anisotropy of an etch.

A = 1 - Vl/Vv

Where A is the anisotropy, Vl and Vv are the lateral and vertical etch rates respectively. A purely anisotropic etch has A = 1.

WET CHEMICAL ETCHING VS PLASMA ETCHING

There are three predominate methods to delayer a device. They are wet chemical material removal, Plasma material removal, and lapping to remove the material. For the purposes of delayering, sawing and grinding to reduce sample size or to permit quicker access to the defect of interest shall be considered under the category of lapping.

There are new methods that will be discussed latter which are gaining ground in the delayering arena. They are Focused Ion Beam (FIB) milling, Laser delayering, and Ion Beam milling (the beam is not focused but rather is made up of ions moving in parallel with one another.)

Let’s start by comparing wet chemical delayering with Plasma delayering. Figure 7 shows the comparison.

Delayering Etching Options

------------ PROS --------------

WET ETCHING

- Fast Etch Time
- Ease of Adjustment
- Low Cost

DRY ETCHING

- Variety of Modes of Operation – Barrel, Planar, ECR, RIE, Downstream
- Automated Recipes
- Repeatable Process

WET ETCHING

- Too Many Variables – Time, Heat, Light, Chemical Composition, pH
- Variable Results
- Etch Artifacts

DRY ETCHING

- Heating of Devices (Failure Recovery)
- Damage From Antenna Effects
- Redeposition/Sputtering of Heavy Metals (Au, Pb, etc)
- Etch Artifacts

Figure 7. Delayering options comparison Wet VS Dry.

The chemistries for some selected wet chemical processes are given in the following section. While reading through the different formulations remember that wet chemical delayering can be tailored to be fast or slow by changing the concentration of the etch constituents or by buffering the etch. A special comment on the buffers will be given where appropriate.

Access to the semiconductor device (die) may require the removal of a plastic mold compound. Detailed instructions for decapsulation are given in other papers so only an overview will be given here.

The plastic mold compound removal is generally performed by hot nitric acid. If the mold compound is especially difficult to remove, hot sulfuric acid may be used. Some mold compounds require a mixture of hot nitric and hot sulfuric acids.

The acid temperature is very important. Temperature ranges from 25C to 180C have been reported depending upon the application.

For most applications, the water content of the acid is critical. If the water content is high, the acid will react with the metallization in the wire bond region. Reaction of the acid with the wire bond region may result in a loss of electrical continuity. For some applications electrical continuity is necessary as the device being deprocessed must be electrically tested after the decapsulation process to isolate the fail site to a specific location.

The best nitric acid to use is red fuming but due to limited supply the semiconductor industry has moved away from red fuming nitric acid to yellow fuming nitric acid. Use fuming sulfuric acid to avoid any moisture in the chemical.

Nitric acid can be used to dissolve Pb-Sn solder bumps after the package has been removed on a Flip Chip device. There are also premixed eutectic solder removal formulations, which work well.

If the device has gold bond wires that need to be removed a solution of Aqua Regia (HCl: HNO_3) or a mixture of Potassium Iodide and Iodine in DI water will dissolve the gold.

As a rule most noble metals such as Au, Pd and Pt dissolve in Aqua Regia while most transition metals such as Ta, Ti and Nb do not.

Each metal requires a different etch formulation. Figure 8 gives a list of possible metal etches.

Wet chemical etching of silicon relies on an oxidizing agent to convert elemental silicon to a higher oxidized state. In the formulations, nitric acid and hydrogen peroxide are used as oxidizing agents.

The silicon oxide must be converted into a solution. Hydrofluoric acid used for this purpose.

Possible Metal Etches

- **TITANIUM - H_2O_2:H_2SO_4**
- **TUNGSTEN - H_2O_2:H_2SO_4**
- **SILVER - NH_4OH: H_2O_2**
- **GOLD - HCl:HNO_3**
- **NICHROME - H_2SO_4**
- **BARRIER LAYERS**
- **TiN: NH_4OH, H_2O_2 and H_2O in even proportion**
- **Ti:HNO_3:HF::100:1 or H_2O_2:H_2SO_4**
- **TiW: H_2O_2**
- **Refractory Metal Silicides: Mixture of HF, HNO_3 and NH_4OH**

Figure 8. A list of Possible Metal Etches.

A monitoring agent such as acetic acid is used to control the rate of the etch. An interesting trick to help improve the etch quality of a silicon etch is to season the etch with a sacrificial piece of silicon. So add a little piece of silicon into your etch 30 minutes before use to make the etch quality higher.

Polysilicon etches are used to expose the gate oxide for inspection. A high degree of selectivity of polysilicon to gate oxide is required. One etch composition is HNO_3:HF:H_2O (25:1:16). Other silicon etches work on polysilicon as well. Choline(Trimethyl Hydroxyethyl Ammonium Hydroxide) and TMAH(Tetramethyl Ammonium Hydroxide) are also useful. They have high selectivity ratios of polysilicon over oxide. These two etchents are also used for backside etching of the silicon to expose the gate oxide.

An alternate method to etching silicon with acids or solvents is to use alkaline etches. Depending upon the crystal structure the etch can be anisotropic with superior smoothness to the acid etches. One etch is 45% KOH in H_20. The etch can be monitored by the evolution of H_2.

Silicon etches can be based on an HF-HNO_3-CrO_3 system. In this case the etch rate depends upon the orientation of the crystal. Such etches are used as stains to determine the crystal orientation.

Defects in the silicon substrate may have a different crystal orientation than the bulk silicon. The shape of the etch pits for stacking faults is dependent upon the crystal structure (100) and (111). A crystal structure of (100) will show rectangular pits while the (111) crystal will show triangular pits. The defect is enhanced during etching by the use of Copper or Chromium to control the direction of the etch.

The etch rate of the silicon can be controlled by the $HF:HNO_3$ ratio. By changing this ratio, preferential etching of p-doped over n-doped regions can be obtained. Care should be taken when etching silicon with different doped regions to control the light incident on the surface of the sample. The light in combination with the diode junction formed by the different diffused regions forms a photo battery. The change in the potential at the surface of the doped regions can cause one region to etch faster than another with unexpected results.

Oxide glasses are all etched in approximately the same way. They use an HF etchent. Ammonium Fluoride is one possible buffer for HF etchents for oxides. As a general rule of thumb for VAPOX 3 minutes are required for every 1 micron of interlevel oxide if the oxide is undoped. Boron doped oxides require twice as much time.

Nitrides can be etched with a mixture of nitric acid, ammonium fluoride, Ethylene glycol and acetic acid.

Wet chemical etches for delayering are cheap and cost effective. The equipment is not difficult to implement and is inexpensive. When implementing a wet chemical process caution should be taken to understand all safety aspects of the chemicals and processes being used. The chemicals used in deprocessing can be flammable, toxic, reactive, carcinogenic or some combination of all the previous conditions.

Wet delayering has a few problems. There are multitudes of variables that must be controlled to "standardize" the process. Most labs just assume the chemicals they use are appropriate for the job without controlling the conditions of the etch. Lack of control causes variable results and in some cases etch artifacts. See Figure 9 for some items to control when wet chemical etching.

Items to Control in a Wet Chemical Delayering Process

- **Process time or etch stop**
- **pH**
- **Temperature**
- **Ramp time to temperature**
- **Light**
- **Chemical composition**

Figure 9. Items to control in a wet chemical delayering process.

Typically, when trying to set a "standard" delayering process the comment is made "I can't have a standard process. The material varies too much."

The material may vary but what the process is trying to say is that all the variables have not been taken into account. In the case where the material varies from sample to sample the analyst should make an attempt to understand why the materials vary. Then change the process accordingly. One approach to controlling the process with a wide variation in materials is to implement a method of end-point detection.

PLASMA PROCESSNG

Plasma processing can be accomplished in a variety of tools such as a barrel, planer, ECR, down stream, and inductively coupled plasma (ICP) reactor. Because the processes require a vacuum and require some electronics to control the plasma the equipment designers have tended to add more automation to the delayering process than is found in wet chemical processes or lapping equipment. The processes may therefore be more consistent.

Anisotropic etching permits the removal of the oxide around the metal lines without undercutting the metal. Figure 11 shows a metal stack that is still intact after plasma etching.

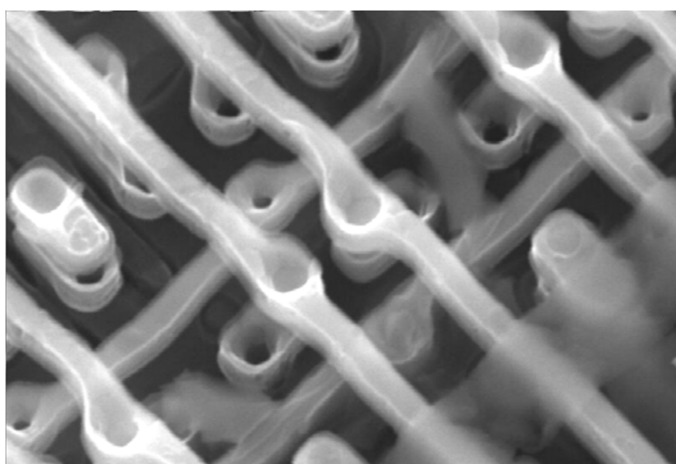


Figure 11. Metal stack as seen after plasma etching. The metal stack is now ready for SEM inspection. Courtesy of Michael Strizich of Analytical Solutions, Inc.

Improper plasma process selection can still cause problems such as the lifting metal seen in Figure 12 that is due to an over etch of the oxide. Improper process selection can be seen when the etch is incomplete as in Figure 13, polyimide residue. Once the plasma process is adjusted the problems with out of control processing are less frequent than with other types of deprocessing, which are mainly manual.

Even a good plasma delayering process has it own set of issues. Plasma delayering can be a hot process. The device temperature can rise to several hundred degrees Centigrade unless the reactor has methods to control the device temperature. When deprocessing units where the electrical signature must stay intact throughout the procedure, heating of the device becomes a problem as high temperature can cause the failure to heal. A temperature-controlled plasma reactor is required when temperature must be controlled.

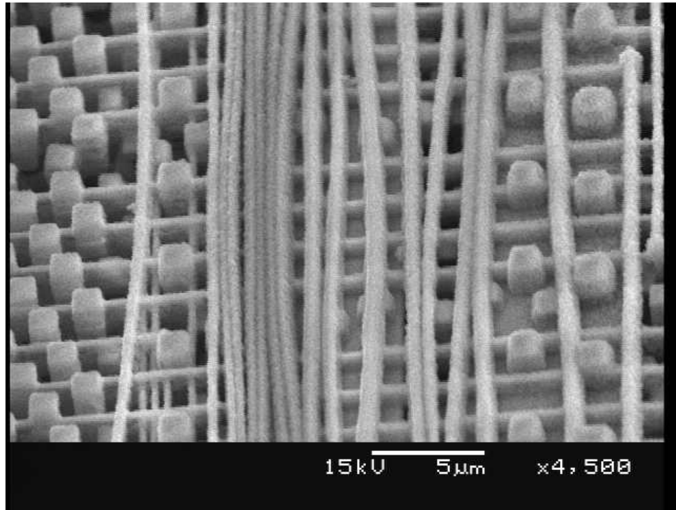


Figure 12. Lifting metal due to plasma over etch of the oxide. Courtesy of Trion Technologies, Inc.

The plasma generates a UV light that can cause healing of the failure. Units where the electrical signature is known to be UV sensitive must not be plasma etched if electrical testing is to be performed after the delayering.

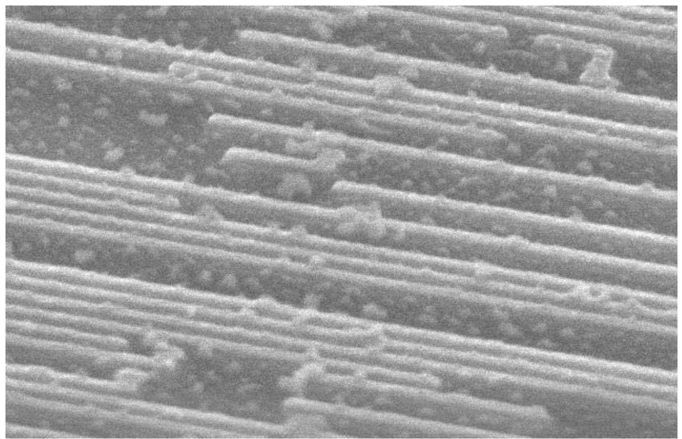


Figure 13. Polyimide residue due to incomplete plasma etching. Courtesy of Trion Technologies, Inc.

The high Radio Frequency (RF) electrical field caused by the plasma generation technique forces the integrated circuit designers to add special structures to the design, antenna diodes, to prevent electrical over stress (EOS). EOS can cause unwanted artifacts in the device or can cause the failure to change characteristics.

Heavy metals when present, such as Au can be sputtered back onto the sample if the RF power at the device surface is to high. Other etch artifacts such as RIE grass as seen in Figure 14 can be generated by micro masking of the device being delayered due to deposition on the surface of the device by contamination in the plasma chamber.

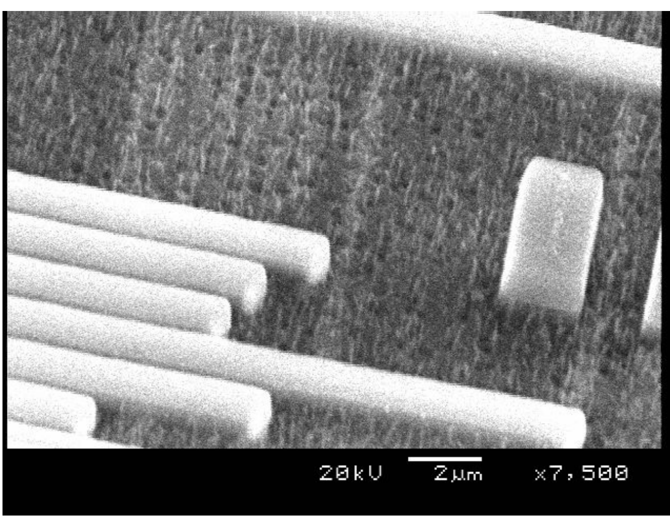


Figure 14. RIE grass due to micro masking. Courtesy of Trion Technologies, Inc.

There are several sources of contamination in a plasma chamber. The heavy metal on lead frames or other locations on the device can be sputtered back onto the die. The substrate on which the devices sits can be a source of contamination that causes micro masking by sputtering of the substrate material onto the device. Then there is the material deposited on the chamber walls due to the plasma processing that can be deposited on the device causing micro masking.

Some plasma etch processes intentionally generate a Teflon like compound on the vertical surfaces of the material being etched. The floro-carbon material acts like a lateral etch mask to prevent etching. This gives a anisotropy value of nearly 1. That is no lateral etching. If the ion bombardment of the etch is not sufficient to break down the polymer on the surfaces to be etched away then a Polymer grass is formed as seen in Figure 15. The same polymer can contaminate the chamber. When etching the polymer deposited on the chamber sidewalls can be deposited on the device being delayered causing RIE grass as seen in Figure 14.

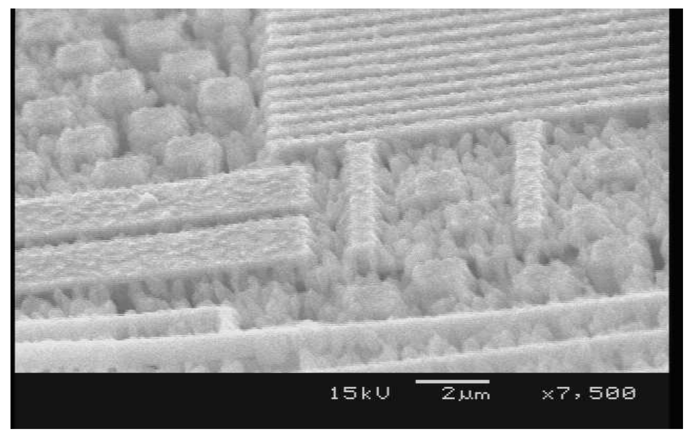


Figure 14. Polymer Grass. Courtesy of Trion Technologies, Inc.

There are so many different etch processes that trying to describe each ones individual characteristics would take too much space. A table of processes is given in Figures 15 for delayering. In Figure 16 are deposition processes. When delayering there times when sacrificial materials need to be deposited on the surface of the device to protect layers which are to remain after the deprocessing step. As an example deposition an oxide after plasma etching the Multi Level Oxide (MLO) can protect the substrate when the surface is lapped to remove the polysilicon gate.

Process for Dry Deprocessing of Semiconductor devices

Material	Gas Used	Rie Power Range	ICP Power Range	Pressure	ICP	Turbo	Selectivity	Etch Rate
Aluminum (Al)-anisotropic	$Cl_2 + BCl_3$	50-200	No	30 mTorr	no	yes		4-5000 A/min
Aluminum-isotropic	Cl2 + BCl_3	50-200	No	180 mTorr	no	no		5-10,000 A/min
Aluminum Gallium Arsenide (AlGaAs)	$SiCl_4 + BCl_3$ + Ar	50-200	0-300W	5 mtorr	yes	yes	PR (7:1) SiN (25:1)	10000 A/min
BPSG	CF4/(O2 or Ar)	20-50	300-500W	5-75 mTorr	yes	yes		1500A/min
Cadmium Telluride, and other II-VI's	H2 + CH4 + Ar	65-150	0-300	20 mTorr	yes	yes	PR (1.5:1) SiN (>50:1)	up to 10000 A/min with heat added
Carbides (i.e. SiC, AlTiC)	Cl2 + SiCl4	100-250	0-500	5-75 mtorr	maybe	yes		varies
Carbon (C)	O_2 + Argon	100-300	No	250 mTorr		no		
Chrome (Cr)	$Cl_2 + O_2$	50-150	No	50 mTorr	no	yes		600 A/min
Copper (Cu)	$BCl_3 + Cl_2$ (heat 200C)			180 mTorr	no	no		2000 A/min
Epoxy	O_2 + %5 CF_4	20-50	300-500	250mTorr	yes	no		10,000 A/min
Fused silica	CHF3 + 10 % O2	200-600	No	10 mTorr	no	yes	PR (1.5:1)	1000 A/min
Gallium Arsenide (GaAs) thinning	$Cl_2 + BCl_3$ (isotropic)	20-200	0-500W	30-200 mTorr	yes	yes		30000 A/min
(GaAs) profile	Cl_2+BCl_3 (anisotropic)	20-70	0-250	5 mTorr	yes	yes	PR (4-16:1)	15000-20000 A/min

Process for Dry Deprocessing of Semiconductor devices- Continued

Material	Gas Used	Rie Power Range	ICP Power Range	Pressure	ICP	Turbo	Selectivity	Etch Rate
Gallium Nitride (GaN)	Cl2 + SiCl4	20-300	0-600	5 mTorr	yes	yes		5000-6500A/min
Gold	HBr + BCl3 (heat 50-90C)	200	0-600	6 mTorr	yes	yes		1600 A/min
Indium & Aluminum containing III-V's	HBr + BCl3 (heat 50-90C)	200-300	0-600	5 mTorr	yes	yes		1000-1500 A/min
Indium Phosphide (InP)	$CH_4 + H_2$	20-200	0-300	5 mtorr	yes	yes	SiN (50:1)	3-500 A/min
Indium Phosphide (InP)	HBr + BCl3 (heat 50-90C)	20-70	300-500	5 mtorr	yes	yes	PR (2:1) SiO (20:1)	1500-2500 A/min
Molybdenum (Mo)	SF_6	50-250						
Oxynitride	CF_4 + 5% O_2	100	No	150 mTorr	no	yes		1800A/min
Photoresist	O_2	10-200	0-500	50 mTorr	yes	yes		800 A/min
Platinum	Cl_2, Ar (heat 90C)	100-200	No	5 mtorr	no	yes		1600 A/min
Polyimide	O_2 + Ar	50-500	0-500	20 mTorr	yes	yes		5000 A/min
Polysilicon (isotropic)	Cl_2	50-250		180 mTorr	yes	yes	SiO (20:1)	5000 A/min
Polysilicon (anisotropic)	Cl2 + HBr	20-200	0-500	30 mTorr	yes	yes	SiO (100:1)	3000 A/min
PSG	CF4/(O2 or Ar)	20-50	300-500W	5-75 mTorr	yes	yes		1500 A/min
Quartz	$CHF_3 + CF_4$	200-600	No	5 mtorr	yes	yes	Cr (20:1) PR) (1:1)	3000 A/min
Silicon (Si)	SF_6 + He + O_2	200	No	10 mTorr	yes	yes		40,000 A/min
Silicon Dioxide (SiO2)	CF4/O2 or CHF3/O2	100	No	50 mtorr	no	yes		1800 A/min
SU8	$O_2 + CF_4$	20-70	500	100 mtorr	yes	yes		13000 A/min
Silicon Nitride (Si3N4)	SF_6/O_2 or CF_4/O_2	100	No	150 mTorr	no	no		2200 A/min
Tantalum (anisotropic)	$CF_4 + O_2$	50-250	No	20 MTorr	no	yes		2,000A/min
Titanium (Ti)	$Cl_2 + BCl_3$	20-70	0-500	5 mTorr	yes	yes		600 A/min
Tinitride (TiN)	$SF_6 + O_2$	150	No	100 mTorr	no	yes		1500 A/min
TiTungsten (TiW)-iso	SF_6	200	No	200 mTorr	no	no		5000A/min
TiTungsten (TiW)-aniso	$SF_6 + O_2$	200	No	5 mTorr	no	yes		10000A/min
Tungsten (W)-isotropic	SF_6	200	No	250 mTorr	yes	no		6000 A/min
Tungsten (W)-anisotropic	$SF_6 + O_2$	200	No	5 mTorr	yes	yes		10,000 A/min

Figure 15. Plasma delayering processes, 1. Courtesy of Trion Technologies, Inc.

Process for Dry Deposition of Semiconductor Materials on Semiconductor devices

Material	Gas Used	Rie Power Range	ICP Power Range	Pressure	ICP	Turbo	Selectivity	Etch Rate
Amorphous Silicon	SiH4	25-100	no	600mT	No	no		1200 A/min
Nitride	NH_3 + DES	50	No	600 mTorr	No	no		50-1000+ A/minute
Nitride (faster deposition)	$NH_3 + SiH_4$	70	No	625 mTorr	No	no		100-1000+ A/minute
Nitride (Ammonia Free)	SiH4 + N2	25-100	No	700mT	No	no		200-800 A/min
Oxide	TEOS	40	No	1 Torr	No	no		200-1000 A/minute
Oxide (faster deposition)	$SiH_4 + N_20$	60	No	900 mTorr	No	no		200-1200+ A/minute
Oxynitride	SiH4, N2, + N20	25-100	no	900mT	No	no		200-1000 A/min
Silicon carbide	trimethylsilane or CH4/SiH4	25-100	no	600 mTorr	No	no		500-1000 A/min

Figure 16. Plasma deposition processes, 1. Courtesy of Trion Technologies, Inc.

Deprocessing of Low K Dielectric materials presents a special problem to the analyst. The Low K Dielectric Materials are grouped into 4 different categories. Each one has their own special properties. The groupings are shown in Figure 17.

1. Doped oxides:
Silicon Oxyfluoride (FSG), Hydrogen Silsesquioxane, (HSQ), MSQ, HOSP, CVD.
2. Organic dielectrics:
DVS-BCB (divinyl silozane bis-benzocyclobutane), SILK, FLARE, PAE-2.
3. Highly Fluorinated dielectrics:
Parylene AF4 (aliphatic tetrafluorinated poly-p-xylylene), a-CF, PTFE (Polytetrafluoroethylene).
4. Porous Dielectrics:
Aerogels, Xerogels, Nanoglass (Nanoporous silica).

Figure 17. Groupings of Low K dielectric Materials.

The mechanical properties of the Low K Dielectric films are of great concern. The films are high in tensile stress. When overlying compressive layers are removed, the Low K dielectric layer may weaken and crack. The crack, which is really a deprocessing artifact, may be misinterpreted as a real defect.

Delayering of Low K Dielectric Materials can be accomplished by any of the means discussed. One of the most prevalent methods to remove the Low K Dielectric is to use a plasma etch.

The plasma etch recipe for doped oxides (FSG, HSQ, and MSQ) uses a variant of the SiO_2 plasma recipes. The plasma etch generally involves CHF_3, CF_4 and O_2. A common mixture is 80% CHF_3 and 20% O_2. The ratio of CHF_3 to O_2 should be varied for the particular material being etched.

Organic/Highly Fluorinated Films (DVS-BCB) typically use O_2 or H_2 plasma etch chemistries. An example of a typical etch would be 90% O_2 with 10% SF_6 at 200Watts and 400mTorr. This process can also be used to etch highly fluorinated films.

As the need for lower K dielectrics becomes more important, porous films may enter the market. Made from basically the same materials with different porosity to lower the dielectric constant value the etch process of these new materials will remain some what unchanged. The added porosity should induce a higher etch rate within the material. The porosity may cause un-wanted etch artifacts for plasma etching such as pitting of the underlying materials due to punch through of the plasma etch. Other techniques may be required to etch the material of interest.

PARALLEL POLISHING

Parallel polishing is used to mechanically remove overlying materials such as metals and insulators by using a polishing wheel rather than wet chemistries or plasmas. Parallel polishing can be quicker than ether wet chemical or plasma processes when setup time and cleanup are considered. Parallel polishing can cause less damage to the underlying layers than the wet chemical or plasma processes. As an example of when parallel polishing is most helpful consider an

EOS event. Such an event causes extreme heating of the materials in the device. The heating changes the characteristics of the material in the region of the EOS event.

When a wet etch is used to deprocess the device the region of the heating etches differently than the unheated region leaving etch artifacts. The same etching differences are seen with plasma etching. What is worse in the case with plasma etching is that the etch can pit the substrate leaving etch artifacts which look like the location of the root cause of the failure.

The disadvantage to using parallel lapping is that the defects that occur in the material being removed may be missed if the device is not frequently examined. In fact stopping in the middle of a layer being removed by lapping is critical to finding defects. The appropriate use of optical microscopy such as, bright field microscopy, dark field microscopy and phase contrast when inspecting the material removal process is critical. Don't use just one inspection technique as each technique can see a different aspect of the material removal process.

Bright field microscopy can see the surfaces but can be blinded by the glare from the same surface. Dark field or phase contrast has the advantage of being able to see through the glare of the surface to underlying material edges but loses the color information of bright field microscopy.

Scanning electron microscopy (SEM) inspection of partially lapped or delayered materials requires an understanding of the proper beam accelerating voltages which should be used to penetrate the un-removed material. As an example if the dielectric overlying a metal line were to be removed completely a metal filament from one metal level to another might be lost. As happened in one device the W contacts were shorted together through filaments in the oxide. "Standard" deprocessing lapped through the oxide as there had never been any defects found in this layer. Inadvertent inspection of the oxide layer, when it had only been partially lapped, revealed shorting of the W contacts. To inspect the layer the SEM voltage, 1KeV, was set high enough to penetrate the oxide remaining above the W filament. A higher voltage would blast through the defect because of its small size. A lower voltage would not penetrate the oxide.

So why should you parallel lap? Figure 19 gives some reasons to parallel lap the device.

One important reason to parallel lap modern devices is that the metal layers are already planer. Plasma etching leaves ridges that cannot be removed by more plasma etching or wet chemical methods. Only parallel lapping removes the edges.

WHY PARALLEL POLISH?

- **Multilevel metallization**
- **Eliminates Etch Artifact**
- **Superior Surface planarity**
- **Cycle-Time**

F*igure 19. Reasons to parallel polish.*

Lapping can introduce its own artifacts, see Figure 20. In Figure 20 the lapping artifacts at the bottom of the image are caused by the round off of the edge of the die due to the lapping technique. The streaks in the lapping are caused by the presence of Al pads from the solder bumps. The Al pad locations lap faster than the rest of the passivation surface. The lapping artifacts are easily recognized so they may present less of a problem than wet chemical artifacts or dry etching artifacts.

Lapping artifacts can be limited to the surface. If the lapping process is performed with care a lapping artifact can be removed by the time the next layer is exposed. Plasma etching and wet chemical etching tend to propagate the etch artifact into the underlying materials.

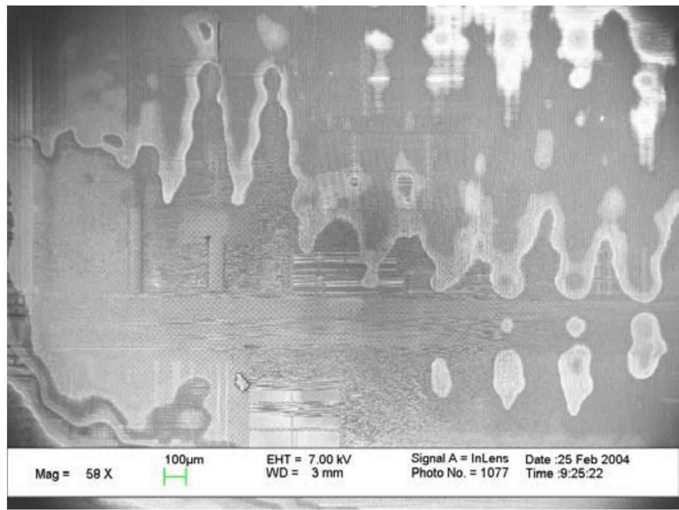


Figure 20 shows lapping artifacts at the edge of a die due to round off of the edge and the presence of solder bumps across the surface of the die.

Figure 21 shows a lapping process where the sample is placed in an encapsulating material. The resultant puck of material is mounted in the lapping system where it is moved across the surface of the lapping plate in an oscillatory pattern while the plate rotates underneath.

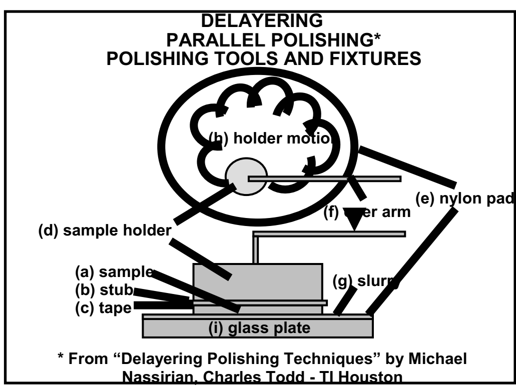


Figure 21. Parallel lapping tool and fixtures.

Current lapping technologies can be much more sophisticated. Some companies use sample holders that hold the device in a fixed position similar to the manner in which a diamond is cut. Other companies use sample holders that are self-leveling. Yet other companies have sample holders that ride on the lapping platen. There does not seem to be an end to the variety of ways companies can find to mount the device being lapped.

Figure 22. gives a process flow for parallel polishing. In Figure 22 1um Alumina is used to lap the surface. There is much discussion about the use of 1um alumina. Some say that diamond is better than alumina. Only trial and error on the particular material being lapped will answer the question.

**DELAYERING
PARALLEL POLISHING FLOW**

- Remove die from package.
- Remove wire bonds or bumps.
- Remove any organic die passivation.
- Lap hard passivation (nitride, oxide) with 1um alumina paste.
- Lap metal with 1um alumina paste.
- Repeat until at substrate.

Figure 22. Parallel polishing flow.

Now that wet chemical, plasma and lapping techniques have been discussed please take note that setting a "standard" delayering process may require a combination of wet, dry and lapping techniques to be take advantage of the benefits of each type of technique while eliminating the delayering artifices generated by the other technique used.

Modified Process Flow

- Remove die from package with a wet chemical process as appropriate.
- Remove wire bonds or bumps by lapping.
- Remove any organic die passivation with nitric acid.
- Lap hard passivation (nitride, oxide) with 1um alumina paste.
- Remove dielectric around the metal line by plasma etching.
- Remove metal by lapping.
- Repeat until Metal 1 has been removed.
- Plasma or wet chemical etch the multilevel oxide over the poly. Depending upon need this layer could be lapped until the poly is exposed.
- Wet chemical etch the poly with Choline.
- Wet chemical etch the oxide with 49% HF.
- Etch the substrate with Wright-Jenkins etch or other appropriate defect delineation etch.

Figure 23. Modified process flow allowing for complimentary processes.

A CROSS SECTION PROCESS

To understand the delayering process a look must be taken at the layers that are to be removed. Lets start by looking at the cross section of a solder bump, Figure 24.

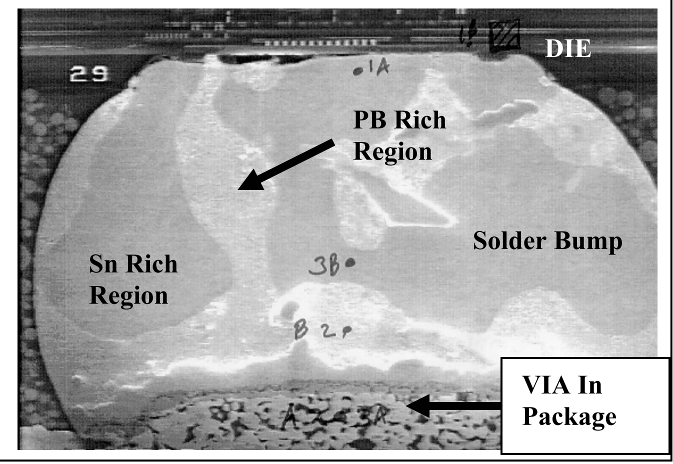


Figure 24. Cross section of a solder bump in a package.

In Figure 24 the package was lapped with 180 grit emery paper until the bump started to show. The paper was then changed to 360 grit emery paper until near the target. An 800

grit emery paper was used to come up to the target. When the target was at hand the grit was again changed this time to 1500. A final polish was performed using colloidal silicon with a pH of 10.5. The colloidal silicon acts as a stain to bring out as a relief the Pb, Sn and inter-metallic.

The cross section of figure 25 was performed in the same way as that of the cross section of the solder bump shown in Figure 24.

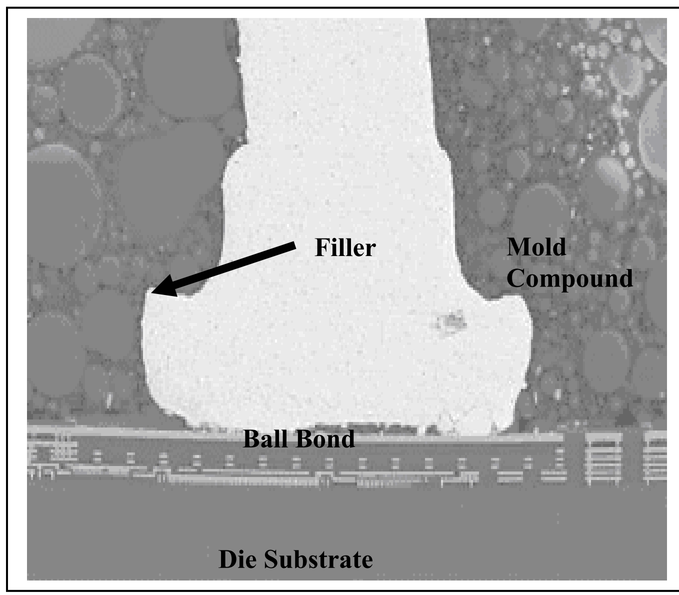


Figure 25. *Cross section of a ball bond in a package. Image courtesy of Omar Diaz de Leon of Texas Instruments.*

Now lets look at the cross section of the semiconductor device. To properly view the cross section the section must be stained to bring out the detail. In the case of Figure 26 a wet silicon etch, stain, was used to highlight the implanted junctions. The stain brings out the diffusions, well structure, EPI layers and stacking faults

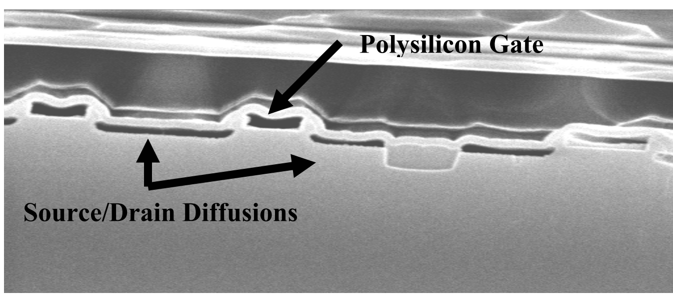


Figure 26. Cross section of a device showing the junctions highlighted by one of the stains. courtesy of Michael Strizich: Analytical Solutions, Inc.

A plasma etch can be used to stain the cross sections. The etch uses CF_4 at 100 Watts for 30 sec. at a gas pressure of 300 mTorr.

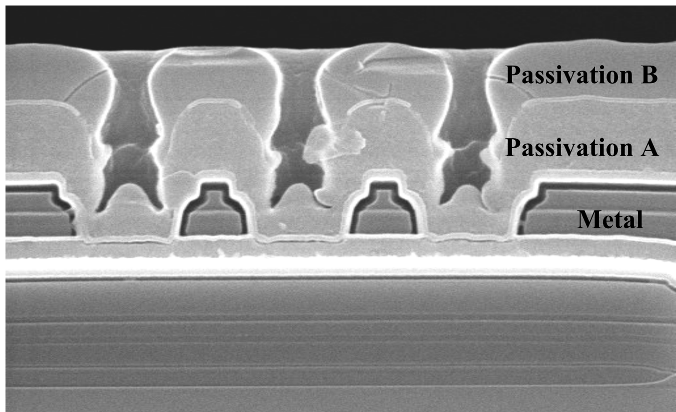


Figure 27. Image shows a cross section etch by a buffered oxide etch. Image courtesy of Michael Strizich of Analytical Solutions, Inc.

Figure 28 shows a typical cross section of a device that might be deprocessed.

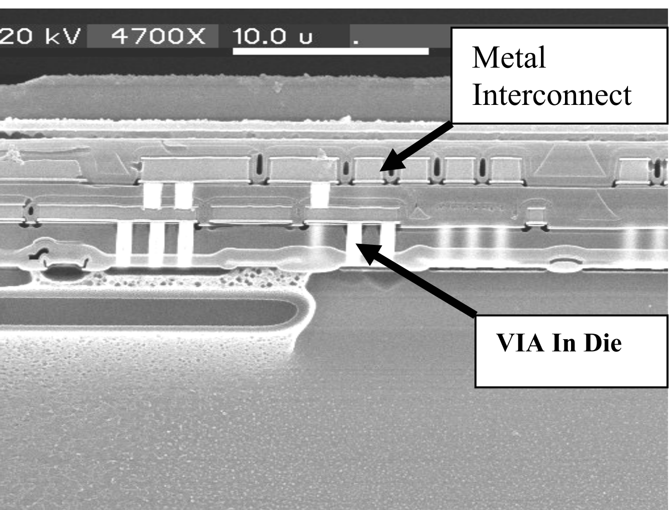


Figure 28. A typical cross section of a device that might need to be delayered. Image courtesy of Michael Strizich of Analytical Solutions, Inc.

Deprocessing of a 5 metal level system.

Now that wet chemical etching, plasma etching and lapping has been introduce as material removal techniques and a knowledge of the layers to be deprocessed has been gained we can deprocess a device. For the purposes of this deprocessing recipe the information was compiled by Charles Todd and Rina Chowdhury of TI Houston

The first level to deprocess is the passivation. There may be an organic passivation above a hard passivation. The organic is removed by nitric acid. The hard passivations can be made up of Nitrides, oxides, oxynitride or a combination of layers. For the purposes of this recipe the top level passivation will be a nitride on top of an oxide.

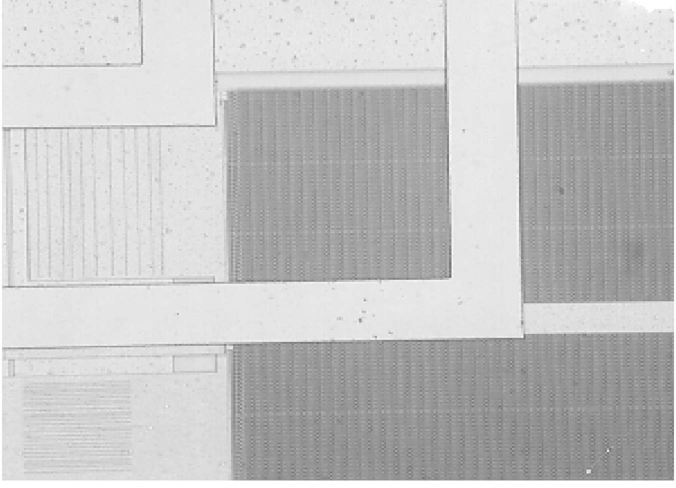


Figure 29. Top metallization before removal of the passivation.

The hard passivation is normally removed by a plasma etch. The first step is to remove the nitride.

Step1: Pressure: 320 mtorr
Power: 150 watts
Nitrous oxide: 20
CF_4 : 80 sccm
Time: 600 seconds.

The second step is to remove the oxide.

Step2: Pressure: 200 mtorr
Power: 150 watts
CHF_3 : 50 sccm
CF_4 : 10 sccm
Time: 500 seconds.

The nitride passivation layer has a yellow tint. After a complete etch the unit should appear silver due to the exposure of the Al level. Note: The passivation removal process will also remove top Ti and TiN layers of Metal 5.

There are two alternative methods of material removal. A wet etch could be used. Step one of the wet etch would be to remove the nitride as described previously.

Step two would be to use 5% HF to remove the oxide.

Lapping could be performed at the passivation layer. To lap off the passivation use 1um Alumina.

Next the metal layer must be removed. Figure 30 shows the die before metal removal.

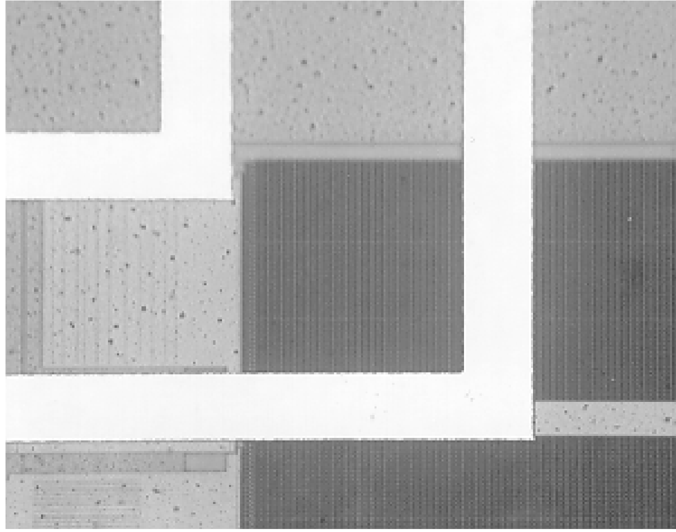


Figure 30. Image of the device after Passivation etch and before Metal 5 etch.

The plasma recipe for the Al etch is:
Power: 120 watts
BCl_3: 40 sccm
Cl_2: 10 sccm
Time: 200 seconds.

After metal etch place the unit in COE for 30 seconds for clean up.

There are two alternative materials removal processes for Al. The first is to use a wet etch: Warm HCl for Al combined with NH_4OH: H_2O_2 for TiN.

The Al can be removed by lapping with 1um Alumina paste. If the metal is Cu it will most likely require a lapping technique to remove it. At this time plasma and other Cu removal techniques are experimental.

Once the metal is removed the oxide is exposed. If a plasma etch or a wet chemical etch was performed there may be a need to lap the top surface of the oxide. Both processes can leave a ridge that must be removed. Figure 31 shows the die after the metal etch ready for the oxide etch.

A plasma etch can be used to remove the oxide. The plasma recipe is:

Pressure: 300 mtorr
Power: 200 watts
CHF_3: 100 sccm
CF_4: 50 sccm
Time: 700 seconds.

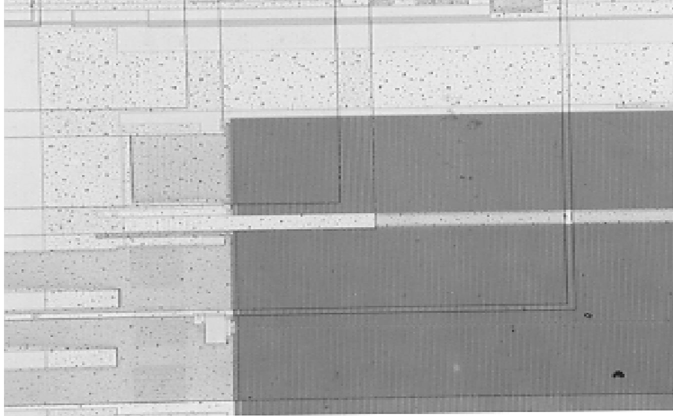


Figure 31. Image of the device after removal of the metal before oxide removal.

Again there are two alternate processes to remove the oxide. The first is a wet etch: Buffered oxide etch or VAPOX etch. The second option is to remove the oxide by lapping. A lap has the advantage of removing any etch artifacts left by the previous process.

We are now back at a metal level. To remove Metal 4 the same steps are used as Metal 5. Figure 32 shows the device ready for Metal 4 etch.

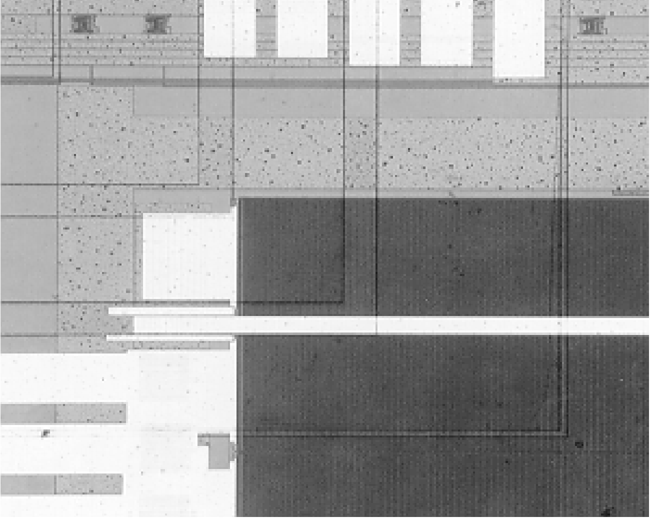


Figure 32. Image of the device ready for Metal 4 removal.

When removing the various levels do not assume the levels are the same thickness or material. For example the metal levels are thinner when closer to the device substrate. The metal lines are also smaller the closer you get to the substrate.

Oxides follow the same type of rule, the closer to the substrate the thinner the oxide. Also the oxide structures are smaller in width closer to the substrate. Oxides have another important distinction. The type of oxide may change from layer to layer. Top layers of oxide are typically harder to provide the strength required for bonding of the device to the outside world. The oxides may be a Low K Dielectric closer to the substrate. The oxide over the poly gate will be a doped oxide. The oxide around the poly gate itself will be a thin cap oxide that is a hard un-doped thermal oxide. Each oxide will require a different method to adequately remove the material.

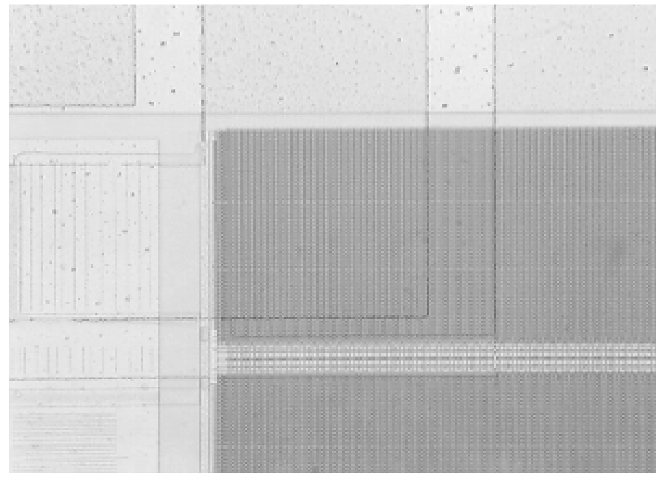


Figure33. Image of the device ready for Dielectric 3 removal.

The standard oxide etch could be used. There are a couple of alternatives. The first is a plasma etch if the dielectric is a Low K Dielectric. The process is:

Pressure: 50mtorr
Power: 400watts – ICP
Power: 200watts - RIE
CF_4: 45sccm
O_2: 15sccm
Time: 30 seconds

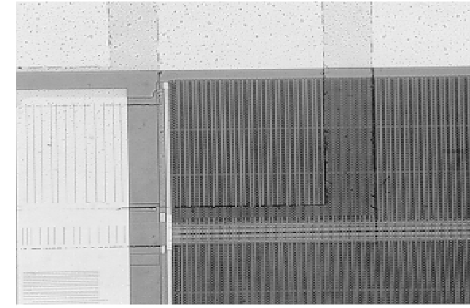


Figure34. *Image of the device ready for Metal 3 removal.*

The same processes can be used for Metal 3 as the previous metal levels.

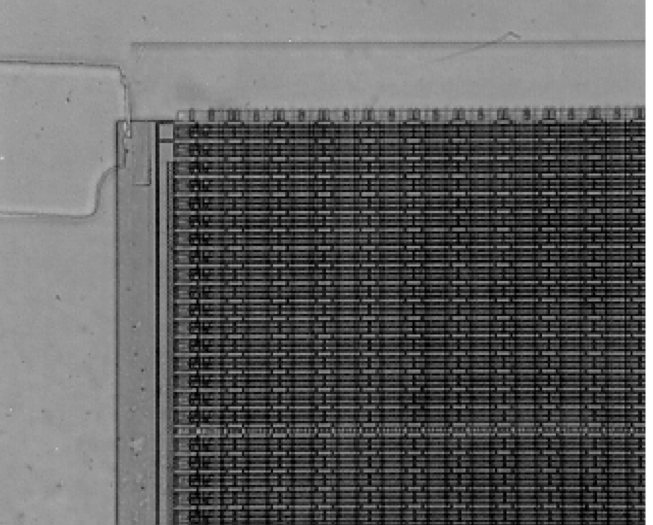


Figure 35. Image of the devce ready for Dielectric 2 removal.

The Dielectric 2 can be removed as in the previous steps.

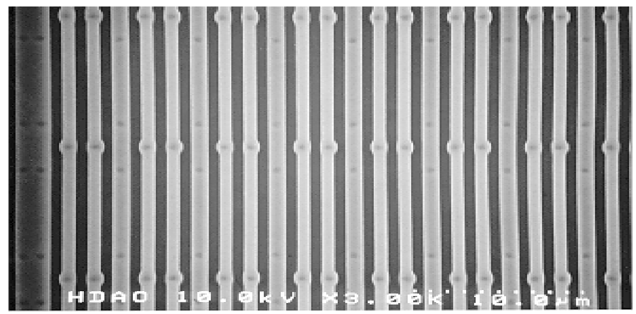


Figure 36. Image of the device ready for Metal 2 removal.

The same metal removal processes are used to remove Metal 2.

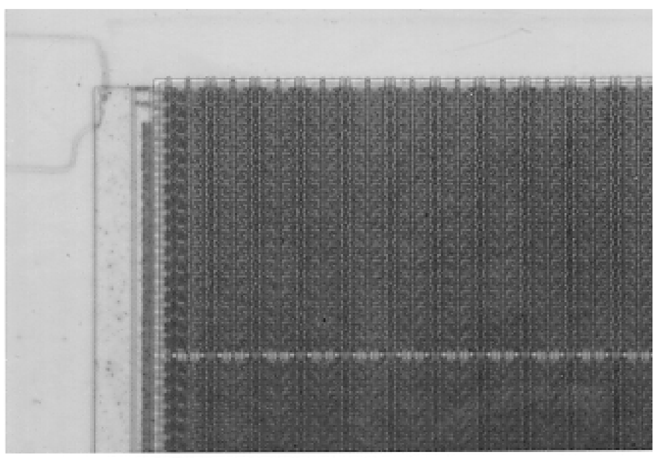


Figure 37. Image of the device ready for Dielectric 1 removal.

The Dielectric 1 can be removed in the same manor as the other dielectrics in the stack.

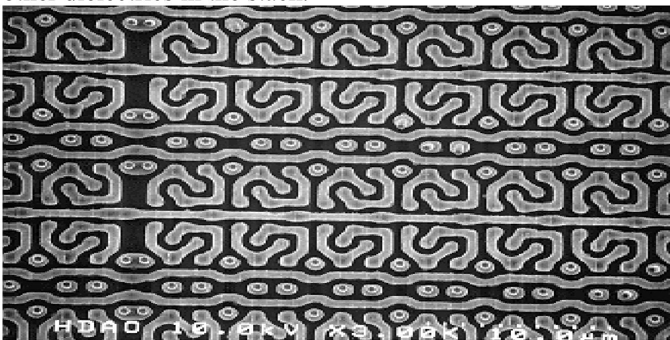


Figure 38. image of the device ready for Metal 1 removal.

The Metal 1 level can be removed the same as the previous levels.

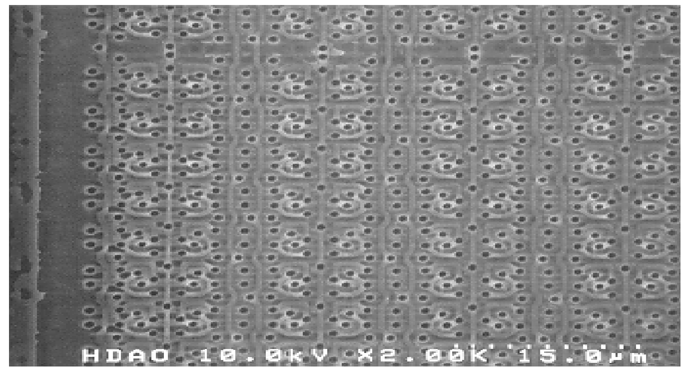


Figure 39. Image of the device ready for multi level oxide removal.

Use 5% HF for 7 minutes at room temperature to remove MLO.

Take special precautions to etch the oxide slowly to ensure that the sidewall oxide remains intact. After etch, optical/SEM inspect the device to make sure all oxide is removed. Low SEM beam energy is important to prevent annealing the semiconductor materials so the etch properties do not change.

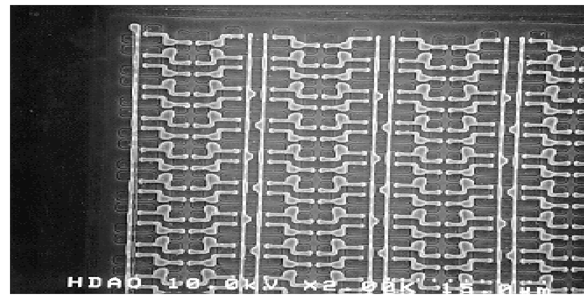


Figure 40. Image of device ready for Polysilicon removal.

For polysilicon gate removal use dilute poly etch for 4 second intervals for a total of 8 seconds: Inspect the device between etch interval to make sure that poly is etched.

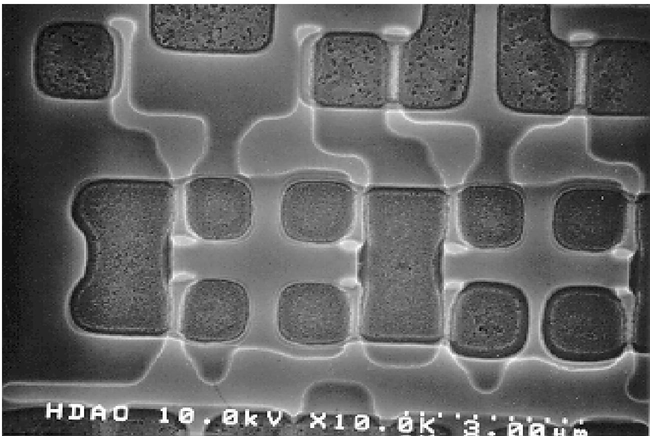


Figure 41. Image of the device at Gate Oxide.

Use 49% HF for 5 minutes at room temperature to remove the remaining oxides. The device will be at substrate level. Using Wright Jenkins, etch the die for 10 seconds to decorate the structure.

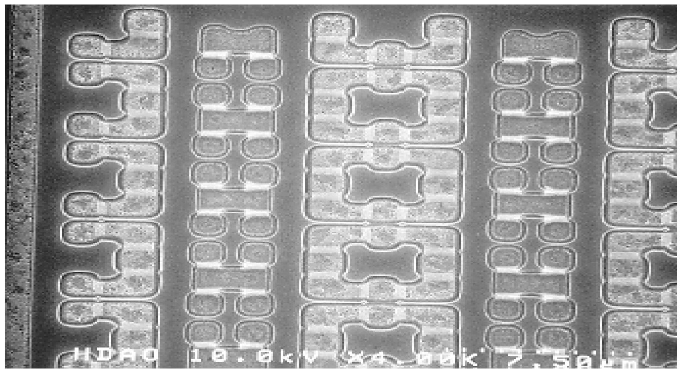


Figure 42. Image of the device at the Substrate level.

The substrate is etched to bring out possible defects such as Oxidation Induced Stacking Faults (OISF) in <100>Si. To bring out the OISF use Wright Etch for ~30secs. Wright etch is a mixture of HF, HNO_3, $Cu(No_3)_2$, CrO_3 and CH_3COOH.

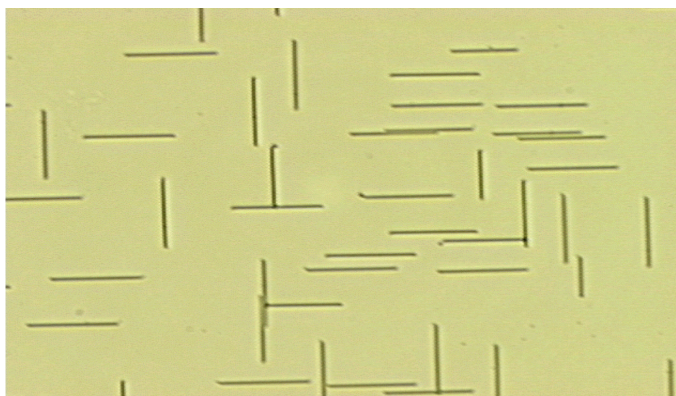


Figure 43. Image of the dislocation found after Wright etch

Dislocations in <111>¤ and <100> Si are stained out using Sirtl etch. Sirtl etch is a mixture of HF and CrO_3 in DI water The Wright etch and Sirtl etches can be found in *VLSI Tech., McGraw-Hill, 1988.*

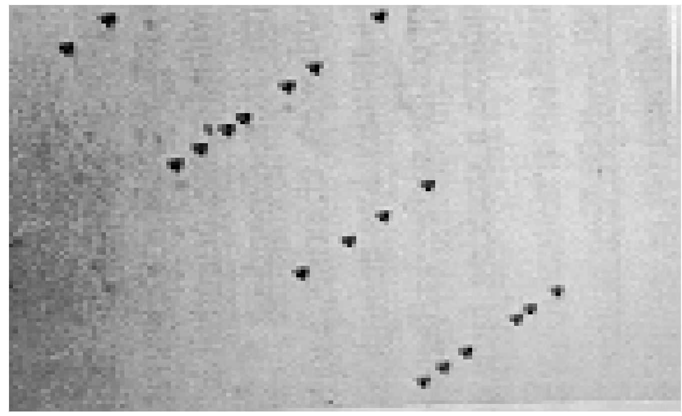


Figure 44. Image shows dislocations in <111> Si stained out using Sirtl etch.

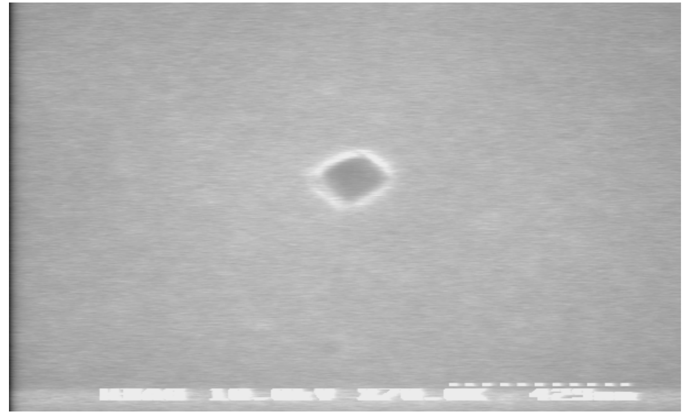


Figure 45. Image shows dislocations in <100> Si stained out using Sirtl etch.

Backside Silicon etching can be an important step in the overall deprocessing flow. When devices have multi-levels of metallization access to the gate and diffusions areas may be difficult. The structure of the semiconductor may preclude top down deprocessing to see the gate oxides, as is the case with trench capacitors. For Flip Chip devices access to the front side of the device is blocked. The easiest way to reach the transistor level is through the backside of the device. Also due to the possibility of deprocessing induced artifacts from the topside a backside approach may be more appropriate. The backside etch may be more appropriate for special failures such as gate oxide defects where the top down deprocessing will remove the evidence of the defect.

A typical backside approach would start by parallel polishing the silicon substrate with 1200 grit emery paper down to 100um of the active transistors. At that point the wet silicon etchents mentioned earlier can be used. There are many etchents which can be used but the two preferred are Choline(Trimethyl Hydroxyethyl Ammonium Hydroxide) and TMAH(Tetramethyl Ammonium Hydroxide). Figures 46 – 48 show the results of a backside etch.

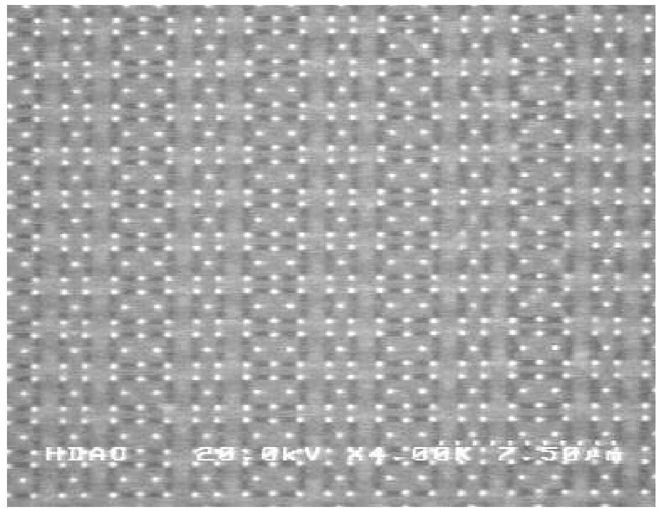


Figure 46. SEM micrographs of the backside of a device showing the silicon etched away from the SRAM array. The etchent was TMAH.

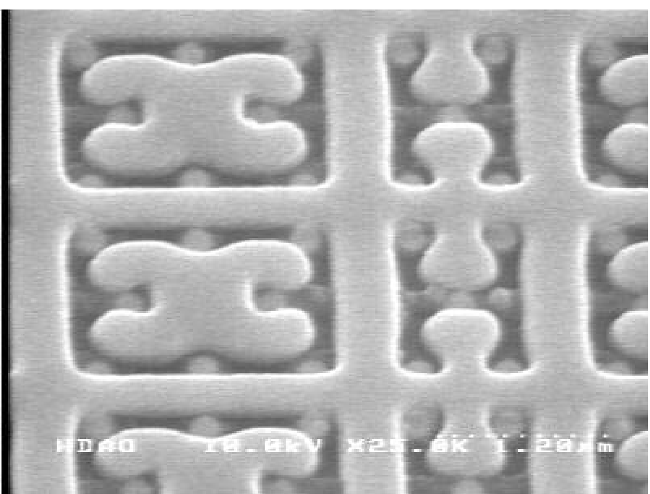


Figure 47. SEM micrograph of the backside of a device at a higher magnification showing the silicon etched away from the SRAM array. The etchent was TMAH.

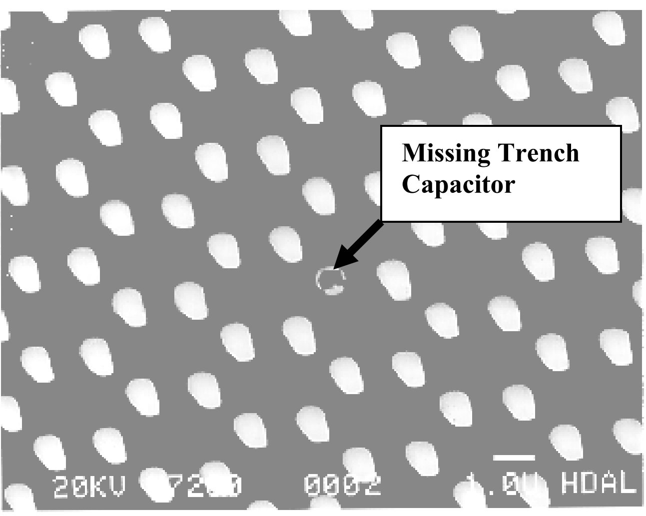


Figure 48. SEM micrograph of the backside of a device showing the trench capacitor is missing from the DRAM array. The etchent was Choline.

SPECIAL PROBLEMS – COPPER ETCHING

Copper etching presents some unique problems. First it is incompatible with conventional dry etching. Second for both wet and dry etching techniques there are no natural etch stops. A new approach on copper deprocessing is required.

Wet chemical approaches to copper etching are given the Figure 49.

Wet chemical copper removal recipes

1. **Ammonium Hydroxide 10 ml**
   Hydrogen Peroxide 10 ml
   DI Water 10 ml
2. **Potassium Dichromate 100 g**
   DI Water 100 ml
3. **Ferric Chloride 5-30 g**
   DI Water 100 ml

Figure 49. Copper wet etch recipes.

There are commercially available etchents for copper from the printed circuit board industry. Transcene printed circuit copper etchents and times are given in Figure 50.

Transcene Printed Circuit Copper Etchents

	CE-100	CE-200	APS-100
Type	Immersion	Spray	Immersion/Spray
Temp	40-60 C	40-60 C	30-40 C
Rate (@40C)	1 mil/min	0.5 mil/min	80 A/sec
Capacity	19 oz. Cu/gal	14 oz. Cu/gal	0oz.Cu/gal

Figure50. Transene copper etchent and etch parameters.

Copper etching based upon a plasma process has been suggested by the Korean Advance Institute for Science and Technology (KAIST). The Institute suggests using a high intensity UV light source, ICP and heated chuck. The theory is that UV light stimulates de-adsorption of the Copper Chloride from the device surface replacing the need for high temperatures. Unfortunately, a plasma Cu etch process will have the same problem with the lack of an etch stop as a wet etch.

A possible plasma etch process for Cu would be:

RF Power ICP: 1300 - 1600 Watts
RF Power RIE: 250-300 Watts
BCl3: 40 sccm
Cl2: 10 sccm
Electrode Temp.: 180 Degrees C.
Chamber Pressure: < 50 milliTorr
UV Light: 300 Watts/sq

Figure 51 shows Cu removed using the plasma etch process.

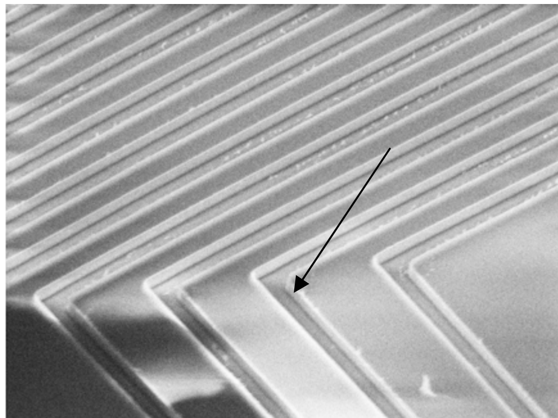


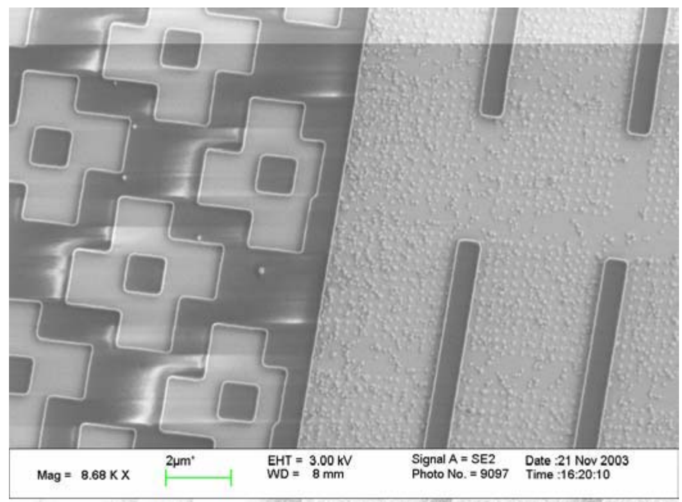
Figure 51. Image of plasma etched Cu metal lines. Arrow points to the Cu line etched. Image courtesy of Trion Technology, Inc.

The current accepted Cu removal process is mechanical Polishing. Lapping has the advantage of accommodating inspection in mid-layer removal. The disadvantages of lapping are the increase in labor requirements and the need for constant inspection to ensure complete material removal.

NEW TRENDS IN DELAYERING/DEPROCESSING

Wet chemical deprocessinng, lapping and plasma etching will remain the main stays for delayering for a long time to come. New fabrication processes, new materials and the need to deprocess without artifacts on new technologies will drive the development of new techniques.

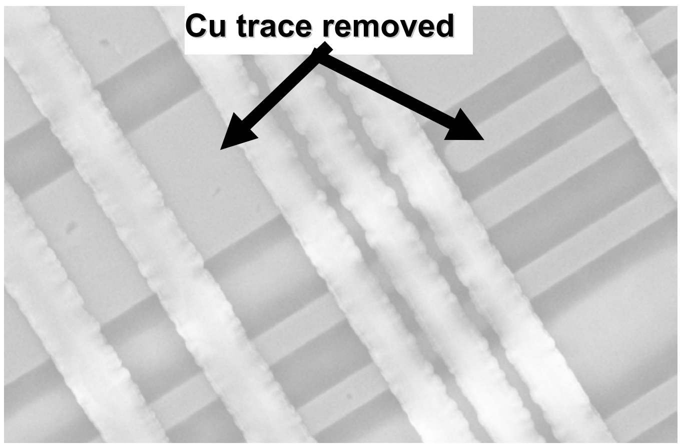
Even the tried and true wet chemical techniques can be improved. There is a new etch chamber which offers a controlled way to etch materials. In Figure 52 the passivation was removed by lapping. The Cu was removed by the new chamber etch process.



52. Image shows Cu etched in a new etch chamber using a new wet chemical process. Image courtesy of Nisene, Inc.

One technique, which has been around for a long time but never fully developed, is Ion Beam Milling. There have been big tools for milling substrates and little tools for dimpling samples for transmission electron microscopy. There has never been a tool for Failure Analysis deprocessing until recently.

A new tool is being developed which floods the surface of the device with ions. Unlike the RIE, which also uses low energy ions, 200 eV, to etch the energies of the new ion etch technique uses ions that are in the 1000's of eV. Like the RIE the energy of the ions can be tailored to meet the deprocessing needs. An example of the results obtained for a Cu etch using a ion beam is shown in Figure 53.
Cu is difficult to remove as has been discussed earlier. As can be seen in the SEM image of Figure 53 the Cu is cleanly removed without damage to the surrounding oxides.



Fig*ure 53. Optical micrograph of Ion Beam Milled device showing removed Cu. Process by Advanced Technologies, Inc.*

Speed of processing is important in backside silicon removal. Parallel lapping techniques are fast but lack the ability to be specific. There is also no end point detection for a lapping process. The analyst must constantly measure the thickness of the sample ether by optical means or by mechanical means. Optical measurements rely on the ability of the microscope to distinguish between the focal point of the device surface and the surface of the backside of the silicon. This technique is error prone due to the inaccuracies of the optical microscope. Mechanical measurement of the thickness of the silicon is hampered by lack of information as to the original thickness of the silicon. This is especially true for flip chip devices where the solder bumps are part of the measurement. The standoff of the bumps is not constant from device to device.

To alleviate this problem laser backside etching has been developed. A small region as shown in Figure 54. can be removed. If required larger regions can be etched.

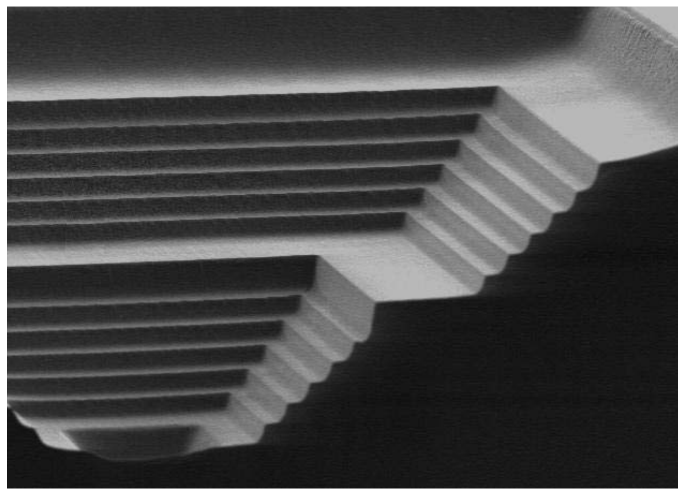


Figure 54. Backside Silicon Etch by laser processing. Photo courtesy – FEI, Inc..

The laser has the advantage that there is a natural end point detection system possible. Optical beam induced current (OBIC) measurements can be taken from the semiconductor device. As the OBIC current rises the surface of silicon being etched is closer to the active transistors. When the OBIC current reaches a threshold value known for the specific device the laser process is stopped.

The most common laser chemical backside removal process uses Cl_2. Other chemicals could be used such I_2, Br_2 and H_2. One company uses only laser ablation, no chemicals, to remove the backside silicon.

Of the alternate deprocessing techniques being developed the FIB deprocessing with gas-assisted etching is becoming mainstream. The tool is expensive from a monetary point of view but from a results point of view it pays for itself.

The FIB can use several gasses depending upon the need. For example XeF2 is used to improve the mill rate of oxides. To remove metals I_2 or Br_2 can be used to improve metal removal as well as Cl_2. For removal of polymers H_2O or O_2 can be used. Figure 55 shows the corner of a semiconductor device. The silicon substrate has been removed showing the device structure in the corner.

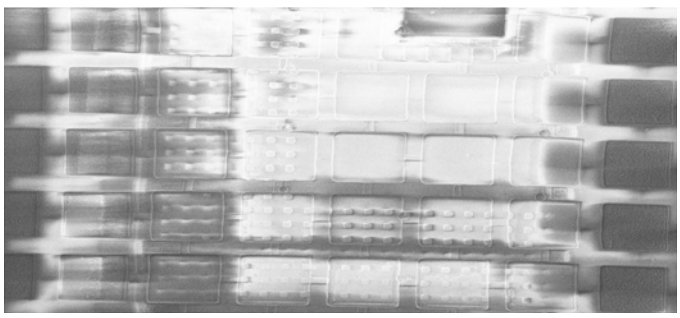


Figure 55. Backside Silicon Etch by FIB processing. The image shows the ability of the FIB to expose an entire device level for inspection. Image courtesy of Reena Agarwal of Texas Instruments.

ACKNOWLEDGEMENTS

- Seshu Pabbisetty of TI-Houston for technical advice in preparing this paper.
- Michael Strizich of Analytical Solutions for his help concerning chemical and plasma etching.
- Charles Todd, Rina Chowdury of TI-Houston for deprocessing photos and info on 5 level systems.
- Michael Nassirian, Charles Todd of TI-Houston for parallel polishing information.
- Steven Nguyen of TI-Houston for photos of parallel polishing process.
- Darwin Rusli TI-Houston for help in preparing the paper.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

The Art of Cross Sectioning

B. Engel, E. Levine, J. Petrus, and A. Shore
IBM Microelectronics, Hopewell Junction, New York USA

Introduction/Laboratory Logistics:

A modern semiconductor facility that is involved with day-to-day production as well as development is highly dependent on a facility that can rapidly produce high quality SEM images of products being manufactured. The ability to cross-section targets such as individual contacts as small as 0.15 microns (or less), while often taking several sequential sections through such targets, is important when looking for small defects and process detail. Such data provides input to engineers and technicians on the quality and general characteristics of the structures being built as well as the origin of defects that detract from yield.

The laboratory where such work occurs can be organized in many different ways, however, at IBM East Fishkill we have found that the most productive method is to have a well-trained cadre of technicians, each with the ability to prepare specimens using a variety of techniques. The most important of these techniques, namely cleaving and polishing of cross-sections, will be described in this paper. We find these two techniques to be the most cost effective and accurate means to accomplish the described objective.

The first step of the process is sample submission. The system is designed to allow an engineer/technician to simply fill out an on-line request form specifying the product, details of the process, where the job was intercepted, and desired result. The output requested is usually a cleave or polished cross-section. The customer supplies documentation that specifies the exact location on a product layout map, chip location on wafer, and usually a schematic drawn to show what the section should look like. Additionally, details of proposed surface preparation can be given. After the request is logged in, the completion of the process can take a few hours or up to 5 days depending on the severity of the problem as determined by the priority of the request. The final product is generally a set of micrographs that is given to the originator to keep as a permanent record. Additionally, the results may be posted on-line for ease of communication to the team members.

Work requests can come from a variety of sources, but typically they will originate from the process community who simply want to assure themselves that the structures being built meet specifications, i.e. the proper slope on via sidewalls, structure shape, trench dimensions, etc. More directed requests occur when a particular structure or group of structures has failed a required specification during in-line process testing. In this case, the requester may need to cross section an individual contact that is deemed bad or perhaps just sample the entire structure to see if a systematic problem is present. Another frequent source of sample input originates from defects found in line by inspection techniques. Here the requester need only give the chip and x,y location of the defect. The SEM lab technicians are trained to locate and laser or FIB mark the defect location in preparation for cross sectioning. Both the analyst and the submitters must have a high degree of familiarity with the product structure and layout to assure a satisfactory output. It should be noted that the laboratory described runs somewhat independently of specialized failure engineers whose job it is to localize and analyze fails of different types. These failure analysis engineers will utilize the lab like any other engineer submitting their request in the usual manner.

The skills necessary to complete an analysis must be possessed by a sufficient number of technicians to keep the analysis output times reasonable. Our site, which is heavily involved in development of multiple technology products, requires 10 full time analysts and 10 fully occupied SEM's (24 hr coverage) to keep the samples flowing. This group of men/women and machines supports roughly 1500 development/production engineers at this facility. Training of these physical failure specialists (PFS) is a major challenge since we are continuously seeking new talent. It is hoped that this paper will assist in the

training and development of routine sample preparation and manual cross-sectioning skills. This of course, must be followed by hands-on practice. It is for this reason that we have prepared this paper with painstaking detail, trying to point out particulars both large and small that need to be considered in the quest to achieve a topnotch image.

Preparation techniques may vary from analyst to analyst. This paper presents a polish technique that produces high quality SEM images on a routine basis, in anywhere from 30 minutes to several hours per specimen, depending on the skill level of the PFS and the difficulty of the job. Cross-sectional polishing is one of the most highly skilled operations performed (in the lab), and is constantly faced with increasing complexity as feature sizes continuously decrease. However, the polished cross-section remains a major part of our laboratory output (> 50%). It should be recognized that it is difficult to describe how to perform an operation such as a polished cross-section in words and illustrative pictures alone. To help make the instructions clearer, we will often describe what may seem to be very mundane aspects of the operation. These details are better said than left out since one never knows what minor parts of the operation can give a person trouble. We do not expect that you will be able to perform these operations in initial attempts, but we hope this foundation will eventually lead to success.

Structures:

For the purposes of this article, we will define the semiconductor chip as a build that consists of two portions. The first is the Front End Of the Line (FEOL), which deals with the devices, the Silicon, Silicon implants, trench isolation structures, and gates. The second section is the Back End Of the Line (BEOL), which contains the connections to the FEOL and the outside world. In general, the BEOL involves multiple levels of metal lines embedded in a dielectric, interconnected with metal vias. This entire structure is a three dimensional building, so to speak, however, the analyst who does the cross section is viewing this structure in a two dimensional view. In order to be a successful PFS, one must realize a two dimensional view with the realities of the three-dimensional build. We will give one example here that will allow the trainee to understand the concept.

Simple Lines and Via structures- 2 vs. 3 dimensional visualization

Two basic structures in a semiconductor chip are lines and vias. Metal lines are contained within one horizontal level of the chip and may traverse large distances. In cross-section, they are usually simple rectangles or squares. Vias are metal structures that connect lines at one level to those on the level above or below. In cross-section, these structures may be conical or rectangular with the long edge in the vertical direction. If one looks optically top down at a via chain, it appears as a series of small lines at multiple levels. These levels are connected by vias that cannot be seen top down because they are obscured by the metal lines. A via chain will repeat this pattern over and over again in an attempt to mimic the product where current is similarly carried from level to level. A top-down SEM of a via chain is shown in Figure 1a. Only the dog-bone shapes of the upper-most lines are visible since the SEM sees only the specimen surface.

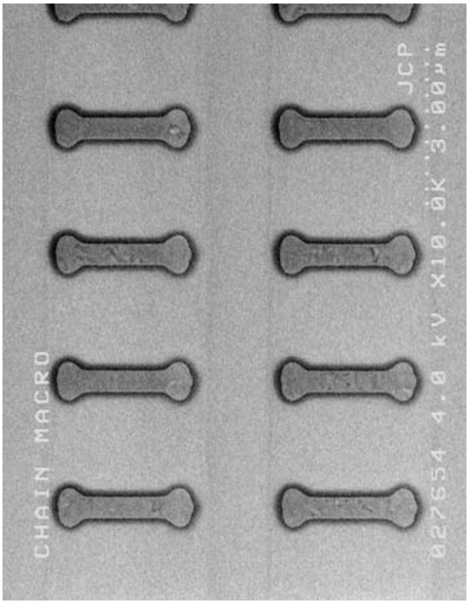


FIGURE 1a: Top-down SEM of via chain

The optical view of the same via-chain is shown in Figure 1b. Here, we can see multiple levels with a varying degree of clarity.

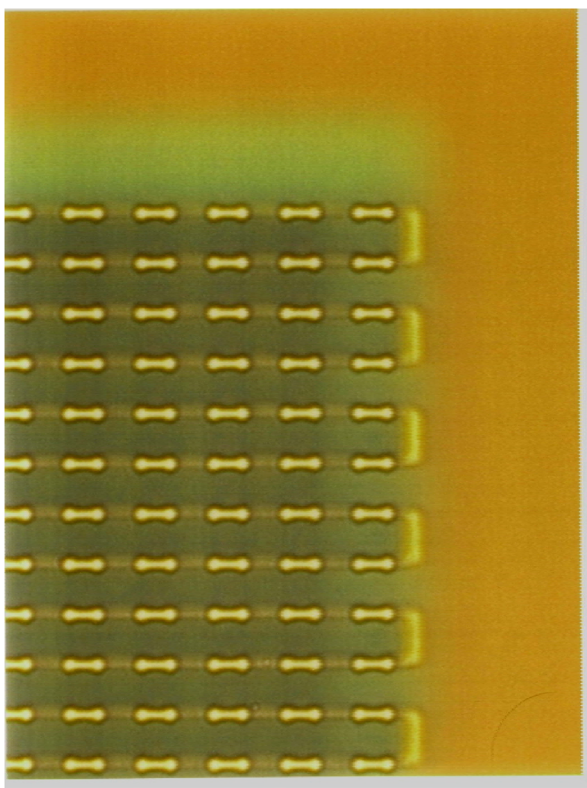


FIGURE. 1b: Optical top down of via chain

In an optical microscope, the buried lower level will focus somewhat below the upper metal because it is on a different focal plane. Again, the via cannot be seen from top-down because it is obscured by the upper-most metal lines. If however, one were to cross-section this structure, the actual build of the vias and lines would become immediately obvious. We then see that the wire is a finite conductor and the via is a vertical connection between levels, as shown in the SEM in Figure 2.

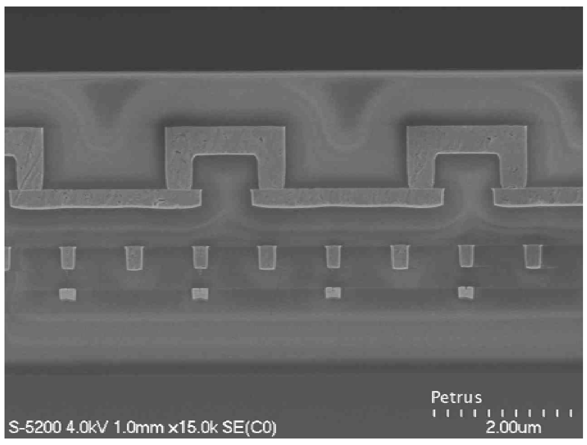


FIGURE. 2: Cross section of a via chain.

In the image above, one cannot tell that the via is, in reality, a cylinder. If however, one were to look at sequential images taken as we section through the contact, the true structure would become clear. In the sequential images, the lines in cross section will remain a constant size. On the other hand, the via will start out small as we begin to intersect it and gradually increase in size until we work our way to the center, where it reaches a maximum. Of course it will then begin to decrease in size as we pass through the center. Judgment as to when one is at the center of a via in section is difficult but can be done sometimes by merely looking for when the bottom of the vias begin to flatten. Alternatively, the analyst can cue on voids which may form in the center in the via which are visible in the SEM. This seam is a fill-related problem and is usually seen with tungsten vias. The seam may not occur with other metals with better fill characteristics.

Sequential sections are often specified for a variety of reasons. One common reason is when you are sectioning through a defect found randomly in the middle of the build and you want to know what level the defect was introduced. As you have just learned, a single cross section is only in a two-dimensional slice of space, and may not reveal all relevant information. In one plane, a defect may appear higher up in the build than it really is. It may have been caused by a problem at a lower level from an earlier process step. In order to obtain root cause information, it is imperative to determine where the defect was introduced. Often, only by careful sequential cross sectioning can that level be found. In any case, it is essential for the PFS to comprehend and harness the difference between the three dimensional actuality and the two-dimensional view obtained in cross section. Only then can the PFS correlate where they are in cross-section relative to a top down perspective. This is a necessary skill since the PFS generally work by themselves and must make critical decisions where to stop the cross sectioning procedures and take photographs.

Determining Method of cross-section (cleave or polish)

Let us assume that we want a cross-section of a via chain in the BEOL. The requester will usually specify the location where the via chain cross section is to be obtained. The first question that we must ask ourselves is whether a simple cleave will produce the desired result. All samples cannot be cleaved to give a reasonable result in the SEM. This is due to the differing build materials and the dielectric make up. Most BEOL samples requiring analysis on specific devices must be polished! Depending on material compositions, when a BEOL structure is cleaved, the metal lines will often pull out or fracture in ductile fashion. In such cases, the cleave will result in poor surface quality which is undesirable for SEM

analysis. This problem can however be corrected with a polished cross-section. FEOL structures can be cleaved unless the size of the structure is too small to see in the optical microscope. This is a key limiting factor in the cleaving technique. Structures and devices which are roughly 10um and larger can be cleaved with nearly 100% accuracy because they can be seen in an optical microscope. Structures and devices that are smaller than 10um should be polished to achieve the same accuracy. Of course there is no guarantee that within the 10 microns you will hit the local structure of interest, but a skilled analyst should be able to cleave into a macroscopic area of this size.

Fracture cleaving techniques

A wafer will normally cleave easily in the directions that are parallel to low index fracture planes. Fortunately, most chipmakers make use of this fact and design their wafer layout such that it is relatively easy to cleave chips and wafers both parallel and perpendicular to general structure. This is extremely advantageous, because cleaving in some form is always required to prepare cross-sections. The analyst is constantly performing either a precision cleave through a particular structure, or simply cropping a larger chip or wafer down to a manageable size for mechanical polish.

If the structure of interest is a via chain, for example, containing fairly large size vias, and the via material lends itself to a cleave, cleaving is a quick means of obtaining cross sectional information. This method is often used when the interconnect structures are not yet filled with metal and we simply want a view of the pattern prior to its' metal fill. If one intends on making accurate dimension measurements, one must be certain that he has cleaved through the center of the via. In this particular case (no metal fill), the situation is made easy by virtue of the fact that you can compare a tilted top down view of the via with the plane of fracture. From this vantage, it is easy to determine one’s exact location. One should note that for structures with resist on top, post develop for instance, a cleave is almost the only way to get quick information about the structure. Due to the material properties of resist, mechanical sections of resist are not done. In the future, due to decreasing feature sizes, it is likely that FIB milling will be relied upon to obtain information about shape and size of resist images. Previous work has shown that the FIB mill can be used to prepare critical cuts through this material, but it is far more time consuming.

The cleave is also a good option for obtaining information in structures that have a repeating linearity to them. An example of such structure is a maze, which is essentially an array of parallel lines. This type of structure lends itself to a cleave since the cleave will easily pass through much of the linear structure. In this case, one is able to get an idea of the appearance of the structure with a minimum effort (subject to the limitations of distorting the metal). If the cleaved structure involves ductile metal, the metal will often pull and distort so that details cannot be seen. However, structures made with harder metals such as tungsten can be cleaved to produce a clean flat fracture. When the cleave is done properly, details such as the facets of the W (Tungsten) grains can often be seen. General shape and details of the surrounding dielectric as well as gross problems such as voiding can also be imaged. Note that while a cleave section can result in distortion of metal structures, when utilized as a technique in the FEOL, gates, shallow trenches and other front end structural elements will show up nicely. This is the preferred technique for a quick view, when it is not critical to hit a particular structure, and all of the structures of interest are big enough to easily intercept.

In general, rough cleaving involves scoring the edge of the wafer with a Tungsten carbide (or other suitable material) tipped scribe and than applying a bending moment to either side of the score. This can be accomplished by simply placing the tip of the scribe underneath the score and applying pressure to either side. This will result in a fracture or cleave as the score propagates into a crack across the wafer. It should usually break parallel to your score, but may stray slightly. There are many variables, such as applied pressure and wafer build, that will affect the quality of your cleave. A similar process can continue until the wafer fragment is of the desired manageable size. It should be noted that this technique is normally used solely for downsizing parts, and not for subsequent high resolution SEM imaging. Additionally, rather than perform these tasks on a hard surface, a clean twill jean cloth should be used. This mat provides a soft particle free surface to lessen scratching, and will help absorb applied pressures to alleviate chipping and shattering. Additionally, it will add the necessary friction needed to prevent your chip from sliding around. These twill cloths have a limited lifetime as they begin to accumulate glass fragments and silicon dust that can affect your cleaved surface.

If, however, one wishes to cleave through a particular structure and obtain a high-quality surface finish (for immediate SEM imaging), another technique is useful. Again, we begin by first scoring the top edge of the chip in a location leading up to, but not into, the area to be SEM analyzed, as seen in Figure 3a.

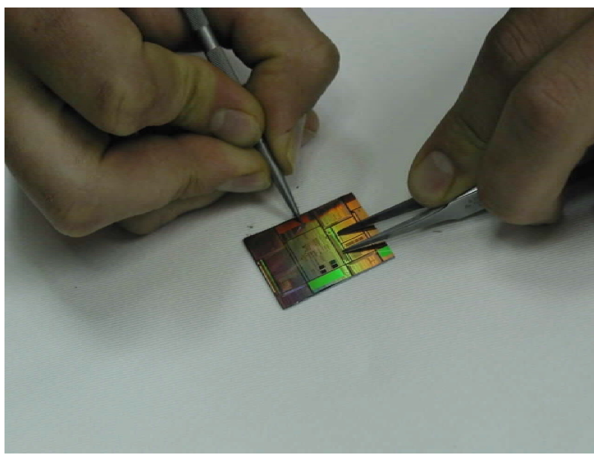


FIGURE.3a: Illustration of scoring the edge of the chip

The score should be in-line with the area of interest so that after propagation it will fracture through this structure. The scoring may be done under a stereoscope or with purchased scoring/scribing equipment (which incorporates a viewing microscope and a scribe). After scoring the topside of the fragment, a corresponding nick should be made on the backside of the chip segment so that area of interest can be located once the chip is placed silicon side up. Remember to use a clean jean cloth to help protect your sample.

The chip should now be oriented such that the scored/nicked side is toward the analyst, with the silicon side (with nick) up. Using ones finger on the edge of the sample furthest from the analyst (no nick/score), gently lift the sample so that it is supported at a 45-degree angle with the scored/nicked side still touching the jean cloth. Now gently place the back of the tweezers against the chip until you can feel the groove created by the nick, as seen in Figure 3b.

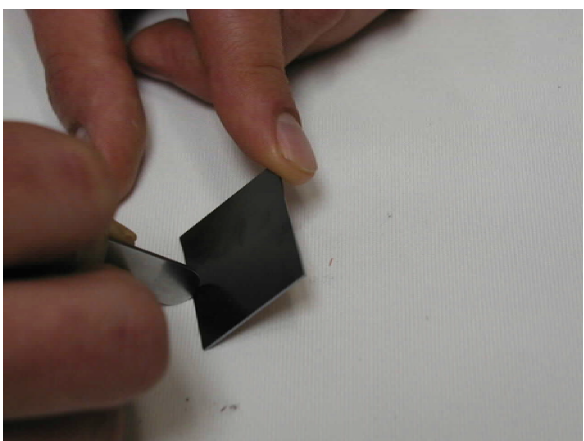


FIGURE 3b: Finishing the cleave from the backside.

Just touching the base of the nick with the back of the tweezers head or scribe is not enough to finish the cleave. A small amount of force must be applied to allow the score to propagate. Moreover, one may need to move their finger to slightly adjust the angle at which the chip is held. Only hands on experience will allow repeated success. This is a quintessential case of “practice makes perfect.” If this technique is performed properly, after the cleave is completed, the side of the chip which was supported by your finger should still be in the same angled position. The cleaved edge on this fragment will have been untouched and should have a mirror finish on the cross-sectioned edge. This is the half of the sample that will be used for SEM analysis. One should now be cautious in handling the chip as to not damage the cleaved edge. The other half of the sample, which may (or may not) have been damaged, as it broke free, should be put aside in the event that future analysis is needed. The sample for the SEM should now be optically inspected to be certain that the cleave is located where it was intended to be and that the edge is free of any unwanted debris.

When cleaving, one generally tries to cleave a sample in half since equal forces on each side of the score will result in an optimum cleave. Unfortunately, this is often impossible to do. The techniques described herein will however work when there is an uneven proportion on either side of the score.

Often, after a successful precision cleave has been made, the chip is still too large to fit into the sample holder. In such an instance, one must further downsize the fragment, while being cautious not to damage the cleaved edge. To do so, take the already cleaved fragment and place it bulk silicon side down on a clean jean cloth. The side that was just cleaved should be facing away from you. Taking a pair of blunt-tipped tweezers, one should place the tips down on the surface of the chip to keep the chip in place. The tweezers arms should be slightly opened such that the real estate between the tweezers tips is the region of interest. In other words, the piece of chip we want will be defined by the space between the tweezers tips. Additionally, the tweezers tips should situated near the edge closest to the analyst (away from cleaved edge). Holding the chip firmly in place, take the scribe and nick the edge of the chip closest to you, just outside one of the tweezers tips. In doing so, the chip will fracture, and a fragment will “fly” away. Only the piece of chip being held down by your tweezers will remain. Repeat this procedure on the other side of the chip your tweezers is straddling. Be certain not to nick the chip inside the boundary created by your tweezers tips. The cleaved edge will remain untouched and the sample should now be ready for direct examination in the SEM.

If one intends to downsize a chip in preparation for a polish section, a chip fragment approximately 4 mm

(long) by 5 mm (wide) is reasonable. A width of 5mm is desirable since it will help maintain proper force on the edge of the sample during polish. The target site should be roughly 100-300 microns from the edge. If the chip is to be kept on the stainless polishing stud and imaged in a standard below-the-pole-piece SEM, the size is not as critical. If however, the subsequent imaging is to be done on a high-resolution in-lens SEM, keep in mind that the sample size is far more restrictive. In such instances, the width (the edge to be polished) of the sample should be less than 5mm and the final length should be roughly 2mm.

Pre cross-section surface preparation

Sometimes, in order to reduce distortion in the cleave or to prevent delamination of a weak interface, the analyst will decide to coat the specimen with a layer of SiO2 (quartz) or TEOS. In this case the structure of interest will be buried under the quartz layer when examined in cross-section the SEM. Examples of a cleave through a line structure without a protective layer of quartz is shown in Figure 4a & 4b. Figure 4a clearly shows the three dimensional nature of the line because when the SEM micrograph is taken at an angle we can see the top as well as the cross section. One should also notice that the edges of the metal lines are not smooth. This appearance is a good example of ductile metal fracture one might see resulting from cleaving of softer metals. If this artifact is undesirable, the distortion can be corrected by polishing, at the expense of turn-around time.

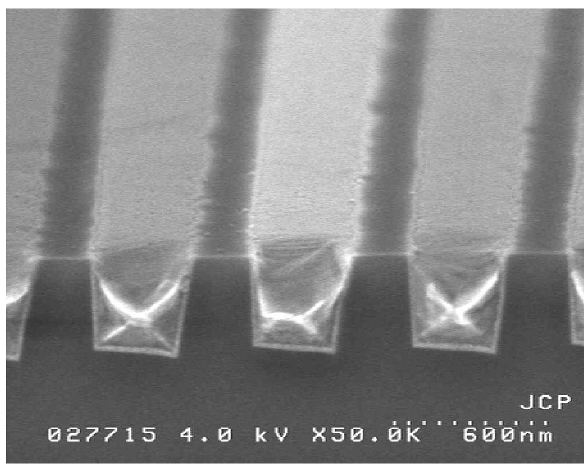


FIGURE 4a. Tilted view of a cleaved section through a single layer of parallel lines. Tilting allows one to see both the top of the lines and the cleaved edge.

The 90 degree view of the same sample shown in Figure 4b is considered the two dimensional view.

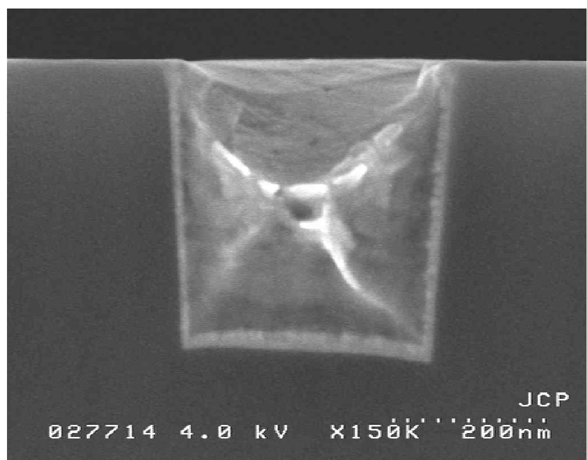


FIGURE. 4b: 90-degree view of same specimen seen in Figure 4A.

The same sample with quartz on top would negate our ability to see the top surface of the line in the SEM, however optically we can still see through the sample so we can monitor the progress of sectioning. We call these views the top down tilted view vs. the 90 deg view respectively. Top down views in the SEM can be taken tilted at shallow angles or at purely right angles to the surface where we can see the structures without any foreshortening or distortions due to such tilt.

In almost all cases, prior to beginning a polished cross-section, we deposit a TEOS layer to protect the uppermost level of our sample. In fact, it is prudent to deposit this layer prior to downsizing the chip. Without this protective layer of TEOS our samples would be damaged severely by the grinding procedures used in mechanical cross sectioning. As an alternative, a glass cover slip epoxied to the surface sample can be used. This method is more time consuming and makes it somewhat more difficult to optically inspect the sample from the top down. Additionally, the added thickness makes subsequent FIB almost impossible. For these reasons, the cover slip is not often used. In cases where the build has progressed far beyond the level we are interested in, we may skip this passivation step. Our lab utilizes a PECVD (Plasma Enahnced Chemical Vapor Deposition) process that deposits TEOS conformally across our specimens. We use 1 to 3 microns of TEOS since we find this to be optimum for ease of viewing and preservation of the structure. As a rule of thumb, the thicker the TEOS the less the chance of damage to the specimen but the more difficult to see the structures in an optical microscope.

Delineation layer for enhanced SEM contrasts

Sometimes we need to deposit a layer of Chromium (or any suitable metal coating), before the TEOS deposition, so that we can clearly delineate the surface of interest from the TEOS used for surface protection. A common example of the need for this technique is when lines and vias require imaging before they are filled with metal. Here they will exist as simply empty pattern in dielectric. A cleave sample of such an empty mold is shown in Figure 5.

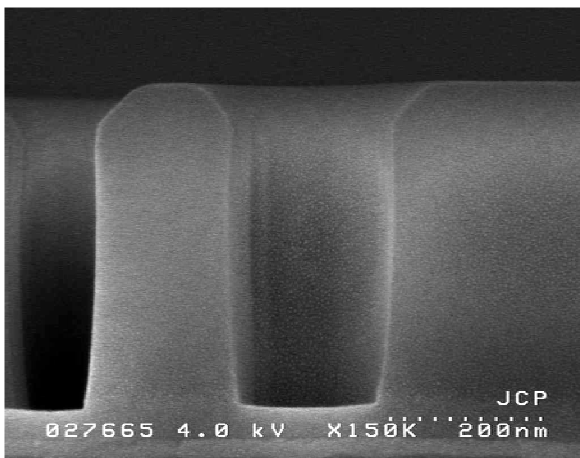


FIGURE 5: Cleave sample of a patterned via, pre-oxide fill

With such a cleave, it is difficult to hit the via center and often difficult to get a nice view of your structure. As a result, a polished cross section may be required. In this case the PFS will coat first with about 300 angstroms of Cr, followed by 1 to 3 microns of TEOS. A section with Cr delineation is shown in Figure 6 (The void seen in this picture is in the protective TEOS, and results from "pinch-off" during deposition).

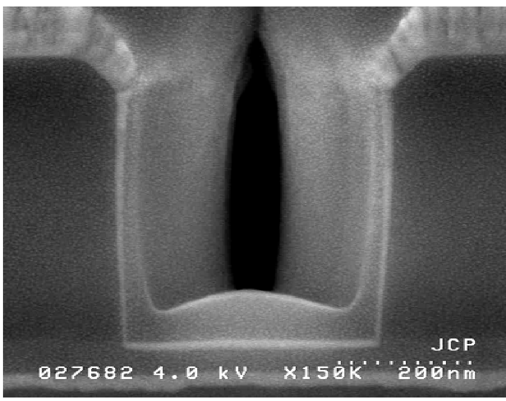


FIGURE 6: Cleave section of patterned line with the addition of a thin Cr delineation layer.

Without the Cr delineation layer it would be nearly impossible to delineate between the BEOL process oxide and the TEOS protection layer (assuming no chemical assistance). Additionally, because Cr sticks to a number of surfaces, it is also beneficial as an intermediate adhesion layer, and will help keep the protective barrier in place. It should be noted that Cr deposition should not be more than about 300 angstroms in thickness to help ensure its optical transparency. It is essential for this layer to remain optically transparent so optical inspection of the structure can still be done during mechanical cross sectioning.

Cleaning / Oxygen Ashing

If top-down SEM analysis has been performed on a specimen requiring a polish cross section (pre-TEOS deposition), it is important that you oxygen ash in a plasma environment to remove any organics that may have been deposited. If you do not clean the top surface of the sample, the quartz or TEOS may not properly adhere, causing the TEOS to delaminate during polishing. This will effectively destroy your cross section. There are many makes and models of suitable ash tools; however, one must be cautious that other residual gasses are not present in the chamber. If, for example, the tool doubles as a RIE (Reactive Ion Etch) tool and often sees fluorine chemistries, there may be unwanted damage to the exposed chip surface.

Laser or FIB Marking

In some instances, it is desirable to mark the area you are interested in sectioning. This is a good practice when the structure is too small and difficult to see optically through a microscope. These marks will denote the area of interest and give a good landmark for optically locating the fail site during the polish. Two common methods to achieve this goal are with a laser or a FIB It is important to consider the size of the mark, taking care to leave as small a mark as possible. The larger the mark the easier it is to see but the more chance of delaminating during the polish. The laser mark is perhaps more common due to ease of use and time considerations. It is more than acceptable for everyday samples and non-critical applications where the goal is to put landmarks in general areas. The laser mark however, can be quite cumbersome as it is difficult to completely control its' size and location. We find the FIB provides more accurate marking, particularly for small isolated contacts. Moreover, because the FIB can make small marks in an extremely controllable fashion, it less often results in subsequent problems during polish. Additionally, with the FIB we can mark sub micron areas that cannot be optically seen under a

microscope, but are useful during subsequent SEM imaging.

Mounting prior to Polishing

Let us assume we have successfully cleaved the sample to the proper size and we need to prepare a polished section. Figure 7a and 7b show alternate views of the polishing block with the specimen mounted to the removable stud. Note that these pictures are for illustrative purposes only. One should never rest the sample as seen in these pictures lest the sample will be damaged.

FIGURE 7a: Overall view of polishing block with illustrative sample mounted. The side screws hold the removable stud in. The adjustable Teflon heel is seen on the other end.

FIGURE 7b: Alternate view of polishing block with illustrative sample mounted.

The polishing holder is a machined block of stainless steel 2 inches by 2.5 inches with an adjustable Teflon heel at one end. At the other end is a removable sample holder (stud) on which we have placed the specimen to be polished. The removable holder will fit into any below- the-pole-piece SEM. This system allows you to remove your sample and place it in an SEM for examination, where you can either check the progress of your section or take the required micrographs of the final plane. For higher resolution images at higher magnifications (approximately 200k and above), one may opt to remove the specimen and place it in a smaller in-lens specimen holder. This holder is necessary when using in-the-lens systems such as the Hitachi 5200 and 5000 series SEM's. Additionally, there are special sample holders and polishing blocks designed to accommodate smaller, in-the-lens-samples

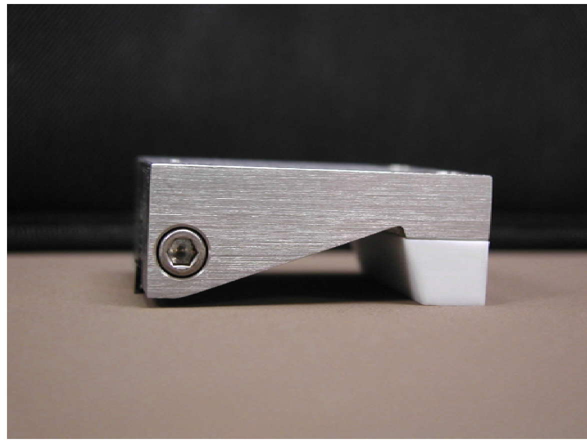


FIGURE 8: Side-view of polishing block showing adjustable Teflon heel. Specimen holder screw is also shown in this figure

The Teflon heel allows for low friction and minimal material exchange during polish. The heel also enables fine adjustments to be made to the polishing angle. This control is necessary when one is attempting to polish into a long metal line, or row of contacts. The polish angle is controlled using counter sunk adjustment screws located on the top of the block. It is important to remember to reset the heel parallel to the polishing block prior to its use. This will minimize any potential difficulty in straightening the sample later. This can be accomplished by simply tightening the heel screws snug to the block (A worn Teflon heel should be replaced to not complicate the straightening process).

The specimen stud is heated on a hot plate and a small dab of wax is applied to the stud for adhesion of the sample. We use Apiezon wax for this purpose, but there are other suitable adhesives. The key is to have an adhesive that can melt or become viscous at relatively low temperatures, can be reset if necessary, and easily removed with solvents. The cleaved piece of chip is to be attached with the silicon side to the polishing stud. This will allow the back end interconnect to be visible (through the quartz). This is necessary to monitor progress during polish. The specimen is set on top of the melted wax, pressed

down, and straightened with the appropriate interconnect lines running as parallel to the edge of the block as possible. This will help reduce future straightening problems. Be certain that you have mounted your sample using adequate pressure to remove most of the wax between the sample and the stud. This will help reduce any chamfering of the edge during polish, and will reduce the possibility of the sample drifting while the adhesive cools. Moreover, one needs to be certain the target is positioned such that it is hanging off the edge of the stud. In other words, when polishing, one needs to assure themselves that they will encounter the target before running out of real estate, and hitting the stainless block. These operations are made easy with the help of a stereo zoom microscope. The lines on the chip are good to use as reference. One can focus either on setting lines perpendicular or parallel to the stud's edge. During polish, these lines should also be used to help you judge when you are parallel, and how far you are from the intended target.

Next the stud and mounted specimen are left to cool so the adhesive sets. Running under cool water will cause this to occur faster. Be cautious as not to have too turbid a flow, for fear that the water may shift your carefully placed chip.

Any superfluous adhesive should be removed from the surface of the chip, particularly around the area to be polished. The adhesive can partially obscure ones view of the chip surface, potentially affects the polish, and may have adverse charging effects during subsequent SEM inspection. Any excess material can be swabbed away using a Q-tip and acetone (or other solvent).

Optical Inspection and Specimen progress

During the polish it will be necessary to optically view your sample to monitor your progress toward the structure of interest. This will allow for precise straightening of the polish angle as well as judgment of when to change lapping film grits. To do this you will place the sample and polishing block onto an inverted microscope with the TEOS'd surface down and exposed to your lens. The inverted scope is outfitted with a circular support ring to rest the outside of the holder so that you can freely view the sample. Using the inverted scope you may judge how parallel to the edge you are as well as how far you are from the intended target. One can also rotate the polishing block such that the Teflon heel is on the stage and the polished surface faces the lens. In this view, one can make judgments as to whether the polished surface is scratch- free and clean. Unless you are at your target, a cross sectional view will not allow you to see the structure of interest because it is buried within the chip fragment. The structure will not become visible until you are within a few microns. Top-down optical imaging of the structure is possible since you are looking through transparent quartz/TEOS. Judgment of the grind by paper grit and wheel speed is very critical for an accurate cross section. This skill will take time to develop. Somewhat later on in this paper, several sequential optical views of the polished surfaces will be shown at various stages in the polish sequence to illustrate typical polishing progress and status of the edges of interest.

Rough Grind

Diamond impregnated lapping films from Allied High Tech Products, Inc are typically used for rough grind steps (however, there are several other manufactures that make equivalent products). Additionally, more inexpensive papers such as SiC may be used, but they tend to not last as long. We have found that the diamond papers are best for our high output applications. Additionally, they are excellent for evenly cutting systems with hard and soft metals coexisting. The lapping films come in many grits, and we typically stock several from 30um grit down to .1um grit (30um, 15um, 9um, 6um, 3um, 1um, 0.5um, 0.1um). The important thing to remember is that each subsequent paper must remove the scratches induced by the previous coarser grit, and must do so while leaving enough real estate to allow final polishing steps. Each individual has their own personalized system for which films they choose to use. Only hands on practice will acquaint the novice with the material removal rate for each grit, and the dynamics of this rate as a film is used and worn. This knowledge and experience will empower the analyst to make educated decisions as to which paper should be used and personalize their methodology. We recommend that the beginner use only small intervals between papers as they gradually approach the target. In general, switching between films depends on the distance from the target to the polished edge. Analysts that have mastered the art of sample preparation often use a smaller number of grits as they have become comfortable with getting close to the area of interest on rough grits and making last minute adjustments on fine grits. The following example is typical of the experienced technician who may use a 30um film, followed by 6um film, and final adjustments with 0.5 grit.

Our lab uses Buehler polishing tables with 3 standard 8inch wheels per table. The brass wheels are then outfitted with quartz polishing plates. It should be noted that any setup with a rotating wheel, flowing water, and drainage is acceptable. The quartz plate should be cleaned of any material and foreign debris so that the surface is completely smooth and flat.

The lapping film is then placed on a wet quartz wheel. Pulling a squeegee across the film will cause the film to adhere to the quartz wheel. Adhesion of this paper is supplied by the natural adhesive forces of water on the backside of the paper. Each lapping film can be used many times before they are ineffective, and should be stored between uses in drying books. The papers are color coded for identification, and should be stored separately as to not contaminate a finer grit with coarser particles.

During polish, a steady somewhat low flow of DI water is aimed at the center of the wheel so that it will flow outward across the spinning wheel. The water acts to reduce friction between the sample and film as well as to remove the material that is being polished away. Particles of unwanted material can be dragged across the face of your polished surface to create damage and deep scratches. The films can be periodically cleaned off using clean wipes or Q-tips to remove any embedded material. This cleaning will enhance the performance of the film.

The wheel should be spinning counter clock wise, ranging between 300<>600 rpm (lapping film only). In general, the faster the speed, the quicker is the material removal.

Generally the weight of the block is the only applied pressure that is required. The polishing block should be held in a manner as to not add any significant down force. When you put the block down on the polishing film, put it down Teflon side first and than gently lower the samples' edge to the pad. This will prevent the sample from snapping off the stainless stud.

The very first step is to remove any sharp edges from the chip fragment. Otherwise, this edge can grab the abrasive and tear the specimen off the holder or tear the lapping film. If one were to hold the block as indicated by the initial location in Figure 9, these edges will be effectively removed in just a few wheel rotations.

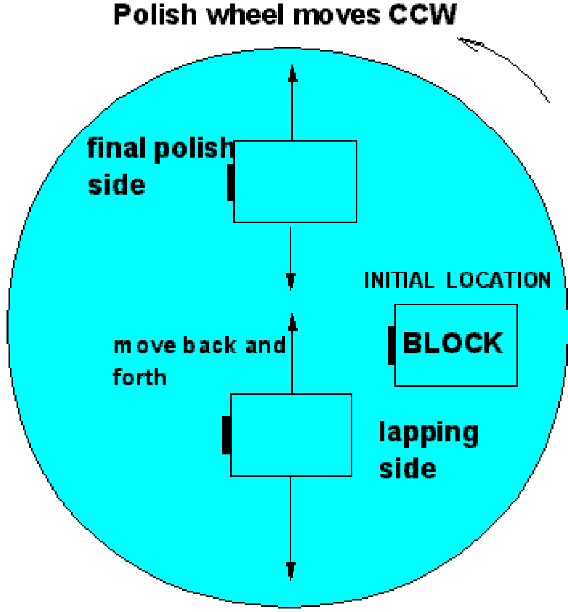


FIGURE 9: Schematic illustration of polishing wheel and location of block for various polishing operations

During rough grinding steps, the polishing block should be held such that the grinding media travels from the protected surface towards the silicon and stainless steel stud. In this manner, the motion of the wheel will be continuously pushing the sample into the stainless block. For example, the block may be held on the portion of the wheel closest to the operator, with the Teflon heel to the right and the sample to the left (assuming counter clock wise rotation). This position is illustrated in Figure 9. While holding the polishing block (without exerting any pressure), it is good practice to move the block back and forth across the grain of the spinning wheel (from edge to center). This will prevent ruts and uneven wear of the paper, which helps extend the films lifetime. Additionally, this technique will also spread the removed material across the paper, resulting in less buildup, and this less chance of unwanted surface damage. Be certain to not cross over the center of the film, where the wheel will be spinning in a different direction relative to the edge of your sample. On this side of the rotating wheel, one encounters the possibility of the rough grit shearing the chip off the stainless stud.

The lapping film is used for the rough grinding only, and will chip and damage the edge of the sample, particularly during usage of the coarser grits. In these early stages, optical interpretation of the structure in cross section will be impossible with the degree of damage to the section. Normally the cross-section is not observed during the rough grind, and will only be viewed when using the finest film and final polish steps. The less abrasive steps allows

precise optical interpretation of the cross sectioned edge.

Depending on how far your cleaved edge is away from the target site will determine which grit of paper you should start with. When we are over 400 um from the target site we use a 30-micron grit. For farther distances, 45um paper may be used. The 30-micron grit will be used to get within 200 microns from the target and for the initial straightening of the polished surface. The sample should be polished only in few second intervals, and checked repeatedly in the inverted optical microscope. In the early stages, this is done to check progress, but mainly to see the degree to which we are parallel to the structural elements. One may judge using any structural elements close to the polish edge for reference. If you look top down and notice your polish is not parallel, simply loosen the Teflon heels' adjustment screw by a small increment (<1/8 turn) on the side that is being polished more aggressively. This will lift the heel on this side, removing this side of the sample off the polishing film. This effectively increases pressure on the other side that has more bulk silicon causing it to polish faster. Normally, this takes 4 or 5 adjustments before the specimens polish surface is parallel to the structural elements.

During polish, one must be continuously aware that polishing stresses can cause weak interfacial layers in the sample, or the deposited TEOS layer to delaminate. Another common problem is local chipping at the polished edge. The earlier these potential problems are diagnosed, the more likely that adjustments and accommodations can be made. Oftentimes, simply changing to a new polishing film, or different grit all together may help. In more severe cases, another TEOS deposition may be required.

When using a 30-micron paper, it is expected that the edge will be visibly scalloped and rough. In this case, it is simply a function of the rough abrasive. Inspection at this point can be done at small, 5x objective magnification. Remember to frequently check the status of your sample, say every 2 or 3 swipes back and forth on the wheel, particularly when using a new film. Adjustments should be made often until we are at the distance needed to step down to the next smaller grit. In this case, we will drop down to a 6-micron paper.

When the sample edge is within about 200 microns of the target one can opt to go straight to the 6-micron film although as stated earlier this somewhat depends on the skill of the PFS. Before placing the next film, one should again be sure the quartz disk is free of debris. Moreover, it is imperative that the sample be rinsed and swabbed, also to remove any loose debris, or larger grit from the previous polish. It is important not to transfer any unwanted material onto the next polishing film. The 6-micron film will be used to bring the polished surface to within 50 um from the target area. Again, regular status inspections are advised.

It should be noted that with each decrement in film size you must plum, or straighten, the surface of the cross section by adjusting the Teflon heel on the back of the block. With small incremental turns one can increase or decrease the pressure to the side of the sectioned edge that requires adjustment. This will gradually make it parallel to the structures of interest using the structures seen top down as a guide to judge the adjustments that are to be made.

After making the initial adjustments on 30 microns begin polishing on the finer grit film. After 5 or 6 back and forth swipes on the 6-micron film, stop and check the top surface again and adjust so that edge is parallel with structures. In the optical microscope, the edge should now be much smoother and more or less parallel to the structure of interest. Continue to polish until the target is within about 50um of the edge. In this example, we will now change to the 0.5-micron lapping paper. As always, make sure the quartz wheel and sample are free of any debris. You should now view the top edge with 50 X magnification and higher while making fine adjustments to the angle using the Teflon heel. Continue to polish, examine, and make any necessary adjustments to your sample. When attempting to hit an individual contact, be certain that your approach does not intercept the contact of interest. A good distance to stop is roughly 2 microns from the target. Figure 10 is an optical top down view illustrating a safe distance to stop outside the target. In this image, the target is the vertical row of links on the far left of the chip surface. Notice how the polishing sequence was stopped immediately before the target, leaving a perfectly smooth edge.

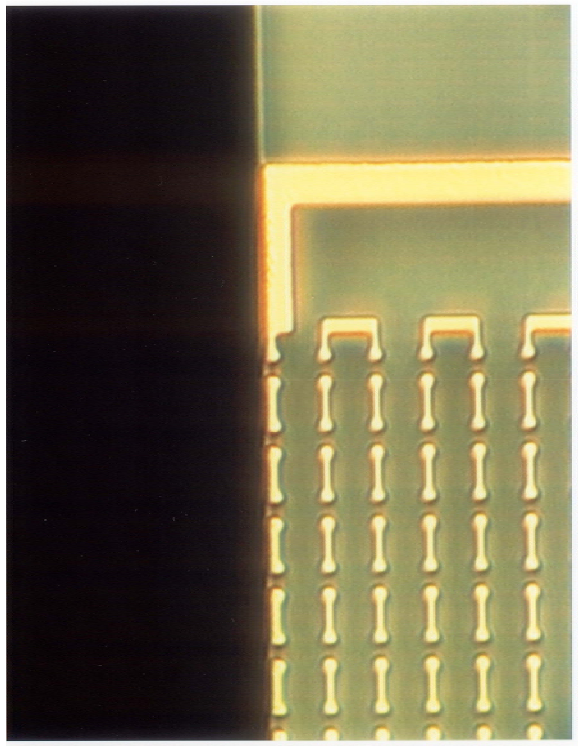


FIGURE 10: Optical view of good stopping position for completion of rough grind step (assuming that the vertical row on the left of the image is our target)

If the structure of interest is a large chain containing multiple vias and no specific contact of interest you can stop inside the structure and be assured that there is still plenty of room for the final polish. By repositioning the block on the optical microscope to view the cross sectioned edge, one can see that at this point, the surface is scratched and damaged. Figure 11 illustrates the damage that is typically seen at this stage. Notice how the bulk silicon (right half of image) is noticeably striated with scratches running right into the back-end features.

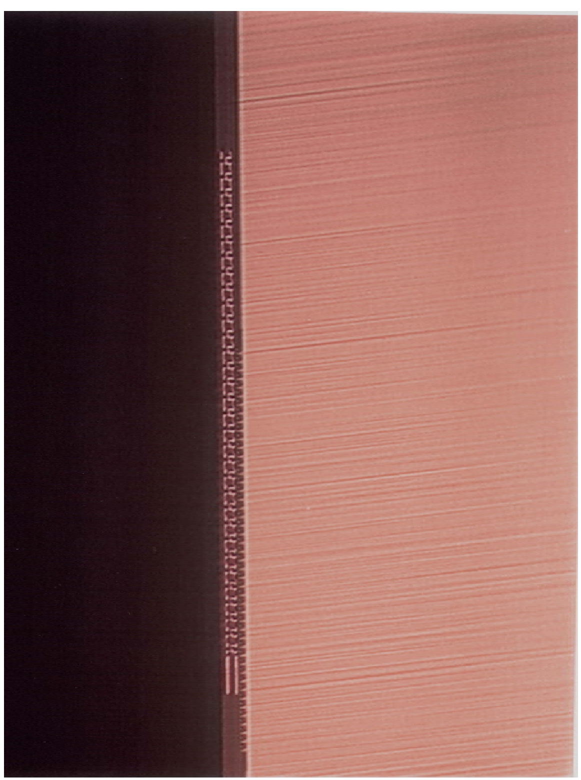


FIGURE 11. Optical view of cross-sectioned edge after rough grind, showing the expected level of scratching

The 0.5-micron film is the last lapping paper that will be used during this section. From this point on, we will rely on a diamond slurry polish to take us into the target, as well as remove all the scratches and remaining damage.

Final Polish

The final polishing steps will be performed using a “Final-A” Polishing cloth with .25um and .05um water based polycrystalline diamond suspension slurry. Allied High Tech Products, Inc., sells the products we use but there are several other suitable final polishing slurries on the market. Along with these slurry's we also use a 50:50 mixture of DI H2O and Glycerin. Polishing wheels should be dedicated to the final polishing pads. The pads typically have adhesive on the back and should not be removed until they are worn and require changing. When replacing a pad, start at one edge of the wheel and slowly roll it over the wheel while removing the adhesive cover slip. The goal is to apply the pad without air bubbles. Additionally, each slurry size should have its own dedicated final-A pad because slurry remains within the pad even after rinsing. Intermixing grit sizes on the same pad will result in increased polish rates and unwanted scratches.

During final polish, the polishing block needs to be held with a different orientation with respect to the

rotating wheel (as compared to rough grinding steps) as shown in Figure 9. We hold the block on the other side of the wheel such that the wheel and slurry will be pushed under the specimen, hitting the bulk silicon first, and then traveling through the area of interest. For example, on a wheel-spinning counter clockwise, hold the block on the far side of the wheel, with the sample to the left and the Teflon heel to the right.

Lets assume the sample has been polished (rough grind on diamond lapping film) to within 2um of the contact or into the chain of interest. Make sure you have rinsed the final-A pad free of any debris. Always check the surface area for disturbances and replace as necessary. Rub your fingers over it to assure that there are no bubbles or impregnated grit. Before using always check it with a slow stream of water to see it is free of bumps.

Begin by applying the 0.25um diamond suspension slurry and glycerin. With the wheel rotating at about 100 rpm, liberally apply about 4 squirts of slurry essentially covering the wheel followed by 4 squirts of glycerin or 50/50 mix of glycerin and water kept ready mixed in squirt bottles. Do not rub fingers across at this point as it is assumed that the wheel is clean from your pre check. During final polish, the wheel must be kept wet at all times. In order to do so, regularly apply both fluids in equal amounts. You can never add too much, however, the diamond slurry is expensive. The closer you came to the contact in the rough grind, the less time spent with these diamond slurries. Additionally, too much final polish will begin to round the edges off your sample. Therefore, it is in ones best interest to get as close as possible to the target using the finer diamond films.

Initially you are going to polish with 0.25-micron polycrystalline diamond silica mixture and glycerin. The glycerin gives some lift to the block during the polishing. This is very important when sectioning some of the softer materials, and will help prevent polishing damage. When dealing with organic dielectric, straight glycerin (as opposed to 50:50) is used to maximize lift and thus minimize damage to the specimen.

Place the specimen down on the rotating wheel with the Teflon heel making first contact and slowly lowering the polished edge onto the pad. Rotation should be between 75-100 rpm. Move the specimen back and forth on the block as indicated on Figure 9 making sure the only pressure exerted is from the weight of the block on the polished edge. Continuously apply your polishing solution adding glycerin per the discussion above as you move back and forth. After about 5-10 swipes you are ready to look at the polished surface at 10-50 x objective magnifications when looking at the edge in cross section, one should key in on the surface scratches. It is impossible to completely remove all scratches from all parts of the sectioned surface. The critical spot to focus on is the interface between bulk silicon and where the active layers of structure begin. We have completed this polishing step and are ready to move on once the scratches from the 0.5 film have been confined to the bulk silicon only. As mentioned earlier, the goal of each subsequent step is to remove the scratches induced by the previous step. Here, the small exception is that scratches in the bulk silicon are permitted, as long as they only lead up to the active structure layers and do not extend any farther. Once this occurs, the layers to be inspected in the SEM should no longer contains any 0.25-micron scratches. An optical photo depicting what to look for is shown in Figure 12. This optical image has been brightened to better illustrate what the scratches should look like. Notice how the scratches in the bulk silicon (right side of image) no longer extend all the way up to the back end features.

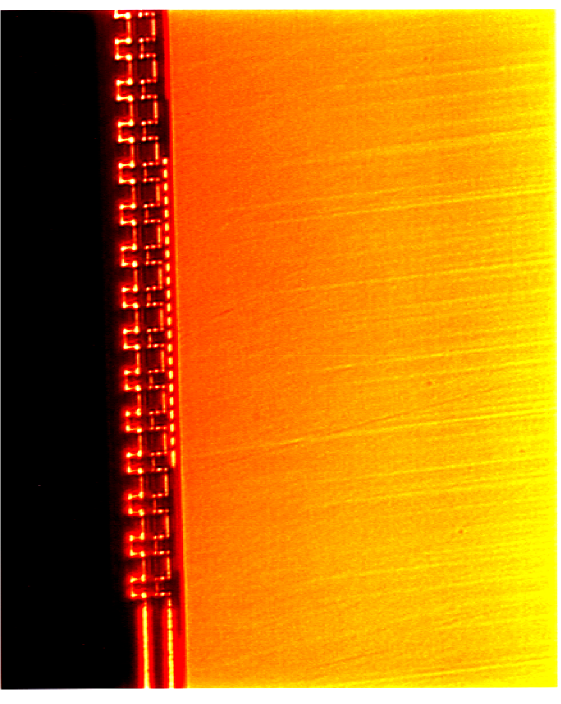


FIGURE 12: First look after completion of 0.25 diamond polish. Edge scratches lead up to but do not extend to the edge of interest

A word about polish rates with diamond slurry: These are somewhat variable depending on materials being polished and other variables such as rotation speed and pad integrity. As a rough guide however, 10 swipes using either the .25micron diamond or the .05-micron can take you through a ¼ micron contact, so one should proceed with caution. Experimentation and experience is the best way to get a feel for polish rates. Initially, go about 5 swipes between optical examinations. Of course if you are

just sectioning through something that does not have a critical contact to hit you can afford to be more aggressive.

After rinsing the sample prior to optical inspection, don't bother to dry the whole block; rather, only dry the specimen by using a nitrogen (N2) blowgun directed at the polished edge. Once this region is dry one may examine in the inverted microscope. It could take you several minutes to reach the edge of the target. Never adjust the Teflon heel once final polishing has begun. Your aim is to use the 0.25 diamond to reach a point just prior to touching the contact. This is difficult to judge optically. For your first few attempts you may want to look at the specimen optically and then compare it to how in looks in an SEM such as Hitachi 4500 that accepts the stud without dismounting the specimen. This way you can ascertain your exact location relative to the target. As long as you do not dismount the specimen from the polishing stud, it is possible to go back and fourth between the polishing pad and the SEM without any problems.

Lastly, the 0.05-micron diamond slurry solution will be utilized for a final clean up of the sample surface. Here is where rounding can be particularly problematic if you are to long on the sample. In general, the less TEOS used, the more rounding will result. Remember that just 10 back and forth swipes can potentially take you through a 0.25-micron contact. If you are just touching a contact and the contact is small, give it 3-5 swipes and look under the microscope to see where you are. In an optical cross sectional view, you should see the area of interest as shown in Figure 13.

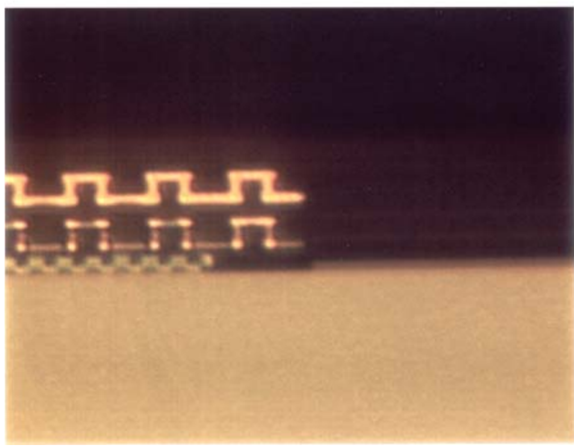


FIGURE 13: Optical image depicting acceptable quality after the final polishing step.

There should be no scratches or delamination in the area of interest. If the specimen is for example a via chain, the situation is less critical because your section will invariably be intersecting many vias at various stages. In the SEM you can always find a via where you are near the center. Likewise if it is a linear array of lines and one doesn't care exactly where the cross section is, you can be somewhat liberal. On the other hand if you are doing a sequential section of a single via, care is needed and only 2-3 swipes should be used in-between SEM sessions. Many sequential cross sections can be made through a contact. This gives the customer a full overview of the contact and it's failure mechanism.

The highest quality images are obtained on SEM's with in lens capability such as the Hitachi 5000 or 5200. To use this SEM type, one must carefully remove the specimen from the stud in order to put the chip into a special sample hold. To do this, first remove the stud from the block. Heat the stud on a hot plate and slide one edge of the specimen towards the front of the stud so it hangs over the edge. Now you can grab it by one edge and slide it off of the stud. Bad technique here can end up with wax or foreign matter getting onto the edge of interest. Alternatively, there are stainless studs that have been specially designed to accommodate small shims that are accepted by the in-lens sample holder. Small chip fragments can be affixed to this shim using an adhesive as described earlier. The advantage of this is that the chip need not be heated and removed prior to imaging. One need only loosen the setscrews that were holding the shim in place. Note that once you have removed a specimen, you can put it back on the holder for further polishing. Simply place the shim back into the holder and tighten the setscrews Critical alignment is not necessary at this point.

Final Surface Preparation

Argon ion milling

The surface quality of the cross-section is critical when attempting to achieve high-resolution SEM images. Little or no surface topography (i.e. scratches, debris, unwanted organics) is desired, especially when dealing with high materials contrast, low KeV, surface sensitive imaging. In order to achieve such a finish, the final preparation for many samples is an Argon Ion Mill. The tool we are using is a Gatan Duo Mill. This tool is a workhorse in sample preparation due to its flexibility. It accepts a wide range of sample sizes and is capable of multiple beam angles and beam conditions. Additionally it can be set up to have gas assisted milling. The theory behind ion milling is to introduce a low energy beam of ions (normally Argon) to the surface of a sectioned sample to achieve a final smooth surface that would be otherwise unattainable. The ion mill will eliminate even the smallest of scratches. There is no polishing slurry that can compare to the final surface quality we get from ion milling.

After dismounting the sample from the polish stud, the sample is placed in a holder that will allow the sample's cross-sectioned edge to face towards the beam. The beam's angle of incidence is adjustable, and we have found that an angle between 10 and 15 degrees to the sample's sectioned surface (75° to 80° from normal) is optimal. Additionally, the gun conditions should be set to 4KV @ 0.5 mA. The specimen should be rotating (2-10rpm) as not to induce a directional mill. Normal times for ion milling average around 20 sec to 2 minutes. The optimum mill time for a particular system must be individually determined. Longer mill times are not always favorable due to differences in mill rates from material to material. For example, if you are sectioning into a tungsten via surrounded by SiO2, you must be careful not to mill too long. Oxide mills at a much faster rate than tungsten; therefore, excessive times will result in rounding of the tungsten via after the oxide is selectively pushed back. Accurate SEM analysis will be made difficult by the resulting "3D effect" created by this rounding. Figure 14 shows a polished contact after 30 sec ion milling.

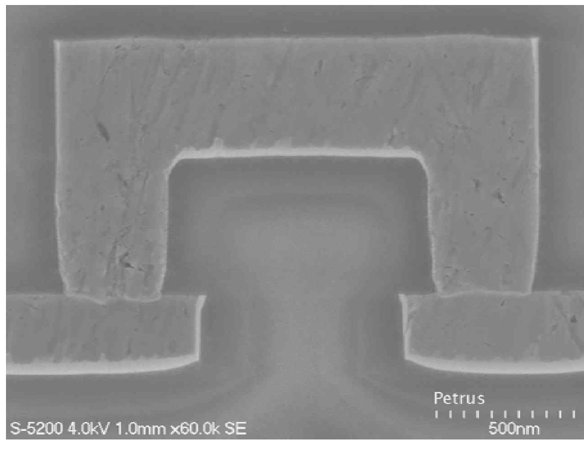


FIGURE 14: SEM micrograph of the contact after 30 sec of ion milling. Often you will need no further decoration

Sample Decoration Techniques (Plasma & wet chemistry)

In our lab, the most common decoration used to achieve materials contrast in the SEM is DE-100. This is a concentration of 10% Oxygen in a balance of CF4 gas. This process is performed in a small tabletop RIE tool that is dedicated to sample decoration. The current tool we are using is a Technics Micro RIE with a PC controller. This tool is capable of handling several processes with numerous gas combinations for the purpose of materials delineation. De-100 will decorate Silicon Nitride and any form of Silicon. It will not etch oxide with the short exposure times we typically use. Normally an etch time of 18 sec is adequate to enhance the materials for SEM inspection. Figure 15 shows a FEOL sample post ion mill and 12 sec DE100. In this image, the nitride has been etched and appears dark (in contrast to the bright oxide) because it has been selectively etched.

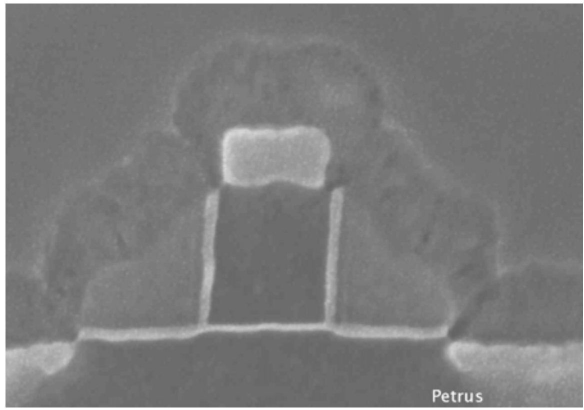


FIGURE 15: SEM of gate region after ion mill and 12 seconds of DE100 RIE etch.

Many samples will require an SiO2 decoration. For these samples we commonly use a 5:1:1.6 glycerated, buffered hydrofluoric (5:ammonium fluoride/1:hydrofluoric/1.6:glycerine) or a 7:1 BHF (not glycerated) as well as other variations of the same chemistry (the 5:1 has the advantage of passivating metal). This decoration will not etch Silicon and provides a good oxide decoration. Normal etch times range from 3 to 15 sec depending on the dielectric material and the concentration of the etch. Figure 16a shows a FEOL structure after a BHF decoration. This image can be compared to the image with the DE 100 etch

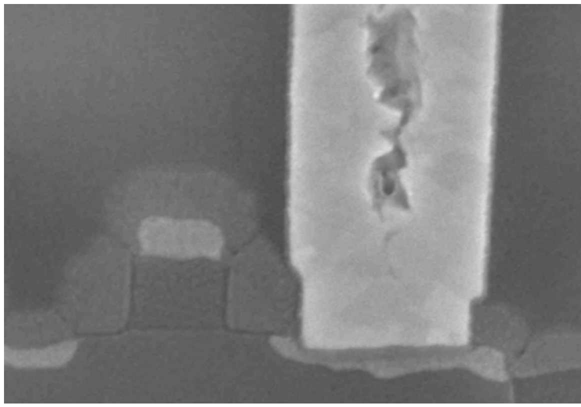


FIGURE 16a: SEM of gate region after 15 seconds of 10:1 BHF decoration. Notice that there is no distinct delineation of oxide from nitride compared to the DE 100 etch

Figure 16b is a BEOL structure where we see that the 5:1 BHF has decorated the nitride barrier that did not show up in the sample that was only ion milled (Figure 14).

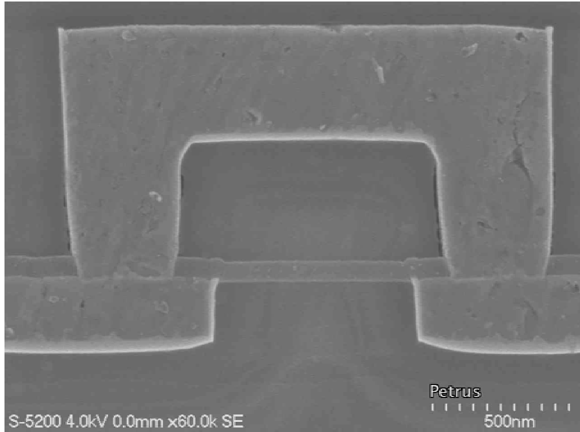


FIGURE 16b: SEM of a via chain link that has been decorated using a 5:1 glycerated. The nitride oxide delineation is clear since the oxide was selectively attacked and recessed.

Additional Remarks

The processes outlined above attempted to cover every detail, from start to finish, that one might encounter during the mechanical polishing of a specimen. It is expected that even with this knowledge, it will still take the novice some trial and error to become proficient. It should take anywhere from 15 minutes to several hours to prepare a specimen prior to the time spent in the SEM. SEM mastery is an entirely different beast, where you can make or break all the effort spent in sample preparation. Here in our lab, micrographs of the final section are always captured digitally, stored, and then transmitted to the requester. Additionally hard copies are provided. Invariably, some image quality is lost between what you see on the SEM CRT and your printed digital image. This issue can however be alleviated by using Polaroid film (which gets very expensive). Alternatively, the requester will sit with the PFS while examination is being done in the SEM. This way he or she gets an uncensored view of the build and any anomalies within.

Focused Ion Beam Cross Sectioning as a Complement or an Alternative to the Conventional Mechanical Technique

Historically, manual mechanical cross sectioning techniques have been the most widely used methods of obtaining alternate perspectives of integrated circuits, structures, and defects associated with semiconductor technologies. To date, the evolutionary development in manual mechanical cross sectioning has managed to keep pace with the continually shrinking geometries associated with integrated circuitry in semiconductor technology. As discussed, in order to produce a high quality SEM image using mechanical techniques, one must employ a number of rough grinding and fine polishing steps, which can often be a lengthy and tedious process. Moreover, in the case where one is interested in securing an image of a specific defect or isolated structure, this procedure can be extremely difficult, depending on the size of said defect or structure, It is imperative that the analyst “time” his/her arrival at the point of interest so that it coincides with the final stages of the cross section. As the industry advances towards yet smaller feature sizes and new low-k dielectrics, mechanical sectioning techniques may need to be supplemented, or even replaced, by faster, more accurate sectioning procedures. Additionally, speedy turn-around-time is of the utmost importance. One can therefore understand the need to drive towards processes that can churn out rapid results with high precision, while preserving the expected high quality. The Focused Ion Beam is powerful failure analysis tool, which, in regard to the specific application of cross sectioning, has quickly become a regularly used instrument, and even a necessity in particular situations.

Complimentary FIB

Below (Figure 1) is an SEM image depicting a typical electromigration failure in a copper/low-k system; a void has formed at the depletion end of this metal stripe near a flux divergence point.

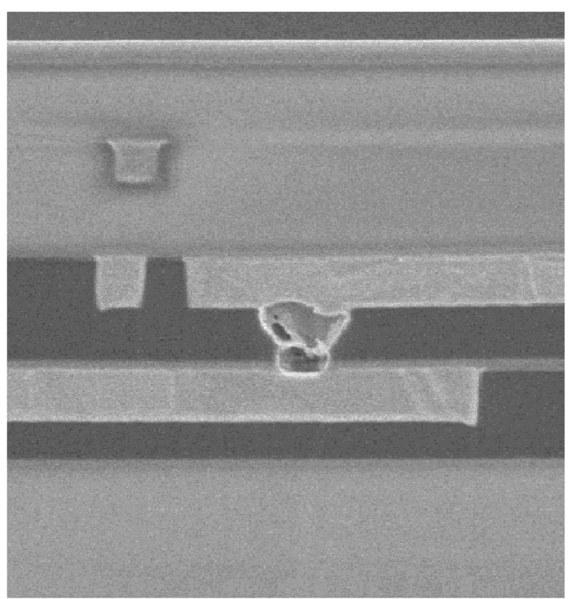


Figure 1: Typical void formation in an electromigration test structure

This is an excellent example of where an analyst greatly benefits from utilizing the FIB to complete a manual, mechanical cross section. Although a

mechanical cross section could have been implemented to image this region, FIB allows us to do so without any adverse effects to the void itself. During a mechanical cross section, it is probable that polishing compounds, grinding media, or loose debris will lodge itself in the void. It becomes extremely difficult to remove these materials once they have lodged themselves within the exposed opening. Additionally, even if the analyst was able to remove most of the contamination, residue is likely to be left behind. This presence of any residual contamination will most definitely impact any subsequent materials analysis.

Figure 1 also helps illustrate the high precision and accuracy which can be achieved with FIB cross sections. Once a close manual mechanical cross sectional approach is completed, leaving the analyst within 2-5um of the desired target, the FIB allows the analyst to incrementally remove very small amounts of material in an extremely controlled fashion. Figure 2 depicts a low magnification FIB image where a manual mechanical cross-section approach was completed using the FIB.

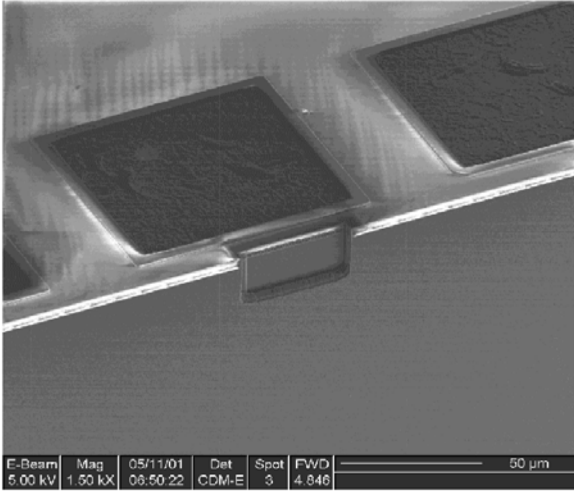


Figure 2: FIB box completing a manual mechanical cross-section

Such control will result in a repeatable method of sectioning small targets, and the capability to stop squarely where one desires. In comparison, even a master technician with cross sectioning expertise can have difficulty manually sectioning into small targets. In today's technology, with ever decreasing feature sizes, it is becoming nearly impossible to actually see what we are attempting to section; we are pushing the envelope with the resolution of today's optical microscopes.

The advantage of being able to incrementally remove very small amounts of material becomes heightened with new dual beam FIB technology (ion and electron columns), which allows for real time SEM updates. Whereas in older, single beam FIB technology, the analyst would have to go through a tedious process of stopping the mill, rotating the sample so that the cross sectional plane is oriented toward the beam, grabbing an image, and rotating back to the original position to continue milling, new dual beam technology allows for continuous cutting while imaging on the fly, There are, however, tricks, that the skilled FIB technician may use to help circumvent this issue with single column FIB's, which will be discussed in a later section.

Another powerful reason for employing the FIB as a complement to a manual mechanical cross section is the arrival of low-k materials on the dielectric scene. In highly general terms, low-k materials can be separated into two groups: spin on films and CVD (chemical vapor deposition) films. Of course these categories can be further broken down into more distinguishing classifications such as organic, porous organic, porous oxides, doped oxides, etc. Regardless of the specific low-k dielectric we are discussing, there is often a degradation of mechanical properties, including modulus of elasticity. As a consequence of this reduction in modulus, it has become extremely difficult to obtain high quality mechanical cross section. Whereas analysts were once able to produce smooth polished surfaces in older oxide technologies, they are now faced with the regular occurrence of smudging and smearing of soft materials, and fracturing of more brittle materials.

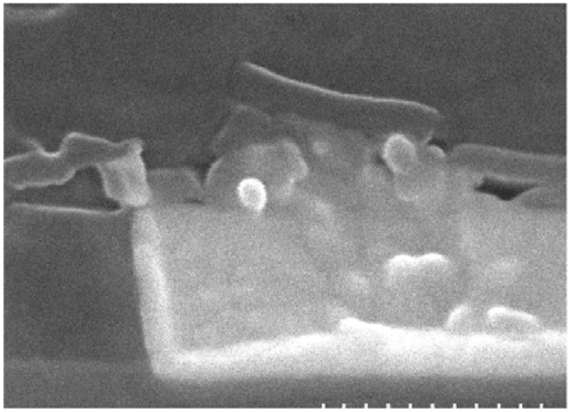


Figure 3: Example of damage created during mechanical polish of a copper/low-k dielectric system

In either case, it has become increasingly difficult to separate actual appearances from sample preparation related artifacts. Additionally, as was alluded to in an earlier section, there is inherent difficulty in mechanical sectioning of porous materials, and many of today's (and tomorrow's) low-k dielectrics are, in fact, porous. In either case, a manual mechanical cross section can be initially used to get the analyst close to the desired target (2-5um). The subsequent FIB slices away the damaged surface and remaining

real estate, affording the analyst a clean finished product.

FIB as an Alternative

Not only can FIB be utilized in conjunction with manual cross sectioning, but it can also be implemented as a cross sectioning technique on its' own. In such instances, the mechanical close-polishing step is skipped altogether, and the sample goes directly into the FIB. Figure 4a and Figure 4b are low magnification ion beam and SEM images respectively, depicting different examples of the resulting structure from this FIB approach.

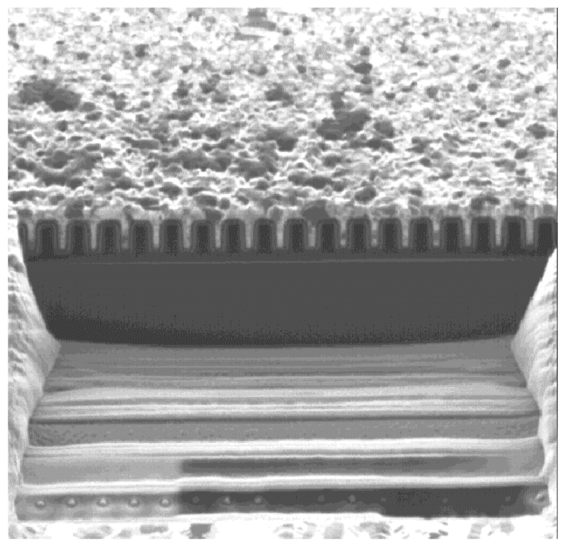


Figure 4a: "Head-on" view of a FIB step-mill box drilled directly into the surface of the chip

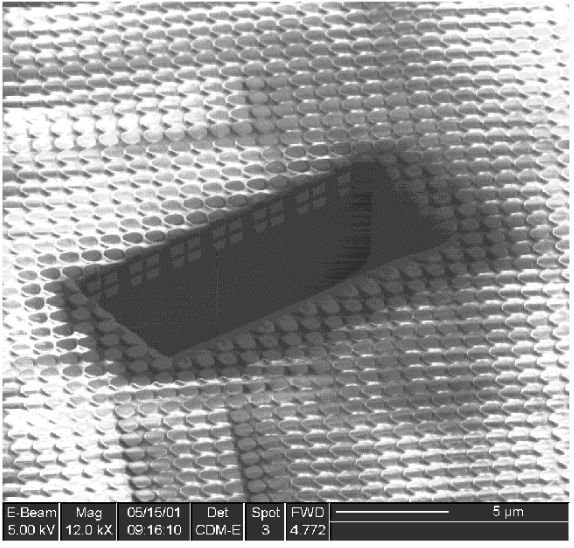


Figure 4b: Angled view of a FIB step-mill box drilled directly into the surface of the chip

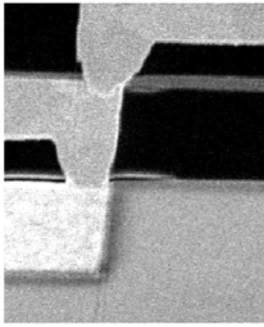
As you can see, this is a technique by which a hole, or a series of holes, is dug down into the top surface of the chip. It has been dubbed a "step mill" because often times there are a series of excavations at different depths that give a staircase effect. The final appearance of the hole will depend largely on the tool, and the speed at which a large hole can be milled. In either case, there must be an opening large enough so that the analyst may look, at an angle, into this hole, at the final cross sectional plane.

There are several reasons or situations where an analyst might opt for this approach. The first reason is speed. Depending on the skill and experience of the analyst, and the difficulty associated with hitting particular targets, a complete manual cross section can take up to 4-5 hours. In the case where only a close polish is being performed for subsequent FIB, this time can be cut in half, but the sample then requires about two hours of completion time on the FIB. Several hours can be shaved off of final turn-around-time by forgoing any mechanical sectioning and going directly into the FIB.

Another reason to use the FIB as an alternative is to help overcome navigation related issues. Defects are frequently identified with high-powered SEM's. Often, these defects are within large arrays that do not offer any good landmarks. In these situations, it is extremely difficult to navigate back to a target location. It is nearly impossible to section a target that one cannot find or see optically. Because the FIB also offers good high-powered magnification capabilities, the analyst can easily locate the area of interest. As soon as this site is found, sectioning can begin. Once again, this advantage becomes amplified by the dual beam FIB which will allow the analyst to maneuver around the chip in SEM mode, without the destructive effects of the ion beam.

Artifacts and Issues

As with any technique, there are of course some shortcomings associated with the use of FIB as a cross sectioning tool. It would be unfair to praise the virtues of FIB without acknowledging the drawbacks. Perhaps the largest problem associated with imaging on the FIB (ion beam only) is that the beam itself is destructive. One must be cautious not to sit in one location for too long, especially in the target area, for fear that it can create irreversible damage. In regard to imaging, there is another problem that is specific to organic low-k dielectrics when using the real-time electron beam update feature on a dual beam FIB. Notice the delamination at the interfaces in Figure 5.



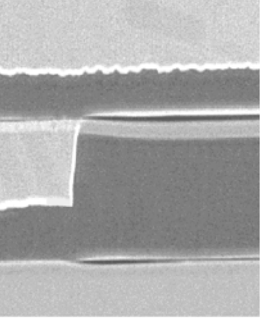


Figure 5: Example of SEM induced delamination occurring in a copper/low-k system

Figure 5 is an SEM image of a BEOL (back end of-the line) copper/SiLK^{TM} (Trademark Dow Chemical Corporation) integration scheme. As has been documented in other papers, SiLK^{TM} suffers from instability under an electron beam during imaging, and can shrink [2]. This problem can be partially avoided by taking fewer SEM images and focusing only when necessary. Another FIB artifact linked to low-k organics is induced conductivity due to gallium implantation in exposed dielectric surfaces [2,3]. This can have an effect if one is attempting passive voltage contrast (based on charging differences) in cross section, and may also affect dielectric contrast during subsequent imaging; this will influence layer delineation relative to certain materials [2].

One of the most common FIB related artifacts are striations as can be seen in Figure 6. This image was taken in a "step-mill" done from the backside of a chip (this is why all the structures are upside down).

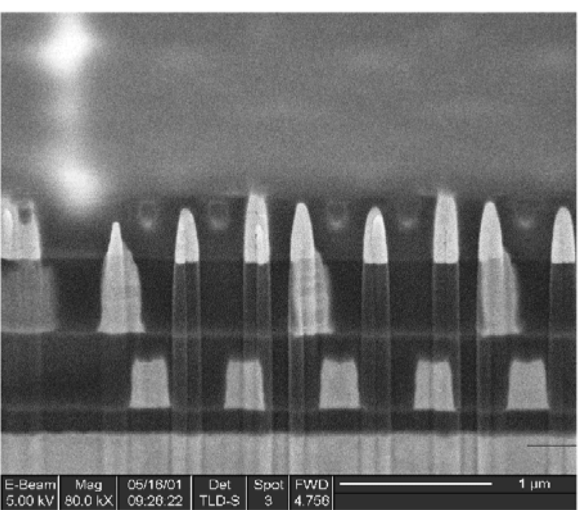


Figure 6: Example of FIB induced striations

These striations occur as a result of uneven mill rates and propagate downward in the direction of the ion beam. They usually coincide with the edges of metal features. Once can almost completely eliminate these surface imperfections by parallel polishing off above levels. This will effectively remove any features that may initiate the striations. Even when present, striations are usually a low level problem and can be ignored.

There are also other concerns that may not necessarily be labeled as disadvantages, but must be considered nonetheless. The first of these is specific to single column FIB's. Although many of these tools are equipped with some type of end point detection system, there is really no foolproof way of stopping at the target because of the lack of real time update of the section plane. This is especially true when the analyst is attempting to stop in a buried structure that cannot be seen from the top. However, there is a technique that can be employed to assist the analyst in discerning where the cut is in respect to the target. By digging "seeker holes" that are in line with the desired target, the analyst can peer down into the buried structure, and monitor a landmark by which to gauge their progress. It is advantageous to dig these holes a small distance from the target, as opposed to immediately on top of the target, as to not alter the appearance with the destructive ion beam.

Another consideration, which is specific to using the FIB as an alternative to a manual mechanical cross section, is the ability to see into the hole the analyst has excavated. If the FIB does not have SEM capability, or if high-resolution SEM (in-lens) imaging is required, the imaging will have to take place on another tool after the FIB section is complete. One must be aware that small holes on the surface of a die may not be "SEM friendly;" It can be difficult to get beam in and signal out. As a result, one must be sure to make the hole large enough, particularly in the direction perpendicular to the section plane, so that the sample can be tilted and imaged. This becomes even more important the further the objective is from the die surface. If the hole is too small, the analyst will encounter difficulty when attempting to image.

Finally, the analyst must always be alert to the aperture that he is currently using as one may not want to image and cut using the same settings. Charging and arcing between structures at different potentials is a definite problem that needs to be avoided. Using a "flood gun" or charge neutralization feature will minimize charging, and creative grounding of particular structures may also help alleviate this issue, however, being cognizant of the aperture is the best defense.

Conclusions

The FIB has proven to be an extremely powerful cross sectioning tool that allows for high precision

and accuracy with exceptional repeatability. It exhibits excellent control, such that it benefits the analyst when faced with cutting edge feature sizes. It is well established that employing the FIB in this capacity can reduce analysis cycle times with little or no sacrifice. In fact, it has been demonstrated that it can be of significant worth in increasing the final quality of the sample when dealing with emerging technologies such as low-k dielectrics. FIB is certain to be an everyday tool and a necessary piece of equipment in every failure analysis lab in the not-too-distant future.

Acknowledgements

The authors wish to thank Bruce Redder, Jack Fitzsimmons, Steve Herschbein, and Terry Kane, all from IBM Microelectronics, for their council and contributions.

References

1) T.Joseph and F.Trudeau, Metallographic Techniques for semiconductor Device Failure Analysis, Microelectronics Failure Analysis Desk Reference, Fourth Edition, pp 253-266

2) H.Bender and R.A.Donaton, "Focused Ion Beam Analysis of Organic Low-k Dielectrics;" Proceedings from the 26th International Symposium for Testing and Failure Analysis, Bellevue, Washington, 12-16 November, 2000.

3) T.Kane, P.McGinnis, B.Engel, "Electrical Characterization of Circuits with Low-k Dielectric Films and Copper Interconnects," to be published in Proceedings from the 27th International symposium for Testing and Failure Analysis.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Delineation Etching of Semiconductor Cross Sections

S. Roberts, D. Flatoff
South Bay Technology, Inc., San Clemente, CA USA

Introduction

Sample preparation of semiconductor devices has been an ongoing challenge in failure analysis since the advent of the first discrete transistor device fabricated in the 1960's. As device geometry, density, and materials have evolved throughout the years, techniques have progressed to allow the maximum amount of information to be obtained from the device using the analysis technique of choice.

Failure analysis will often require a cross section of a particular device to help determine what the failure mechanism and, ultimately, what the source of the failure is. There are numerous methodolgies available for preparing cross section samples, typically using a series of abrasive grinding and polishing steps to remove material in a controlled, precise manner. Ultimately, the cross section is subjected to a final micro-finishing process to create a mirror finish without any variations in the surface of the section. Following specimen preparation the sections are evaluated using the scanning electron microsope (SEM) allowing direct sample imaging at relatively high resolution combined with elemental analysis capabilities.
Samples prepared for SEM investigation often require an additional sample preparation step commonly referred to as staining or delineation. Delineation of cross section samples is required to highlight specific features present and is due to the lack of contrast produced from the secondary electron signal. Delineation methods can also be used for removing specific feautres to enhance other more significant areas of a cross section.

Delineation Methods

There are a number of techniques available for delineation of a cross sectional semiconductor sample. Methods commonly used today are chemical etching, plasma and reactive ion etching, ion beam etching, and focused ion beam etching. Selecting a particular method for a sample depends upon the sample material and complexity, the desired feature of interest, the methodology employed, and acceptable damage to the sample. All delineation methods have advantages and disadvantages that must be evaluated prior to implementation into a given analytical problem. Understanding what effects the delineation will have on the sample and it's potential impact on the overall conclusion should be weighed with considerable thought. Often times a delineation can induce certain artifacts into the sample and should therefore be well characterized before implementation.

Wet Chemical Etching

Wet chemical etching is versatile in that a wide range of etchants are available to select from depending upon the desired feature of interest. Typically etch times used with this method are on the order of seconds but can range up to a few minutes. Although it is a common method wet etching is often difficult to control and over-etching of the feature can occur.

Wet etches are typically categorized by the material they are designed to etch away. These are oxide etches, metal etches, and Silicon etches. Although these etches are normally aimed at a specific material of interest they will attack other materials and require careful control over the etch times and conditions. Wet etches are normally used by dipping the cross section in a beaker of solution for a specified time or using an eye dropper. Careful cleaning of the sample surface following etching using deionized water should be done prior to evaluation in the SEM. Some common etches and their applications will be described, however there are many variations available. For a comprehensive list see the reference guide found later in this publication.

Oxide etches are typically hydrofluoric acid (HF) based and in most cases are buffered with ammonium fluoride (NH_4F) in varying concentrations. A brief dip in potassium hydroxide (KOH) following BOE can be used to decorate the dopant profiles in cross sections as well [1].

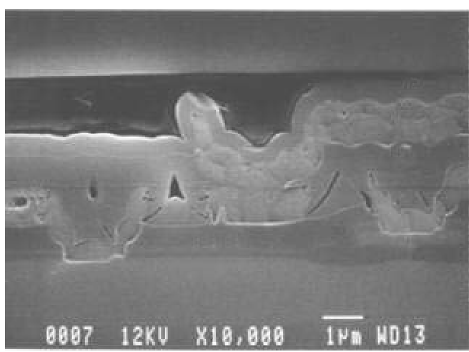


Figure 1: Aluminum based interconnect cross section polished and etched using a 10:1 BOE etch. Defects in the oxide are a result of preferrential etching.

Metal delineation etches require a working knowledge of the materials inside the device. Aluminum metal interconnects are often delineated using hydrochloric or phosphoric acid. Other metals such as tungsten and titanium can be delineated with a solution of sodium hydroxide and potassium ferrocyanide in water. Barrier metal delineation can also be achieved to distinguish it from the silicide using this process as well.

Silicon delineation etches for examination of polysilicon and junctions have basically two types. One is comprised of nitric acid and HF, the other chromium oxide acid and HF. Some common etchants used from these systems are the Dash etch and Sirtl etch. The Dash etch, when used for silicon etching, consists of 1 part HF, 3 parts nitric acid, and 8-12 parts CH_3COOH. Long etching times from 1-16 hours are typical for dislocation revelation when using this etch. The Sirtl etch used for stacking fault delineation in silicon consists of 46 grams of CrO_3 in 100 grams of 40% HF solution. It is typically stable with use and has a wide range of usefulness in various concentrations.(defince the two afore mentioned etches) The aggressiveness of these etches requires short etching times and the realization that the metal and oxide materials in the sample will be sacrificed [2].

Plasma / Reactive Ion Etching

Another method in the delineation of cross sectional features is the use of a reactive gas plasma. Both plasma and reactive ion etching utilize a dry chemical etching process that reacts with the cross section surface in a specific manner to delineate features on the sample. Plasma etching is more isotropic in application whereas reactive ion etching provides directional etching capabilities which can be advantageous in delineation applications. Dry etching provides greater control over the etch rate when compared to wet etching, and in many cases is a highly selective process depending on the process gas used.

Dry etching also encompasses a large parameter set and can sometimes be difficult to characterize for cross sectional samples of heterogenous composition. Parameters such as system configuration (barrel, parallel plate), applied power, vacuum pressure, gas flow, and gas concentration all have an effect on the final sample outcome and can be optimized for a specific material or sample. For the same reasons dry etching presents challenges it also offers a high degree of flexibility that is sometimes required in failure analysis. The ability to control the process and minor consumption of specimen materials during processing enable unknown samples and materials to be etched experimentally and used as a benchmark for other samples.

Etching of polysilicon with selectivity to oxide and silicon substrate has been successful using a tetrafluoromethane (CF_4) process gas. Some metal etching has been successful as well for minor delineation of grains, and the addition of oxygen (O_2) into the process chamber has further improved this selectivity. Other process gases such as sulfur hexafluoride (SF_6) and chlorine (Cl_2) have also been found to be successful in the delineation of polysilicon and silicon with some minor benefits for other constituent materials. Plasma delineation of current Cu low-k device nodes has not been particularly successful due to the inherent problems associated with copper such as low volatility and oxidation.

Delineation etches using dry etching techniques can also create artifacts due to high energy ion bombardment or result from chemical reactions with process gases. Be aware of potential chemistry changes to samples during etching in a plasma reactor of any type.

Although dry etching presents some challenges, the control and selection of a dry etching process can enhance the sample with great benefits.

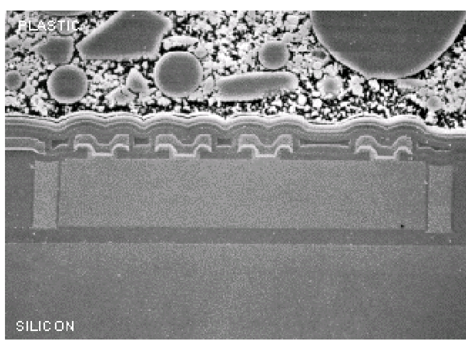


Figure 2: SEM micrograph of integrated circuit following a three minute sulfur-hexafluoride (SF6) plasma etch. 2,500X magnification. (From Analog Devices website, R Burgone. http://www.analog.com).

Ion Beam Milling

Ion beam milling of cross section samples has recently been applied to cross section sample preparation in a variety of ways. Ion milling utilizes a ‘broad’ ion beam of a few

millimeters directed at a predetermined angle to the cross section surface. Sputtering of material from ion bombardment occurrs and is affected by ion energy, ion beam angle to the sample, and sample material. Delineation of features is a result of differential sputtering rates and is material dependent.

Surface cleaning of abrasive contamination as well as the removal of mechanical polishing damage can be accomplished using ion milling for as little as 3-5 minutes. Modern instruments allow for sample rotation, tilt, and oscillation allowing for improved material removal. This technique has been found to be quite useful for copper low-k devices where enhancement of copper grain interconnect contrast and void observations in gap fill have been seen. This technique is particularly useful where particle chemistry analysis is to be done and contamination issues are key. Mechanical preparation is done to within a few thousand angstroms of the particle, followed by ion beam milling into the particle. Chemical information on the particle origin is retained without contamination from the polishing process [3].

Further delineation of surface features can be accomplished using reactive ion beam etching (RIBE) techniques. The implementation of a reactive gas such as iodine into the ion gun can be used to enhance the etch rate further reducing sample preparation time [4].

Ion beam milling can create unwanted surface topography if the proper parameters are not used. Large incident angles, high ion beam voltages and currents, and long milling times create large topographic contrast due to the preferential sputtering effect. Surface roughening can mask features of interest and will ultimately result in a loss of image resolution.

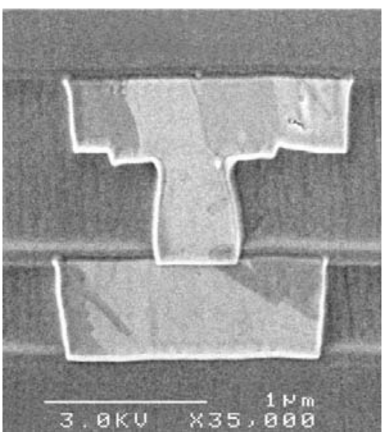


Figure 3: SEM micrograph of copper low-k contact following ion beam etching (From South Bay Technology, Inc.).

Focused ion beam milling can also be used for cross sectional delineation, although the FIB prepared cross section inherently contains contrast from differential milling of materials. Recent developments in process gases used in FIB instrumentation have increased the abilities of the system to mill traditionally difficult materials. Further discussion of the FIB can be found in additional sections of this reference guide.

Conclusion

Sample delineation of cross sections has been a requirement in semiconductor analysis since the first device was fabricated. Recent improvements in this process have followed with technological advancements and enable reliable evaluation of samples.

References

1. F. Trudeau, T. Joseph, 'Metallographic Techniques for Semiconductor Failure Analysis', *Microelectronics Failure Analysis Desk Reference,* p. 265, ASM International, Materials Park, Ohio (1999).
2. S. Perungulam, K. Wills, 'Wet Chemical Deprocessing Techniques', *Microelectronics Failure Analysis Desk Reference,* p. 265, ASM International, Materials Park, Ohio (1999).
3. R. Alani, R. Mitro, K. Ogura, H. Zhang, *New applications for SEM specimen preparation by an ion beam etching/milling/coating system, Proc. ICEM14,* 1, 367-368 (1998).
4. R. Alani, R. Mitro, K. Ogura, *Reactive ion beam etching technique and instrumentation for SEM specimen preparation of semiconductors, Proc. Micro. Microanal. 5,* 2, 912-913 (1999).
5. S. Roberts, D. Flatoff, *Enhanced Sample Preparation of Cu low-k Semiconductors via Mechanical Polishing and Ion Beam Etching, Journal of Micro., Jan/Feb 04,* 41-42 (2004).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Special Techniques for Backside Deprocessing

Seth Prejean, Brennan Davis, Lowell Herlinger, Richard Johnson, Renee Parente, Mike Santana
Advanced Micro Devices, Austin, Texas, USA

Abstract

This paper is about backside deprocessing of CMOS devices. The suite of techniques presented here was developed for bulk silicon and Silicon On Insulator (SOI) technologies using <100> silicon substrate orientation. This methodology was developed on flip chip packaged devices as well as unpackaged die. The application of this backside approach on different packaging technology is also possible. The advantages of using backside deprocessing subsequent to backside FA and fault isolation will be discussed. An outline is given for the tools and techniques used to address both global and local deprocessing from the backside.

Introduction

The idea of physically deprocessing a semiconductor device from the substrate or "backside" has evolved into a technique commonly used for failure analysis. This evolution was brought about by the increased use of backside fault isolation techniques on flip chip packaged devices. Many of the basic sample preparation procedures for backside fault isolation make subsequent frontside deprocessing difficult. The increasing complexity of the metal interconnects has also put limitations on frontside deprocessing techniques. Backside deprocessing provides an effective method for analyzing devices at the transistor level. According to the semiconductor roadmap, devices will continue to increase in complexity thus it is not surprising that current trends are leading to backside deprocessing. (1)

What will be described is an approach to backside deprocessing and how it is specifically used on flip chip packaged devices fabricated in both bulk and SOI technologies. However, these techniques may be applied to other silicon substrate based technologies though they are not explicitly described here. One main advantage of this deprocessing method for flip chip packaged devices is that the device can remain fully functional after substrate removal. Beyond substrate removal the device is still electrically connected to the package pins. The backside methodology exploits another advantage of flip chip packaging technology in that the device can remain packaged for deprocessing. Currently, many fault isolation techniques have been adapted for use on flip chip devices. Some of these isolation techniques require the device substrate to be thinned thus making the die more susceptible to physical stress and possible damage during fault isolation. The die in this fragile state has also been proven difficult to remove from the package for further analysis or deprocessing. (2) A summary of backside fault isolation techniques and tools in conjunction with the use of backside deprocessing for root cause failure analysis is described in "Failure Analysis Turned Upside Down". (3)

Packaged vs. Unpackaged Devices

The ability to deprocess devices in several forms is a challenge for any failure analysis lab. Unpackaged devices may come to the lab directly from the fabrication facility for analysis and deprocessing. Packaged devices may come from customer returns, test fallout, or qualification failures in such form factors as flip-chip, dual inline packaging (DIP), plastic leaded chip carrier (PLCC), thin quad flat packaging (TQFP), as well as many other forms not mentioned here.

As stated above, a variety of packaging technologies exists throughout the semiconductor industry. The types of packaging solutions addressed here are for the flip chip technology on ceramic and organic package substrates. The backside deprocessing technique may be applied to non-flip chip package technologies, provided access to the device backside is attainable. There are several proven methods for accessing the backside of non-flip chip devices. One method is mechanical package milling with a Computer Numerical Controlled (CNC) tool. (4,5) Another method is parallel polishing and chemical etching of the package material, which is used in conjunction with the CNC method described above.

Unpackaged die can come directly from the wafer or removed from the original package substrate. Removing the die from the original package may be necessary if the backside of the die is not readily accessible for backside deprocessing. In either case, the unpackaged die must be mounted to a rigid substrate, which provides structural support to the circuitry enabling backside deprocessing. Figure 1 shows the mounting scheme described here. Possibilities for the substrate are a sacrificial die, a bare piece of silicon, a glass slide or other suitable materials that will provide support for the die. A chemical-resistant bonding agent should be used to attach the

die to the substrate. The bonding agent should also maintain its adhesive properties when exposed to the temperatures encountered during deprocessing. Many common epoxy adhesives have been found to be a suitable choice for the bonding agent.

Die (circuit side on epoxy)
Substrate
Epoxy

Figure 1: Diagram of die to substrate mounting scheme.

Deprocessing Methods

Deprocessing globally from the backside provides a means for circuit inspection of the entire device. The technique is similar for both bulk and SOI devices in that the initial step involves complete removal of the silicon substrate.

The majority of the substrate is removed mechanically by parallel polishing while leaving a minimal thickness of silicon. The remaining silicon is then removed using a highly selective chemical etchant such as Tetramethylammonium Hydroxide (TMAH) or Cesium Hydroxide CsOH. The image in Figure 2 shows an optical photo of the bulk silicon etched by TMAH. The circuitry is visible where the bulk silicon is completely removed. Care must be taken with the etch duration of bulk devices since the chemical etch will remove active or doped silicon regions of the circuit along with the bulk silicon. However, with SOI devices, the Buried Oxide (BOX) layer acts as a deprocessing buffer zone providing a chemical etch stop during the substrate silicon etch. It also prevents propagation of surface scratches created during the polish step.

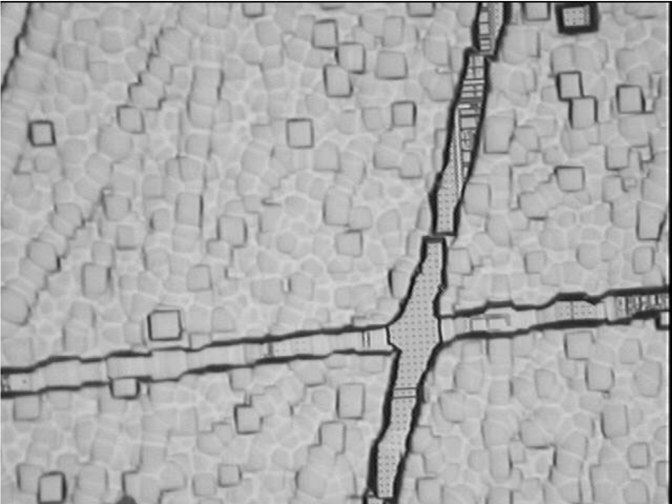


Figure 2: Optical microscope photo of bulk silicon etched in TMAH.

Local bulk silicon removal can also provide access to the area of interest for backside analysis techniques. The local method involves opening a trench through the backside silicon over the area of interest while leaving approximately 20μm of silicon at the base of the trench. The trench dimensions should be large enough to uncover the entire area of interest. This less destructive method can better preserve the functional integrity of the device.

Initial preparation for the local deprocessing requires global thinning of the bulk silicon to a manageable thickness. Trenching through a bulk silicon thickness greater than 300μm becomes time prohibitive for local silicon etching. In addition, a high aspect ratio trench can limit full access to the area of interest. Local silicon etching can be done with a Focused Ion Beam (FIB) tool or a Laser Chemical Etcher (LCE). (6,7) Once the trench is completed, the remaining silicon can be removed with the same chemical techniques as described above for global deprocessing. After the silicon is completely removed from the base of the trench on bulk devices, the gate dielectric is exposed. As previously described, the BOX layer provides a chemical etch stop on SOI devices.

For both local and global deprocessing on SOI devices the BOX layer can be removed using a wet chemical etch such as Hydrofluoric Acid. Chemical exposure time is critical since over etching can damage the active silicon and the field isolation oxide. An alternative removal method can be dry plasma etching with a standard oxide removal recipe. Given the nature and chemistry of dry etch techniques, there are more variables that can affect the etch results. The complexity of the dry etch method is also increased due to plasma interaction with package materials. The repeatability of the wet etch makes it the preferred method for BOX removal.

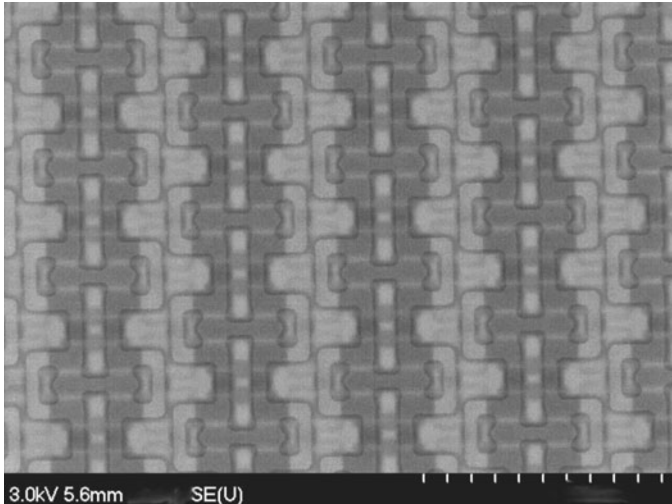


Figure 3: Backside SEM micrograph showing the dopant contrasting effect.

Note: During BOX removal, while a thin layer of it remains, a dopant contrasting effect can be observed in the SEM where p-type and n-type silicon regions are differentiated. The p-type silicon appears brighter than the n-type silicon due to a difference between the work function energies of the doped silicon. An optimum oxide thickness enhances the contrast effect. (10,11,12,13,14) Figure 3 shows the dopant contrast effect on an array of memory during BOX removal.

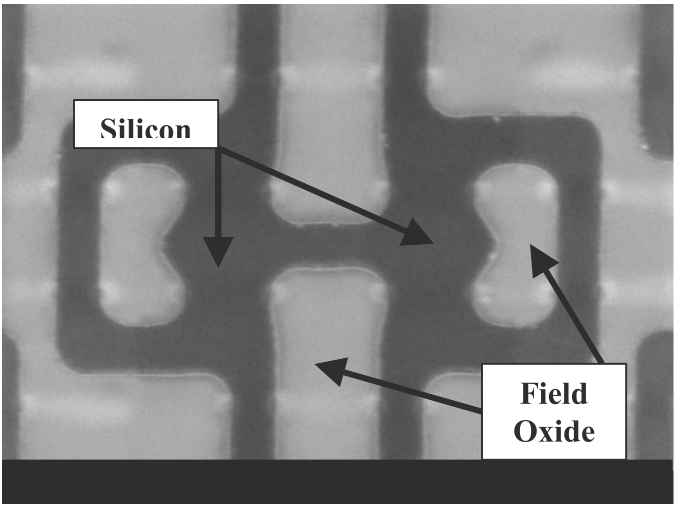


Figure 4: Backside SEM micrograph of active silicon layer.

Once the BOX layer has been completely removed as shown in Figure 4, the active silicon and field isolation oxide areas are exposed. For the purpose of examining the gate dielectric, the active silicon can be selectively removed. The preferred chemical method for active silicon etching is with CsOH because of its high selectivity. However, TMAH or dry plasma etching can suffice for silicon removal. In the case of an oxide gate dielectric, high etch selectivity to silicon over oxide is critical to preserve the integrity of the thin gate oxide. The active silicon removal also allows for inspection of the source and drain silicide. On bulk silicon devices the active silicon is etched during silicon substrate removal. Many anomalies can be observed after the active silicon is removed such as gate dielectric pinholes, ESD damage and silicide stringers as shown in Figures 5,6 and 7.

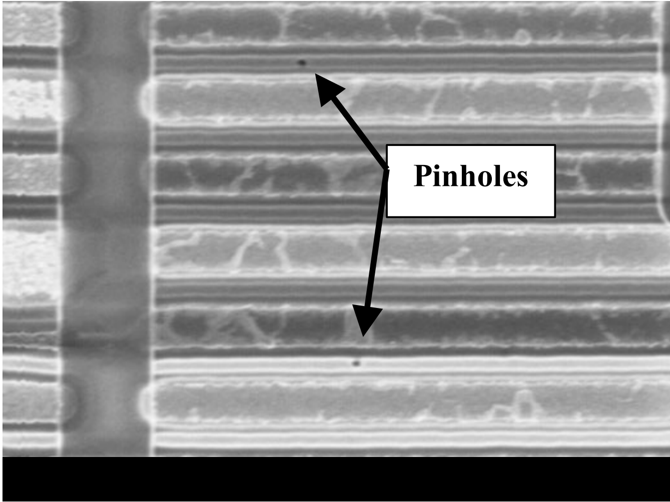


Figure 5: Backside SEM micrograph of gate oxide pinholes after active silicon removal with CsOH.

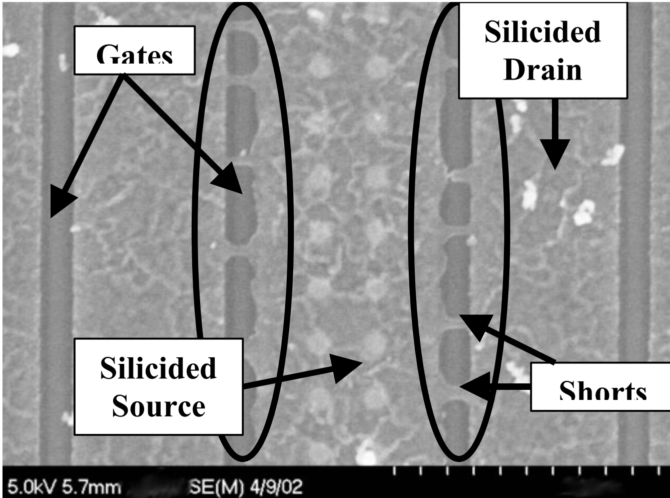


Figure 6: Backside SEM micrograph of Silicide shorts due to high current caused by ESD stress.

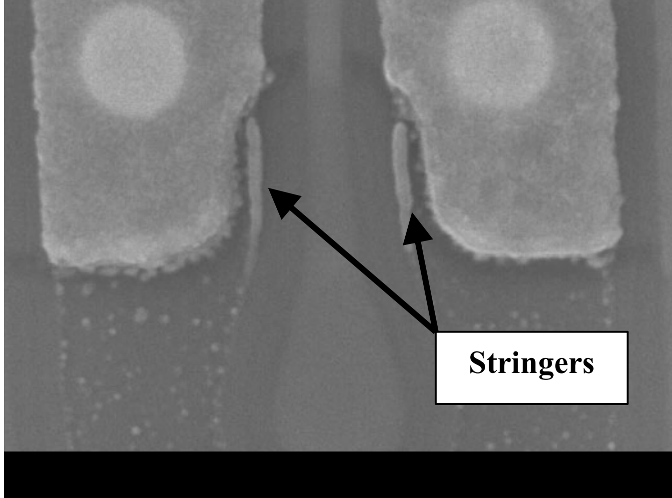


Figure 7: Backside SEM micrograph of silicide stringers along poly.

In order to continue deprocessing, the next step would be to remove the gate dielectric and then the gate itself. Gate oxide (SiO2) can be removed with diluted HF or dry plasma etches. If inspection of the entire polysilicon layer is desired then the field oxide must be removed. This can be done with HF or oxide plasma etching. Then polysilicon gate removal can be accomplished with a standard polysilicon etch, CsOH, TMAH, or dry plasma. Once the removal is complete, the gate silicide can be inspected. Figure 8 shows the polysilicon once the field oxide is removed.

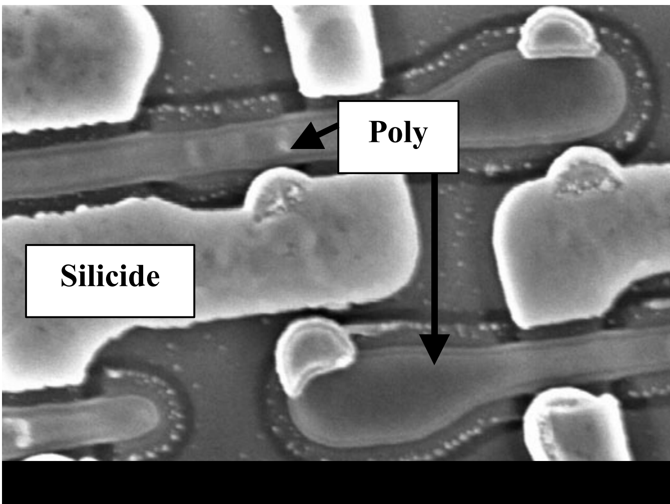


Figure 8: Backside SEM micrograph of transistors after active silicon and field oxide removal leaving poly, spacers and salicided source drain regions.

Traditional frontside deprocessing techniques can then be applied to the backside of the device for continued layer-by-layer inspection of the remaining metal interconnects.

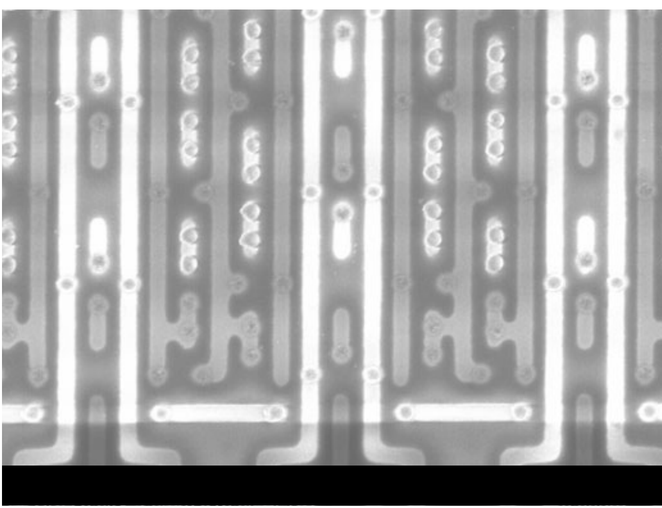


Figure 9: Backside SEM micrograph of polished metal layer.

Conclusion

In conclusion, backside deprocessing is a viable technique for root cause failure analysis. The trend towards flip chip packaging and increased metal layer count for complex semiconductors has created the demand for an alternative deprocessing method. When using this technique, SOI devices have an advantage over bulk devices due to the BOX layer that acts as an etch stop. This provides better control during bulk silicon removal. The technique provides easy access to the transistor active silicon and gate, while maintaining electrical connectivity to the package. In addition, backside deprocessing facilitates circuit edits and parametric measurements at lower metal layers and the transistor level. Local backside deprocessing is beneficial when thermal management is vital for full functionality. Backside deprocessing can be an indispensable technique in the failure analyst's toolbox.

Acknowledgements

The authors would like to acknowledge the AMD Austin Device Analysis Laboratory and Product Development Engineering for supporting the backside deprocessing development efforts.

References

1. ITRS Road Map http://public.itrs.net/Files/2003ITRS/Home2003.htm
2. Dat Nguyen, Cuong Phan, Jeff Conner, Martin Smith and John Drummond, *A Novel Approach to Front-side Deprocessing for Thinned Die After Backside Failure Isolation*, *Conference Proceedings from the 29th International Symposium for Testing and Failure Analysis,* 105-109, (2003)
3. Mike Bruce, Victoria Bruce, Seth Prejean, and Jeffery Huynh, *Failure Analysis Turned Upside Down, Electronic Device Failure Analysis Magazine, Volume, Issue, Month 2003, pp*
4. Silke Liebert, *Failure Analysis From the Backside of a Die, Conference Proceedings from the 26th International Symposium for Testing and Failure Analysis,* 177-185, (2000)
5. Phillipe Perdu, Romain Desplats and Felix Beaudoin, *Comparative Study of Sample Preparation Techniques for Backside Analysis, Conference Proceedings from the 26th International Symposium for Testing and Failure Analysis,* 161-171, (2000)
6. Rama R. Goruganthu, Mike Bruce, Jeff Birdsley, Victoria Bruce, Glen Gilfeather, Rose Ring, Nicholas Antoniou, Jesse Salen and Mark Thompson, *Controlled Thinning for Design Debug of C4 Packaged Devices, Conference Proceedings from the 37th Annual International Reliability Physics Symposium,* 327-332, (1999)

7. D. Davis, O. Diaz de Leon, L. Hughes, S. V. Pabbisetty, R. Parker, P. Scott, C. Todd, J. Widaski, G. Wilhite, K. S. Wills, J. Zhu, *Failure Analysis of Advanced Microprocessors through Backside Approaches, Conference Proceedings from the 24th International Symposium for Testing and Failure Analysis,* (1998)
8. Seth Prejean, Victoria Bruce and Joyce Burke, *CMOS Backside Deprocessing With TMAH / IPA as a Sample Preparation Procedure for Failure Analysis, Conference Proceedings from the 28th International Symposium for Testing and Failure Analysis,* 317-323, (2002)
9. Seth Prejean and Joseph Shannon, *Backside Deprocessing of CMOS SOI Devices for Physical Defect and Failure Analysis, Conference Proceedings from the 29th International Symposium for Testing and Failure Analysis,* 99-104, (2003)
10. S. L. Elliot, R. F. Broom, and C. J. Humphreys, *Dopant Profiling With the Scanning Electron Microscope – A Study of Si, Journal of Applied Physics,* Vol. 91, No. 11, 9116-9122 (2002)
11. C. Schonjahn, M. Glick and C. J. Humphreys, *Energy-filtered Imaging in a Field-Emission Scanning Electron Microscope for Dopant Mapping in Semiconductors, Journal of Applied Physics,* Vol. 92, No. 12, 7667-7671 (2002)
12. M. Aven, J. Z. Devine, R. B. Bolon and G. W. Ludwig, *Scanning Electron Microscopy and Cathodoluminescence of* $ZnSe_x Te_{1-x}$ *p-n Junctions, Journal of Applied Physics,* Vol. 43, No. 10, 4136-4142 (1972)
13. A. L. Bleloch, M. R. Castell, A. Howie and C. A. Walsh, *Atomic and Electronic Z-contrast Effects in High Resolution Imaging, Ultramicroscopy,* Vol. 54, 107-115 (1994)
14. C. P. Sealy, M. R. Castell, A. J. Wilkinson and P. R. Wilshaw, *SEM Imaging of Contrast Arising From Different Doping Concentrations in Semiconductors, Proceedings from Conference on Microscopy of Semiconductor Materials,* Vol. Institute of Physics Ser 146, 609-612, (1995)
15. A. Crockett and W. Vanderlin, *Plasma Delayering of Integrated Circuits, Microelectronic Failure Analysis Desk Reference Fourth Edition,* 243-252, (2003)
16. Alastair Trigg and Loh Peak Yong, *Sample Preparation for Backside Failure Analysis Using Infrared Photoemission Microscopy, Conference Proceedings from the 25th International Symposium for Testing and Failure Analysis,* 117-124, (1999)
17. Frank Zachariasse, *CMOS Front-End Investigation over Large Areas by Deprocessing from the Back Side, Conference Proceedings from the 27th International Symposium for Testing and Failure Analysis,* 237-241, (2001)

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Deprocessing Techniques for Copper, Low K, and SOI Devices

Huixian Wu, James Cargo
Agere Systems, Quality Organization, Product Analysis Lab
Allentown, PA USA

Introduction

Copper interconnects and low k dielectrics have been introduced in advanced integrated circuit (IC) technology to reduce the interconnect resistance, improve the resistance to electromigration and reduce RC delay and cross talk effects. The introduction of new materials in integrated circuits makes physical failure analysis more challenging. Moreover, as technology shrinks, defects can be smaller and more subtle, which generates additional challenges for defect localization and physical failure analysis. In addition, it is important to understand new failure modes and reliability issues related to Cu/low k technology and develop FA techniques accordingly.

In this article, several front side and backside deprocessing techniques will be presented. The deprocessing techniques include wet chemical etching, RIE, parallel polishing, chemical mechanical polishing (CMP) and a combination of these techniques. In addition, characterization results of CMP deprocessing will be presented as well as gate level deprocessing techniques for Cu/low k devices. Silicon-on-insulator (SOI) is becoming a leading technology for high performance and low power IC. Deprocessing techniques for SOI devices will also be presented.

Failure Modes and FA Challenges

Failure Modes

FA challenges will result from new failure modes and reliability issues associated with the integration of Cu and low k dielectrics, barrier metal and high k gate dielectrics. Soft defects and non-visual defects associated with advanced technology brings additional challenges for electrical defect localization and physical failure analysis:

Some failure modes associated with the integration of copper interconnect, low k dielectric and high k gate dielectrics include:

- Metalization level failure modes: *copper corrosion, contamination, interface diffusion, electromigration, stress migration, self heating, joule heating, stress voiding, copper extrusion, via voiding, interface adhesion*

- ILD failure modes: *cracks, interface adhesion, delamination, moisture adsorption, intra-level and inter-level leakage, bias-temperature instability, mechanical weakness, low thermal conductivity, and breakdown of the low k dielectrics*

- Gate level failure modes: *breakdown of the gate dielectrics, soft breakdown (SBD), stress induced leakage current (SILC), direct tunneling, hot carrier injection (HCI), trap assisted tunneling (TAP), negative bias temperature instability (NBTI), interface instability*

- Integration issues: *thermal-mechanical issues, interface adhesion, CTE mismatch, chemical reaction, Joule heating effects*

As the technology continues to scale down, defect size tends to decrease and as the circuits become more complex, subtle statistical variations may become significant failure mechanisms [1]. Any thermal/electrical instability with the increase of bias, temperature, or moisture in smaller features becomes more significant in advanced technologies. Some other failure mechanisms, not visible with microscopes, may occur more often in failure analysis. Detection and characterization of non-visual defects will be a FA challenge. Potential non-visual defects could be associated with the following:

• Non-visual defects: *NBTI (Negative Bias Temperature Instability), Intra-level or Inter-level leakage (Especially for the Cu/low k technology), instability with Cu/low k integration, stress induced leakage current (SILC), trap-assisted gate dielectric leakage current, gate induced drain leakage current, sub-threshold leakage current, interface instability related to high k gate dielectrics, Charge migration, short channel effects, interface adhesive problems and so on.*

FA Challenges
New failure modes and reliability issues associated with copper technology will introduce FA challenges. Understanding of these new failure modes and reliability issues is extremely helpful and highly recommended during failure analysis process. The introduction of copper interconnect, low k dielectric, barrier materials, and high k gate dielectrics will generate additional physical FA deprocessing challenges.

Copper interconnects are softer than aluminum and delayering copper presents several challenges, such as global uniformity problems, dishing of metal interconnect, copper smearing, and erosion of copper and dielectrics. For delayering low k dielectrics, parallel polishing has limitations due to its non-uniform process and low removal rate. The conventional wet chemicals used for ILD etching are not suitable for some low k dielectrics. RIE may be an alternative due to certain advantages. However, there are still some challenges for RIE of low k materials including etch selectivity, RIE grass, surface roughness, and optimization of the process parameters. Nevertheless, several deprocessing techniques such as parallel polishing, wet chemical etching, CMP and RIE have been developed for Cu/low k technology. The combination of the CMP and RIE works well for most situations. RIE is used to remove ILD materials to expose the copper lines and the copper is then removed by CMP.

Deprocessing of Copper Interconnect–Chemical Mechanical Polishing (CMP) and Electro-polishing (EP)

Deprocessing of copper metal layers – CMP
Deprocessing of copper by CMP presents several challenges to the FA analyst since it is easily corroded, is softer than aluminum and smears easily while polishing. Figure 1 shows an example of how copper smears when polished improperly. In addition, copper CMP can create some defects in the inter-level and intra-level dielectric materials. Particle contamination on the polishing pad may result in severe scratches during deprocessing.

Experiments can be carried out and characterized, which result in optimized polishing parameters. Figure 2 shows CMP deprocessing of copper using an optimized process.

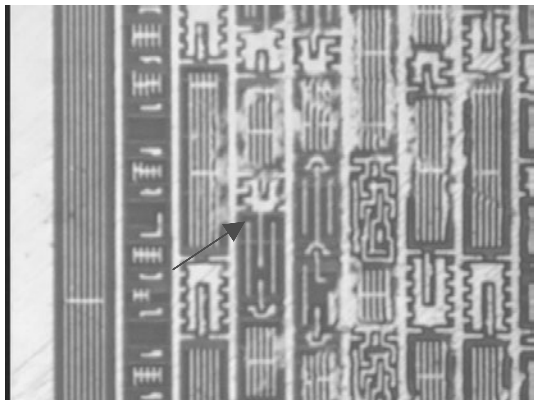


Figure 1. Image showing copper smearing

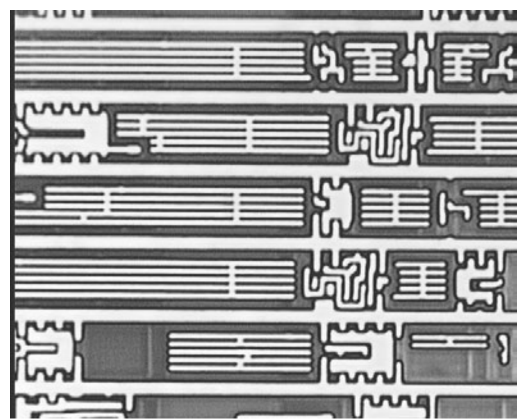


Figure 2. Image showing CMP of copper/low k sample

Characterization of CMP process

The polishing parameters, which need to be investigated, include slurry chemistry, polishing pad type, pressure, and pad rotation speed. The type of slurry and polishing pad type chosen play a significant role in the dishing, recess and erosion performance. Several types of slurry chemistry can be used for delayering copper including 0.05 and 0.06 micron Colloidal Silica Suspension slurry, diamond slurry and copper CMP slurry (Planar Solutions - CU10K2). Copper CMP slurry works the best in our applications. Moreover, pad rotation speed and pressure have a significant impact on the CMP process. Lower pad rotation speeds and lower pressure can reduce problems in the CMP delayering process and avoids over-polishing of copper, thus reducing dishing and erosion. The polishing time to remove the same amount of copper material was recorded at different polishing conditions. Figure 3 shows the impact of pressure on polishing rate at different pad rotation speeds. Lower pad rotation speed and lower pressure result in slower polishing rate.

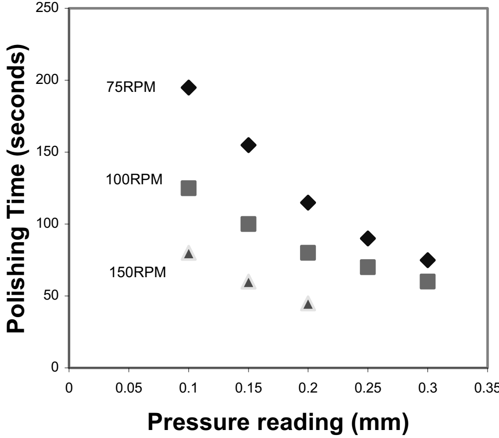


Figure 3. Polishing time vs. pressure reading

It's better to use a lower pad rotation speed (75, 100 RPM) and a lower pressure reading (0.1 mm), but it takes more time. Therefore, the multi-step CMP techniques, with different polishing conditions, can be extremely useful for deprocessing copper layers. Lower pressure and lower pad rotation speed are used in the final step to avoid over-polish of the copper.

Electro-polishing

Dishing and erosion caused by CMP can degrade the interconnect planarity. Electro-polishing is an alternative to remove copper, which is a CMP-free process and can minimize the dishing/erosion and film peeling concerns [5].

During wafer process, electro-polishing removes the copper uniformly across the wafer. However, non-uniform topography of conventional electrochemical plating (ECP) makes electro-polishing not uniform for different areas [5]. A combination of EP and CMP can minimize the dishing/erosion and obtain uniform surface. This method should be appliable to FA deprocessing. For removal of the thick, wide metal buses, this method should be able to achieve the flat, uniform and clean surface. More work should be investigated.

Deprocessing of Copper Interconnect– Dry Etching and Wet Chemical Etching

The degree of anisotropy and etch selectivity are two critical parameters in the etching process. An isotropic etch has the same etch rate in every direction and over-etch may cause undercutting or lifted layers. An anisotropic etch is unidirectional, which is more suitable for deprocessing.

Dry etching such as plasma etching or reactive ion etching (RIE) has the advantages of a high degree of anisotropy. Several process parameters to be considered in dry etching are power, pressure, gas flow rate and temperature, which can be characterized for obtaining a controllable etch rate. The etching chemistries to be used for metal etching are normally boron trichloride (BCl_3), Chlorine (Cl_2), Sulphur hexafluoride (SF_6) and so on. The chemistries used for aluminum etching are BCl_3 and Cl_2.

Normally both physical and chemical etching are involved in dry etching. Physical etching is the physical interaction between ions with high kinetic energy and the surface atoms, resulting in high anisotropy and low etch selectivity. Chemical etching is the chemical interaction between neutral or ionized species and the surface atoms to form volatile products, which results in isotropy and high etch selectivity.

Dry etching normally has a lower etch selectivity when compared to wet etching due to both physical and chemical interaction involved in dry etching.

Dry etching can introduce some etching by-product, typically called “grass”. RIE grass can polymer grass, sputtering grass, or metal grass. Additional cleaning of grass may be necessary for further failure analysis.

Dry etching of copper has specific challenges. Copper can be etched by chlorine-based reactive ion etching (RIE). It is known that the chlorination of copper and the volatilization of $(CuCl)_3$ are essential steps for RIE of copper. The process requires elevated temperature to obtain a high volatility and desorption rate for CuClx and thus achieve high etch rate.

In order to minimize isotropic etching or chlorine residues, it is necessary to reduce the thickness of the chlorinated film, which can be achieved by lowering the Cl_2 partial pressure [6].

Copper is easily corroded during the dry etch process, since copper does not form a self-protective layer to protect against corrosion. Corrosion can mostly likely be attributed to corrosive species trapped in the protection layer formed during the etch process [7]. Moreover, a high oxidation rate of copper brings dry etching of copper additional challenges.

Compared with dry etching, wet chemical etching typically has the advantages of a high etch rate and high etch selectivity. Wet etching can also achieve a clean surface because the etch by-product or debris can be dissolved in the wet chemicals.

However, wet etching is an isotropic process and the resulting over-etch can extend into the layers below through the vias. The etch rate can be varied depending on the composition and concentration of the acid chosen. Characterization of wet etching process used is helpful to get a controllable etch rate.

For wet chemical etching of copper interconnect, choosing the right chemical to give a controllable etch rate, high etch selectivity, uniform etching and a clean surface is of primary importance. Once a chemical is selected that has the required properties, the etching process must be fully characterized to avoid deprocessing included defects.

To investigate defects at barrier layers, highly selective etching/polishing is needed. The slurry used in copper CMP process can attack barrier materials, resulting in poor selectivity. Several chemicals can etch copper such as nitric, hot sulfuric, and a mixture of sulfuric and hydrogen peroxide.

Diluted nitric acid can be used for copper etching due to its high etch rate and high etch selectivity between copper and barrier layer (Tantalum). It should be noted that high concentrations of nitric acid are aggressive for copper etching due to a very high etching rate. In addition, etch selectivity drops in high concentration of nitric acid.

We found that after polishing into the copper layer, wet chemicals can be used to remove the remaining copper layer and leave the barrier layer untouched. Careful optical or SEM inspection of the barrier layer can then be performed for Failure Mode Analysis (FMA). Figure 4 shows where copper was partially removed by wet chemical etching. Figure 5 shows an image of the barrier after totally removing copper by wet chemical etching.

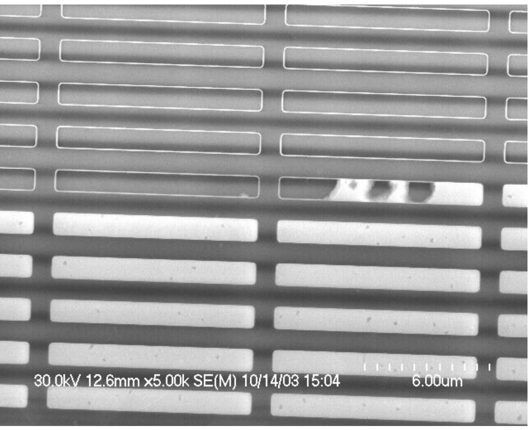


Figure 4. Chemical Mechanical Polishing into copper interconnect

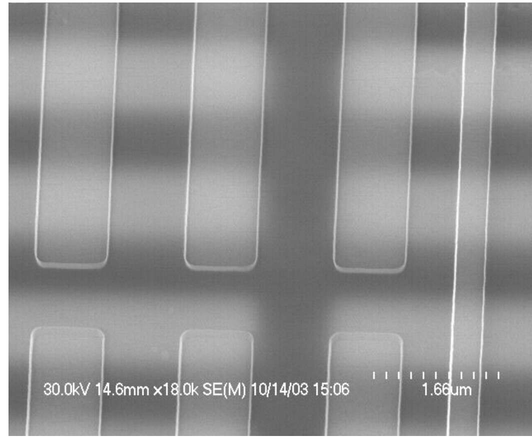


Figure 5. Barrier layer after wet chemical etches away copper

Deprocessing Barrier Layers

Copper readily diffuses into the inter- or intra-level dielectric and is very mobile in the dielectric, causing increased leakage in dielectric and increased junction leakage. Ta/TaN have been used as barrier materials to minimize the copper diffusion into dielectric. CMP, RIE and wet chemical etching techniques can be used for Ta/TaN etching.

Ta is soluble in hydrogen peroxide and HF but insoluble in nitric and other acids. TaN is slight soluble in HF, nitric and Aqua Regia. Ta and TaN can also be etched by CF_4 plasma based RIE process.

Etch selectivity between the barrier layer and inter-level dielectric layer is poor for both fluorine based RIE and HF based wet etching process.

To remove the barrier layer (Ta/TaN) without etching inter-level dielectric, hydrogen peroxide can be used to remove the Ta layer and then diluted nitric layer can used to remove the TaN layer. Silicon dioxide based inter-level dielectric is insoluble in both hydrogen peroxide and diluted nitric.

Deprocessing of Inter-level Dielectric (ILD) Layers – Reactive Ion Etching

With technology continually scaling down, the RC propagation delays become dominant in the overall chip delay. Low k materials are being used for ILD materials so as to improve the RC delay, and reduce the cross talk and power dissipation, as well.

Several techniques such as wet chemical etching, parallel polishing, and dry etching could be used for the delayering of ILD materials for FA. The concerns for a wet chemical process are isotropic etching and difficulty in controlling the etch process. Traditional wet chemical etching recipes for dielectric etching do not always work for low k materials. Parallel polishing has limitations due to its non-uniform removal process and low removal rate for low k materials. With technology continually scaling down, RIE may be an viable alternative due to certain advantages over parallel polishing and wet etching. These include high spatial resolution based on the anisotropy, high selectivity of the etching process, capability of etching devices with much smaller geometries and the ability to accurately control the etching process. However, there is not much work reported for RIE of low k materials for FA and many challenging issues remain. These include but are not limited to etch selectivity, surface roughness, and the optimization of the process parameters. Effective recipes for getting a clean surface, high selectivity and high etch rate are necessary for carrying out RIE of the ILD materials.

Reactive ion etching (RIE)

Reactive ion etching is a combination of physical and chemical reaction. RIE utilizes the synergistic effect of ion bombardment and chemical reaction in plasma. Since the wafer is placed on an electrode, the wafer experiences the bombardment from the ions accelerated to the electrode. The ion bombardment causes physical damage of the wafer material and facilitates desorption of chemical reaction products. As a result, chemical etching in a reactive ion etching process is dramatically accelerated.

RIE grass:

RIE grass is an etching artifact or by-product, resulting in an unclean surface. As a result, the surface needs additional cleaning for further FA. RIE grass occurs when etch resistant material, acting as micro etch mask, accumulate on the surface, resulting in formation of cones [8].

There are basically three types of RIE grass, which are sputtering grass, polymer grass, and metal grass.

When delayering a chip within a package by RIE, package materials will be sputtered and then re-deposited on the surface of the device, causing the formation of sputtered grass. Gold is also easily sputtered during RIE and introduces severe grass. To minimize the sputtering grass caused by package materials and gold bonds, separation of the chip from the package and removing the gold ball bonds before deprocessing is extremely helpful. It's also useful to mask the package materials and gold bonds during RIE deprocessing. Clean sample surface is critical before RIE process to minimize the formation of grass.

Polymer grass is composed of carbon-fluorine polymers, formed on the sample surface during fluorine based RIE. Polymer grass can also act as a micro-etch mask to accelerate the formation of RIE grass. A low fluorine/carbon ratio leads to increased polymerization. Adding more oxygen to CF_4 or other fluorine-based plasma can reduce the formation of polymer grass. Adding H_2 consumes F, increasing the polymer grass. The increase of carbon concentration can also increase the formation of polymer grass.

The addition of CHF_3 makes polymer grass worse. To minimize polymer grass, low chamber pressure with pure CF_4 is highly recommended [9].

For aluminum technology, the oxidation of aluminum by oxygen plasma or sputtering effect will introduce aluminum grass. Adding oxygen plasma increases the formation of this kind of grass. For copper technology, adding oxygen increases the oxidation of the copper, resulting in the etching by-product, which is not as severe as for aluminum technology.

In many cases, very diluted HF based wet chemicals can be used for removing RIE grass to achieve the clean surface.

Characterization of RIE (Etch rate and selectivity)

A RIE process for certain low k dielectrics has been characterized in our previous work [2]. Several RIE process parameters needed to be considered to optimize the RIE process, which are RF power, ICP power, pressure, and gas flow rate. The RIE process parameters also vary with different equipment. In our application, the optimized recipe for RIE of Black Diamond ™ material is: CF_4 of 50 sccm, O_2 of 2 sccm, chamber pressure of 50 mtorr, RIE power of 60 Watts and ICP power of 300 Watts.

With the optimization of the RIE process, high etch selectivity, low damage, and a clean surface can be achieved. Figure 6 and 7 show some SEM images using RIE deprocessing. The dielectric in Figure 6 and Figure 7 is Black Diamond ™ material and FSG respectively.

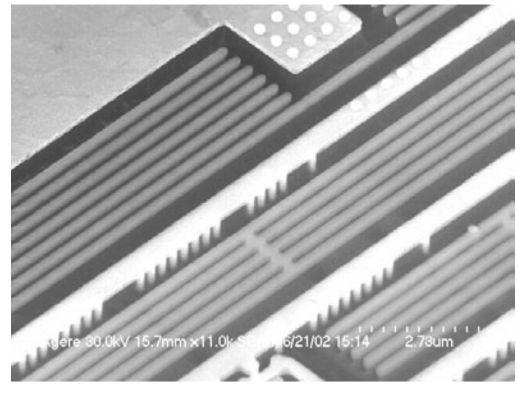


Figure 6. SEM image of 0.13 um Cu/low k technology

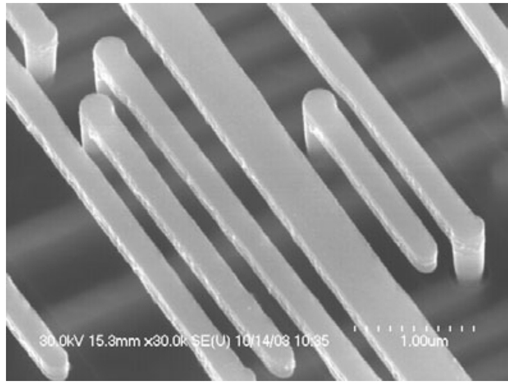


Figure 7. SEM image of 0.13 um Cu/FSG technology

Gate Oxide Integrity and Gate Level Deprocessing

The ability to perform physical FA at the gate level has become increasingly important with increased potential gate level defects, such as soft breakdown of ultra thin gate oxide, SILC, trap-assisted tunneling (TAT) and so on.

Gate Dielectrics Integrity Issues
With the scaling of gate dielectric and the introduction of high k gate dielectrics, soft breakdown of the gate dielectric becomes a serious concern for advanced CMOS technologies. After soft breakdown, there is an abrupt jump of conductivity at the low field but not at the high field. Gate current is noisy after soft breakdown, exhibiting a random telegraph noise behavior [4]. Detection of soft breakdown is a challenge during the FA process. Tunneling AFM (TUNA) is an ultra-low leakage current imaging tool, that might be used for localization of gate level defects, soft breakdown path, detection of dielectric thickness variations, and study of gate dielectric integrity issues. Scanning Capacitance Microscopy (SCM) and Scanning Spreading Resistance Microscopy (SSRM) can be used to obtain the 2-D dopant profile and junction localization.

High k gate dielectrics are being introduced due to the limitation of high gate leakage current for ultra-thin gate oxide. With the introduction of high k gate dielectrics, the stability of the high k/silicon and high k/poly-Si interfaces become a serious concern. High interface defect density affects the device characteristics and reliability. Mobility degradation and hot carrier charge trapping present additional challenges in high k integration. TEM and FIB are very useful analytical tools to study high k interface instability.

To detect physical gate defects and study gate dielectric integrity, a poly-Si etch process used to expose the gate dielectric is important and critical.

Wet chemical etching of poly-Si
Wet etching is normally used to remove gate poly-silicon for FA. Concerns with this method are its isotropic nature and over-etch problems. The correct wet chemical should be selected for high poly-Si/SiO_2 etch selectivity. Characterization of the wet etching process used is important.

A mixture of HF, nitric acid and acetic acid can etch poly-silicon. The etch rate and etch selectivity varies with the ratio of these three acids. The poly-silicon removal rate is high, however the etch selectivity is relatively low.

NaOH, KOH, TMAH, and Chlorine hydroxide can also be selected for poly-Si etching due to controllable poly-Si etch rate and very high poly-Si/SiO_2 etch selectivity.

50% wt. Choline hydroxide aqueous solution has been used for poly-silicon etching at temperature of 55 degree C with the etch rate of approximately 50 nm/min, and 92 degree C with the etch rate of approximately 400 nm/min [11].

In addition, 25% KOH can be used to decorate pinholes in gate oxide.

A mixture of nitric acid and buffered HF with a ratio of 9:1 is highly selective etch between poly-silicon and gate oxide, and it is a suitable poly-silicon etchant especially for removing poly-silicon with/without WSi_x and $CoSi_x$ [12].

A 30% wt TMAH can be used for poly-Si etch to expose gate oxide. The poly-Si etching process has been characterized at different temperatures. Figure 8 shows the impact of temperature on the etch rate of poly-Si and gate oxide. Figure 9 shows the impact of temperature on poly-Si/SiO_2 etch selectivity. Etch selectivity increases as the temperature decreases. [13]

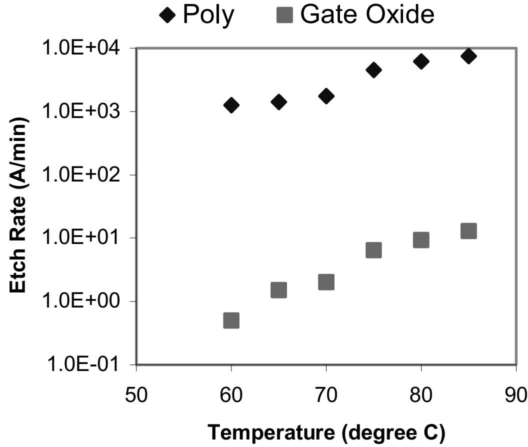


Figure 8. Temperature vs. the etch rate of poly-Si and gate oxide

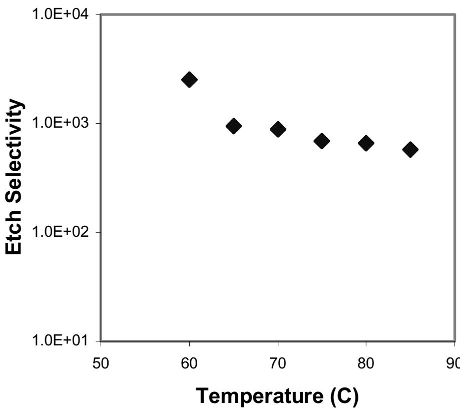


Figure 9. Temperature vs. poly-Si/ SiO_2 etch selectivity

Figure 10 and 11 are images of gate oxide in 0.13 um technology Cu/low k device after poly-Si etch process using 30% TMAH.

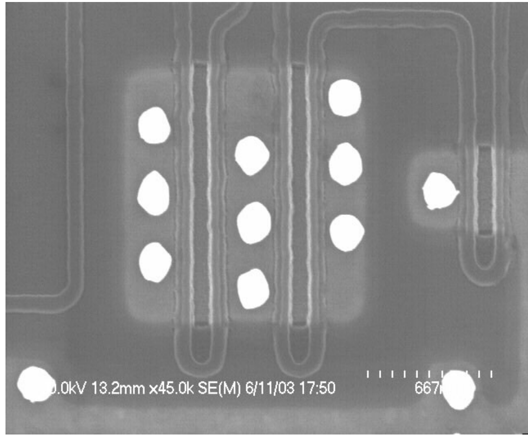


Figure 10. Image of gate oxide after poly-Si etching

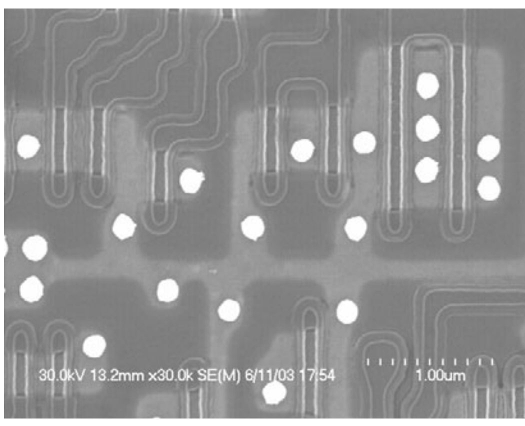


Figure 11. Image of gate oxide after poly-Si etching

RIE of poly- silicon

Etching poly-silicon with SF_6 plasmas is a highly chemical process, resulting in a fast and isotropic etching process. Adding oxygen to SF_6 can reduce the chemical etching rate and thus increase the ion assisted etching with a resulting higher anisotropic profile [14, 15]. To get high etch selectivity, low power SF_6 was used [10]. Highly anisotropic etching can also be achieved by the use of chlorine-containing plasmas due to the low reaction probabilities of chlorine or bromine radicals with silicon at room temperature. Moreover, high etch selectivity can be achieved for chlorine or bromine based RIE [16]. However, the poly-silicon etch rate is low for Cl or Br based RIE.

Backside Failure Analysis for SOI and Cu/low k Devices

Reliability Issues for SOI Technology

Silicon-on-insulator (SOI) is becoming a leading technology for high performance and low power IC due to its suppression of the short channel effect and idea sub-threshold characteristics.

SOI technology eliminates transistor latch-up and provides reduced junction capacitance, which allows the circuits to operate at higher speed. However, for SOI technology, electrostatic discharge (ESD) susceptibility becomes one of the major reliability issues [18]. This is mainly attributed to the poor heat dissipation due to the insulating buried-oxide layer, causing high temperature during ESD events.

SOI devices also suffer from the hot carrier injection induced degradation as the technology continually scales down [19]. The hot carrier degradation in SOI may be dominated by the bipolar parasitic and results in early drain breakdown. Self-heating is also a reliability issue due to the weak thermal conductivity of buried oxide layer. For the SOI devices, both gate oxide and buried oxide can be degraded due to the charge injection.

Plasma induced charge effects becomes more critical. However, SOI is less susceptible to charging damage [20].

Failures related to SOI usually occur at gate or substrate levels. Backside failure analysis becomes more important and critical.

Backside Failure Analysis for SOI and Cu/ Low k Devices

Conventional front side defect localization in IC devices has become less and less effective with an increase in the number of metallic layers for PEM, as technology continually scales down. Electrical defect localization in devices with SOI structures, devices with a high number of metallization levels and Flip-Chip Ball Grid Array (FCBGA) packages, can more easily be carried out from the backside of the die. Development of backside failure analysis techniques is increasingly important. Three major steps are involved in the backside FA process: backside sample preparation, backside defect localization, and backside physical analysis.

A combination of mechanical milling and RIE techniques has been employed for silicon thinning, which is a critical step in the backside sample preparation process [3]. In this process, RIE can be used for final silicon thinning after mechanical milling to achieve more precise control of thickness of the remaining silicon die while still preserving the electrical integrity of the device. Figure 12 shows a low power optical image of the backside PBGA package after the backside sample preparation process.

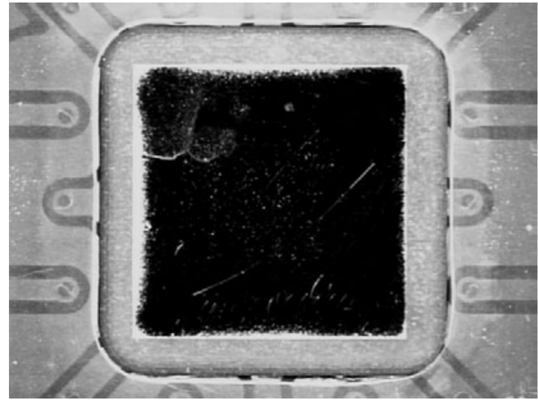


Figure 12. Low power image of package after backside sample preparation

Several backside defect localization techniques and tools have been used in the industry, such as Light-Induced Voltage Alteration (LIVA), Thermally-Induced Voltage Alteration (TIVA), Optical Beam Induced Resistance Change (OBIRCH), Picosecond Imaging Circuit Analysis (PICA), Laser Voltage Probing (LVP), Fluorescent Microthermal Imaging (FMI), and Photo Emission Microscopy (PEM). After backside sample preparation and backside defect localization, backside physical failure analysis may be necessary to find the root cause of the failure.

In certain package types, after the backside sample preparation process, there is a cavity in the backside of the package containing the die. Deprocessing the whole package from the backside is time consuming. Moreover, some artifacts and defects could be introduced during deprocessing of the entire package. A diamond-saw can be used to precisely cut into the cavity along the edge of the die and separate the chip from the rest of the package. Figure 13 shows a low power optical image of the backside of a die after the sawing step.

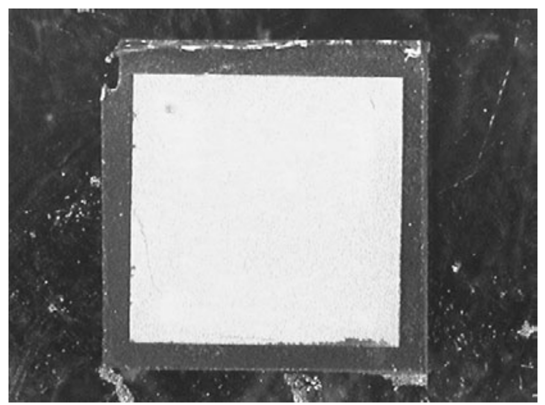


Figure 13. Low power optical image of the backside die after sawing of the package

RIE characterization of the silicon substrate for backside FA has been performed in previous work [3]. Further silicon thinning was performed using a combination of RIE and mechanical polishing until the remaining silicon thickness was approximately five microns. 25% TMAH wet chemical etching was used to selectively remove silicon material over gate oxide due to its high Si/SiO_2 etch selectivity. The advantages of TMAH over mechanical polishing for removing the silicon substrate are as follows: removing silicon without damage to gate oxide, better uniformity across the die, and better control of backside deprocessing. The Si/SiO_2 etch selectivity varies with different temperatures and at different TMAH concentrations. As mentioned previously, etch selectivity decreases with an increase in temperature.

Figure 14 shows a low power optical image of the entire die just after TMAH etching of the remaining silicon. The entire die was exposed uniformly without damaging the gate dielectrics.

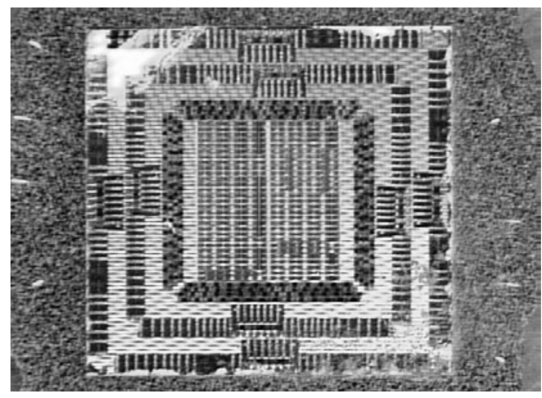


Figure 14. Optical backside image of the whole die

Figures 15 and 16 show high power optical images of the device during backside deprocessing. A combination of CMP and RIE was used to deprocess the device at the metallization levels during backside FA. The CMP method used for backside FA is similar to the one for front-side FA.

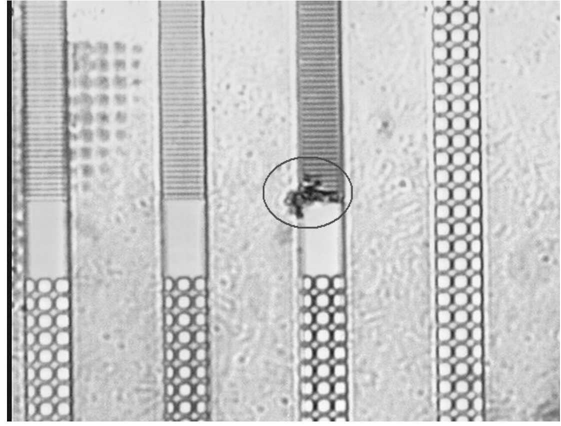


Figure 15. Backside optical image

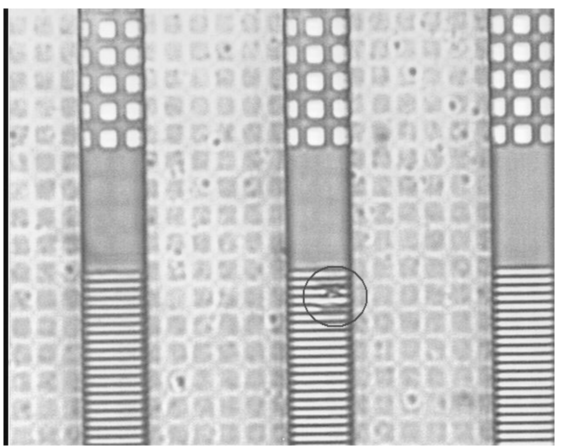


Figure 16. Backside optical image

Figures 17 shows SEM images of the device after RIE etching of the low k ILD to expose the copper metal runners. The anomalies were observed under SEM analysis and were believed to be caused by thermal or electrical stress induced damage.

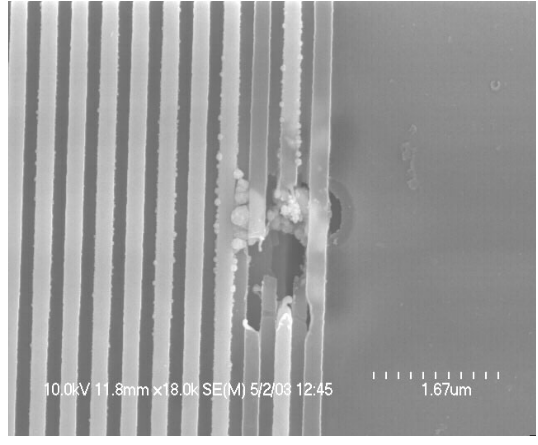


Figure 17. SEM image of M5 runners at defect site

Backside deprocessing of SOI devices has a certain advantage, which is that the BOX (buried oxide) layer acts as a chemical etch stop when etching bulk silicon. This leaves the transistor active silicon intact for analysis. Further delayering can be performed for the inspection of the active silicon, gate oxide, salicide, spacer and poly. Buried oxide can be removed by diluted HF wet chemical or CF_4 based RIE [17].

Summary

Various FA challenges, reliability issues, and new failure modes for copper technology have been addressed. Several front-side and backside deprocessing techniques for copper, low k and SOI devices have been discussed including reactive ion etching (RIE), parallel polishing, chemical mechanical polishing (CMP) and combinations of these techniques. Moreover, gate level deprocessing techniques have also been presented.

Experience has shown that a combination of CMP of copper interconnects along with RIE of low k dielectric works well as a deprocessing technique. In addition, it has been found that wet chemical etching can be a successful way to achieve a uniform and clean surface when etching thick, wide copper buses. For Circuit-Under-Pad (CUP) structures, flip-chip BGA packages or dies with a thick, dense top layer and multiple metal layers, backside FA may be necessary to obtain backside defect localization followed by backside physical analysis. As with all failure analysis deprocessing, the FA techniques chosen depend on the circuit structures, materials, defect types and suspected defect location. It should be noted that the FA recipes mentioned in this paper are dependent on the equipment, materials, and circuit structures under investigation and may need to be modified to some extent to work in your situation. Nevertheless, they can be used as a starting point for your own deprocessing technique optimization

Reference:

[1] Lawrence C. Wagner, “Failure analysis challenges”, pp. 36-41, 2001 IPFA

[2] Huixian Wu, James Cargo, Carl Peridier and Joe Serpiello, “Reliability issues and advanced failure analysis deprocessing techniques for copper/low k technology”, pp. 536-544, 2003 IRPS

[3] Huixian Wu and James Cargo, “Characterization of reactive ion etching of silicon substrate for backside failure mode analysis”, ISTFA 2002

[4] T. Sakura, H. Utsunumyia and et al., “A detailed study of soft- and pre-soft-breakdowns in small geometry MOS structures,” pp. 183-186, IEDM Tech. Dig., 1998

[5] M.H. Tsai, S.W. Chou, and et al., "CMP-free and CMP-less approaches for multilevel Cu/low-k BEOL integration", pp. 80-83, 2001 IEDM
[6] M. Markert, A. Bertz and T. Gessner, "Mechanism studies of Cu RIE for VLSI Interconnections", pp. 155-157, European Workshop on Materials for Advanced Metallization, 1997
[7] Partrice Koch, Yan Ye and et al., "Development of copper etch technology for advanced copper interconnects", pp. 290-294, 1999 IEE/SEMI Advanced Semiconductor Manufacturing Conference
[8] A. Crockett and W. Vanderlin, "Plasma delayering of integrated circuits", pp. 243-252, EDFAS Desk Reference, 2003
[9] Trion Technical Paper – RIE Grass
[10] Huixian Wu, James Cargo and et al., "Characterization of reactive ion etching of poly silicon over gate oxide for failure mode analysis deprocessing", pp. 91-96, Proceeding of 9th IPFA, 2002
[11] Valentina Korchnoy, "Investigation of choline hydroxide for selective silicon etch from a gate oxide failure analysis standpoint", pp.325-331, ISTFA 2002
[12] Y.N. Hua, G.B. Ang, and et al., "Studies on chemical methods to expose gate/tunnel oxide and identification of gate/tunnel oxide defects in wafer fabrication", pp. 695-699, ISTFA 2002
[13] Huixian Wu, James Cargo, and et al., "Interconnect and gate level delayering techniques for Cu/Low k technology failure analysis", pp.90-98, ISTFA 2003
[14] Yao Yahong, Zhao Yongjun, "Anisotropic trench etching of Si using SF6/O2 mixture", pp. 61-66, International Symposium on Micro-Mechanics and Human Science, 1997
[15] Junji, Ohara, et al., "Improvement of Si/SiO2 mask etching selectivity in the new D-RIE process", 2001 IEEE
[16] Kow-Ming Chang, Keiji Horioka and et al., "Highly selective etching of poly silicon and etch-induced damage to gate oxide with halogen-bearing electro-cyclotron-resonance plasma", J. Appl. Phys. 80 (5), pp. 3048-3055, September 1996
[17] S. Prejean, J. Shannon, and et al., "Backside deprocessing of CMOS SOI devices for physical defect and failure analysis", pp.99-104, ISTFA 2003
[18] Mansun Chan, Selina S. Yuen and et al., " Comparision of ESD protection capability of SOI and bulk CMOS output buffers", pp. 292-298, 1994 IRPS
[19] Kangguo Cheng, JinJu Lee, and et al., "Improved hot-carrier reliability of SOI transistors by deuterium passivation of defects at oxide/silicon interfaces", pp. 529-531, IEEE Transaction on Electron Devices, Vol. 49, No. 3, March 2002
[20] T. Poiroux, et al., IEDM Tech. Digest, 1999

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Optical Microscopy

John McDonald
Quantum Focus Instruments Corporation, Vista, California, USA

Introduction

Moore's Law has driven many circuit features below the resolving capability of optical microscopy. Yet the optical microscope remains a valuable tool in failure analysis. It is easy and gratifying to quickly examine a device under a nearby microscope, but effective use of the microscope requires some mastery of concepts, which we describe here. This author consistently hears requests from semiconductor engineers for resolving finer details, most frequently and incorrectly expressed as a request for more magnification. We will describe the physics governing resolution and we will describe useful techniques for extracting the small details. We will start with the basic microscope column and construction. These basic construction elements are hidden away but still fundamental to the most sophisticated million dollar instruments. We will discuss microscope adjustments, brightfield and darkfield illumination, and microscope concepts important to liquid crystal techniques. We will discuss solid immersion lenses, infrared and ultraviolet microscopy and finish with laser microscopy techniques such as TIVA and XIVA.

Figure 1: Typical Lab Microscope showing epi-illumination path.

Some words about light

The term Optical Microscopy is commonly limited to light that can be manipulated and focused with lenses, i.e., the visible light spectrum, plus the infrared and ultraviolet. We will not address non-photonic microscopy, e.g., electron microscopes, focused ion beams, ultrasonic or atomic force microscopes, none of which use photonic light for image formation. All light propagates through vacuum at approximately $3x10^8$ meters/second and nearly the same speed in air. Light travels more slowly in other materials such as glass or water. Light travels approximately 50% more slowly in glass than in air.

If the air/glass interface is bent or curved a plane wave will strike and slow on one side first and thereby itself be bent and curved, turned and focused, giving rise to the optics industry and microscopy. The ratio of the speed of light in vacuum to the speed in some other medium is important to optics and is known as the index of refraction, given by:

$$\text{Index of Refraction N} = \frac{\text{Speed of Light in Vacuum}}{\text{Speed of Light in Medium}}$$

Equation 1:

Some commonly used wavelengths of light as follows:

Ultraviolet		0.2 – 0.4 microns
Visible	Violet	0.4 microns
	Blue	0.45 microns
	Green	0.5 microns
	Yellow	0.56 microns
	Red	0.6 microns
Infrared	Near Infrared	0.75~2 microns
	Short Wave Infrared	~1~3 microns
	Mid Wave (Thermal) Infrared	3 - 5 microns
	Long Wave (Thermal) Infrared	8 – 12 microns

The energy in a photon is related to the wavelength λ is:

$$E = \frac{hc}{\lambda}$$

where:

E is in Joules

$c = 3x10^8$ m/s (speed of light in vacuum)

$h = 6.626x10^{-34}$ joule I s (Planck's Constant)

The Microscope Column

A simple compound microscope consists of an objective lens and an eyepiece. The objective creates an image of the sample in the space between the eyepiece and objective. The eyepiece acts exactly like a simple hand lens magnifier, to magnify the intermediate image for the eye. A well designed eyepiece makes the image appear as though coming from an infinite distance so the eye may work in a relaxed state. The lens in the user's own eye performs the final focus of the image on the retina. If you remove the eyepiece from a common microscope and invert it so you look through the bottom, it makes a fine hand lens for quickly examining an object.

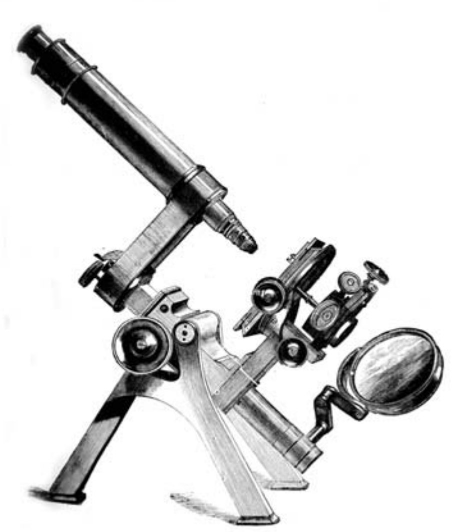


Figure 2: A simple 1920's microscope showing eyepiece, tube, objective, sample stage and illumination source. From Spitta, 3rd edition [1].

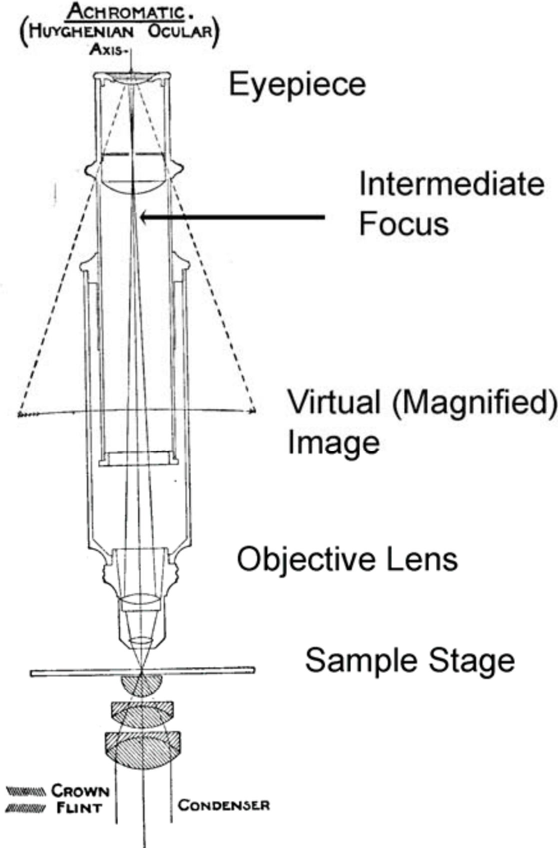
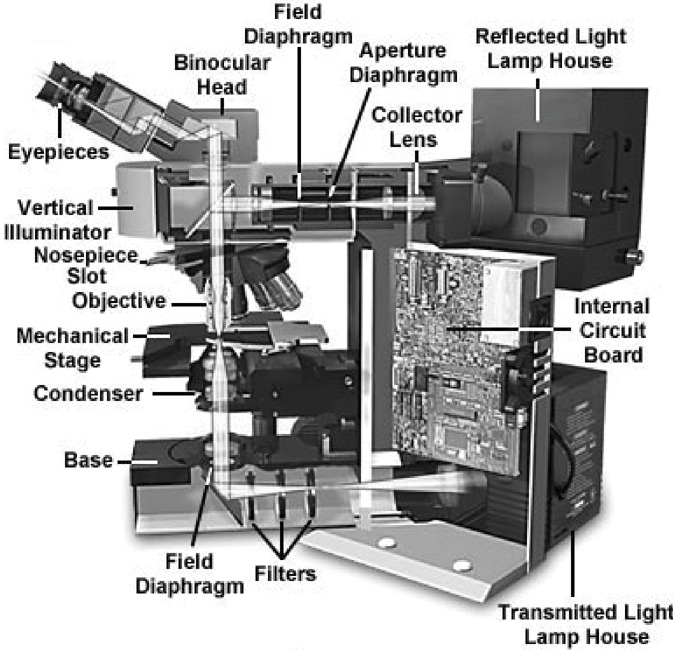


Figure 3: Schematic view of a simple microscope column showing the intermediate focus. In today's microscopes the intermediate focus occurs just inside the bottom of the eyepiece. The image can be directly examined in a dark room by removing the eyepiece and inserting a bit of lens tissue as a screen, into the eyepiece socket. From Spitta, 3rd edition [op. cit.].

A practical microscope requires an illuminator. The illuminator illustrated in figure 1 is an epi-illuminator, epi meaning 'above' in Latin; so it illuminates from above the sample. In a biological microscope the illuminator typically transilluminates the glass sample slide from below. The microscope in figure 1 is also equipped for transillumination as are the microscopes in figure 4. In figure 1 the epi-illumination is coupled into the objective with a beam splitter. A beam splitter is a plane of glass that reflects some fraction of incident light, and passes the rest. Some given portion of the illumination light is reflected into the objective, and some of that will strike and be reflected by the sample. When it returns, a given fraction will pass through the beam splitter to the eyepieces. In figure 1 there is an additional beam splitter in the trinocular head to divert the some of the image light from the eyepieces to the camera.

Microscope Parts and Nomenclature



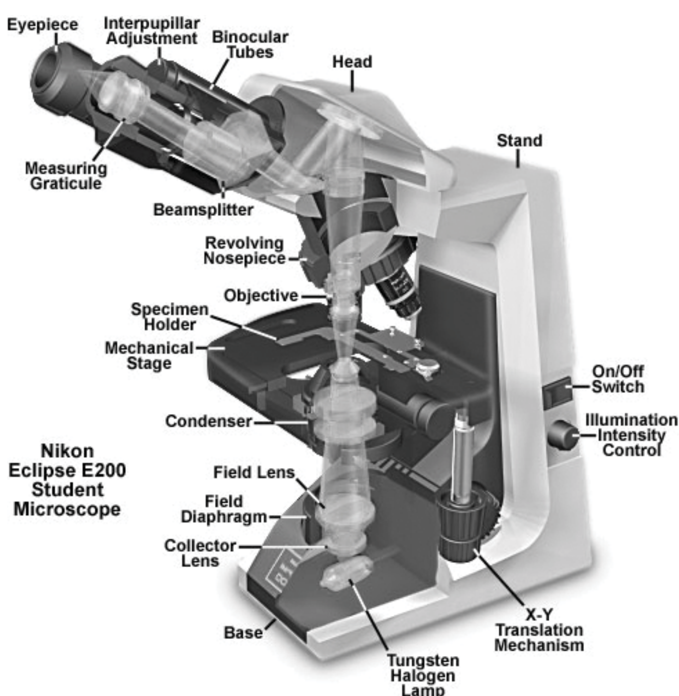


Figure 4: Cutaway views of microscope showing nomenclature. adapted by permission from www.molecularexpressions.com.

Stereo Microscopes

One effective form of depth perception comes from having each eye view from a slightly different angle causing image shift or parallax to each eye. The stereomicroscope creates parallax by actually offering two microscopes in a common housing, one for each eye, each seeing the sample from a different angle. For compactness the left and right eye microscope often share portions of the final lens as shown in the right design in figure 5.

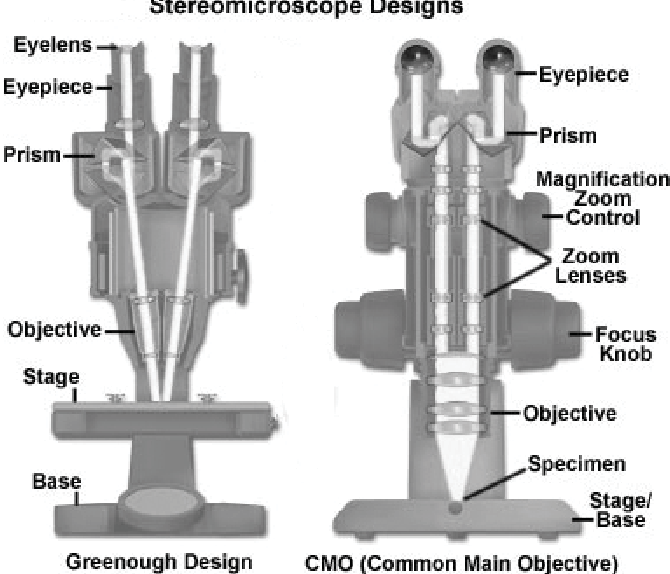


Figure 5: Two stereo microscope designs with erecting prisms. In the CMO design each eye has a distinct microscope path but they share portions of the objective. Adapted by permission from www.molecularexpressions.com.

Today most high magnification microscopes such as those in figure 4 offer two eyepieces, but not for depth perception, rather for the viewing comfort of the user. Many things cue the brain to see depth, but true stereomicroscopes offer depth perception by having separated optical paths for each eye. Nonetheless there are some situations possible with conventional microscopes like those in figure 4 to see true parallax depth perception. The situations arise when the user has improperly adjusted the interpupillary distance between binocular eyepieces. Then each eye will see a slightly different view and the brain will detect some depth perception.

Microscope Objectives

Expensive high magnification objectives have a great many lenses contained within to perfect the image. For any given manufacturer all objectives are deliberately designed so the back focus will fall at a given distance, usually 160 mm to 200 mm. This is done so every lens will remain at the correct distance from the camera and eyepieces when the nosepiece is rotated. Substituting a different manufacturer's lens into a microscope may not work well if the back focal lengths are different. The back focal length is often simply known as the tube length, for the length of tube required to separate the objective from the eyepiece.

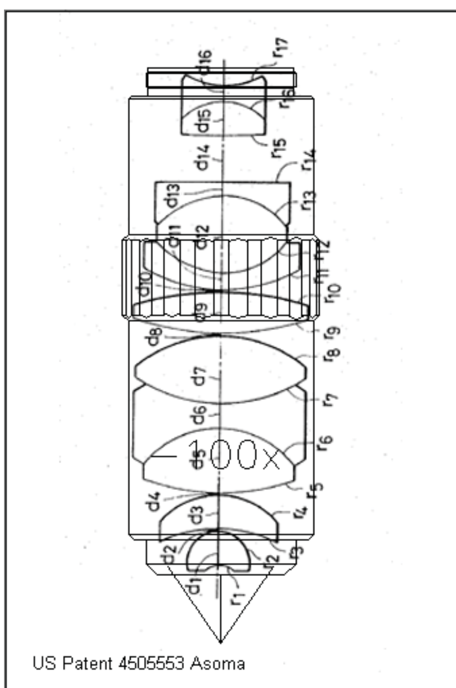


Figure 6: The complexity of a lens may be seen from this 100x Objective from a US Patent roughly superimposed within a common lens housing. [2]

Infinity-Corrected Objective System

Eyepiece Diaphragm Plane
Eyepiece
Intermediate Image
Tube Lens
"Infinity Space"
Parallel Light Beam
Objective
Front Focal Plane of Objective
Specimen
by permission: www.molecularexpressions.com

Figure 7: This schematic shows how an infinity corrected lens works. Note that on the left, the light from any specific point of the field (sample) is parallel in the infinity space. Accessories such as beam splitters are inserted into the infinity space area.

Newer microscopes are equipped with infinity corrected objectives. These objectives project the light in collimated or parallel rays, as though they are coming from infinitely far away. The final focusing is left to a separate lens called a tube lens. A single tube lens can serve all of the objectives in the nosepiece and is often embedded in the head that houses the eyepieces. The collimated or parallel ray space between the infinity lens and the tube lens is optically ideal for inserting options such as beam splitters or filters. Infinity corrected objectives perform better and more flexibly than direct focusing objectives. Infinity corrected objectives must be paired with the correct tube lens. They will not have the correct magnification (stamped on the lens) if they are paired with a tube lens of incorrect focal length. Infinity corrected lenses cannot be interchanged with fixed focal length lenses.

Key Concept: Numerical Aperture

Numerical Aperture (N.A.) is the single most important figure of merit in microscopy. The numerical aperture or N.A. of a lens governs limiting resolution, image brightness, depth of focus, working distance and ultimately the price of the lens; lenses with higher N.A.'s are more expensive. It is a common misconception among microscope users that with a little more magnification they can resolve finer features. It isn't true; magnification has a secondary effect on resolution but numerical aperture determines the limit to resolution.

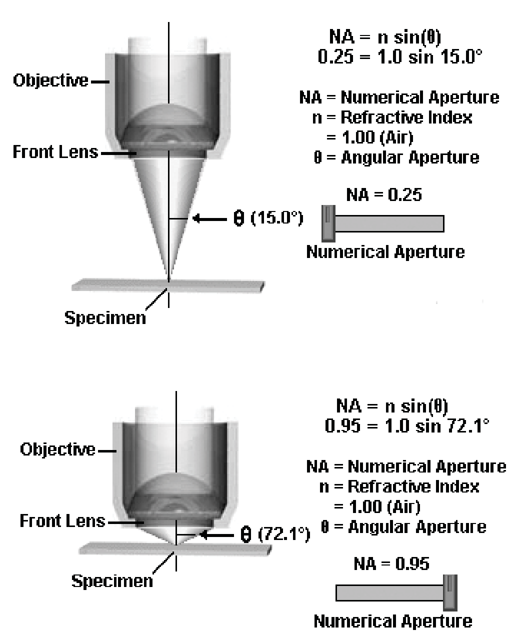
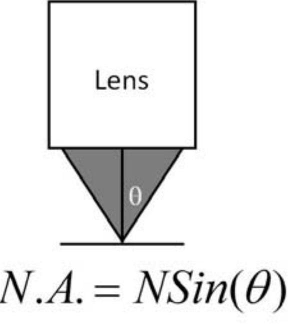


Figure 8: A high N.A. has a short working distance and wide cone angle. A low NA has a long working distance and narrow cone angle. High N.A.'s offer the best resolution.

Most high quality objectives have their numerical aperture value stamped on the side of the lens. N.A. can be thought of as the collection cone of the objective. Large collection angles correlate to large N.A. values, finest resolution and greatest light and signal collection. N.A. is mathematically defined as:

Numerical Aperture is Fundamental to Microscopy

- Numerical Aperture (NA) is the sine of the collection half angle, multiplied by the index of the intervening medium.



$$N.A. = NSin(\theta)$$

3/29/2010 ISTFA 2009 2

Equation 2: Numerical Aperture where:
N = index of refraction (between the lens and the sample)
θ = the half angle of the acceptance cone of the lens

N.A. approaches a limit as the collecting half angle approaches 90º. In air where the refractive index is 1, the maximum N.A. is limited to 1 and in practice rarely exceeds 0.95. A lens with N.A.=0.95 will have an exceedingly short working distance, often impractical for failure analysis. Some lenses are designed for immersion in water or oil and in these cases the N.A. becomes 1.2 or 1.4 respectively.

Well made objectives of 50x magnification or greater are usually designed to be diffraction limited, meaning that the wavelength of the light and the collecting cone of the objective lens govern resolution. How does the diffraction limit work? When light from a point source passes through a circular aperture and is then focused, it spreads with a characteristic pattern as seen in figure 9, due to light bending around the edges of the aperture.

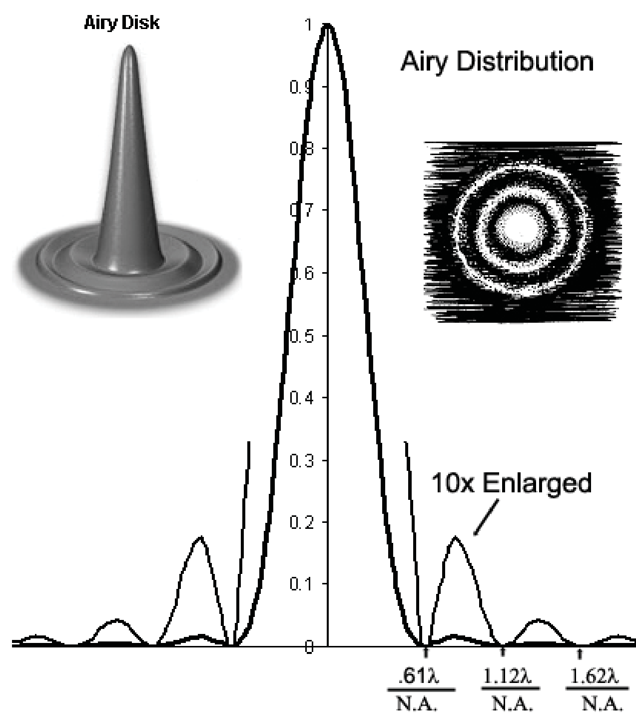


Figure 9: Illumination from a point source spreads at the focus due to diffraction. The spread is least for large N.A. This ultimately limits resolution. On film a well exposed Airy Disk would appear as in the upper right.

Diffraction is derivable from the Heisenberg Uncertainty principle (beyond our scope here) but suffice it to say that the smaller the aperture, the larger the diffraction spread and hence the larger the Airy disk. This means that a point source will always be imaged larger than a point, so diffraction effects ultimately limit the sharpness of an image. A large N.A. minimizes diffraction effects, so it pays to buy good lenses with large N.A.'s. The diameter of the Airy disk is proportional to the wavelength, and inversely proportional to the NA.

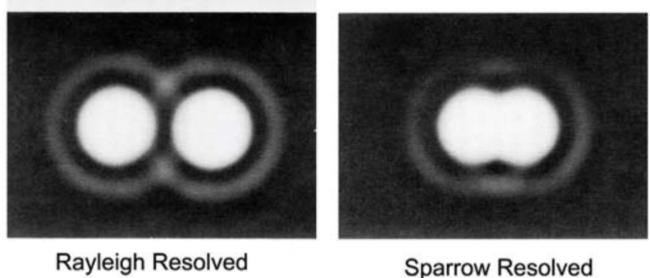


Figure 10: Point sources just resolved by Rayleigh and Sparrow Criteria.

A simple, effective measure of resolution is the Sparrow Criteria, given as the just resolvable separation between two point sources as shown in figure 10. The relation between the Sparrow Criteria and the somewhat better known Rayleigh Criteria is also shown.

Rayleigh Criteria $$D = \frac{0.61\lambda}{N.A.}$$

Sparrow Criteria $$D = \frac{0.5\lambda}{N.A.}$$

Equation 3: Rayleigh & Sparrow Criteria give minimum separation of point sources to be resolved.

The Sparrow Criteria is defined as D=0.5*lambda/N.A., where D is the distance between the points, lambda is the wavelength and N.A. is the numerical aperture. Mid-visible, green light, is about ½ micron wavelength and for an air immersed lens the NA is limited to 1, so using the Sparrow Criteria, typical FA lab microscopes working in air must be limited to, at best, about ¼ micron resolution. More commonly in FA microscopes, the N.A. is traded for increased working distance so the user can place micro-probes between the lens and the circuit. In that case the N.A. is more commonly limited to about 0.5 and the Sparrow resolution is about ½ micron. There are many merit figures for resolution, but it is substantially correct to say that the Sparrow Resolution is the maximum resolution of a lens. Note that magnification is NOT the limiting factor in resolution.

Role of Magnification on Resolution

Magnification doesn't limit resolution, at least not directly, numerical aperture does. But we intuitively know that if we increase magnification we can see smaller details. Magnification is actually about matching the resolution element to the detector, whether eye or camera. Imagine an objective with N.A. appropriate for our resolution task but with no magnification (1x). Then neither our eye nor a camera will be able to resolve the fine detail. The human eye has a resolution limit, often given at 1 minute of arc depending on age and individuals. That corresponds to about 76 microns at a comfortable viewing distance of 250 mm. Likewise cameras have a resolution limit imposed by distance between pixels, known as the pixel pitch, ranging from around 6 microns to 50 microns. The job of magnification is to make the image large enough so the resolved element can be seen by at least 2 or 3 detector elements, be they camera pixels or cones in the human eye.

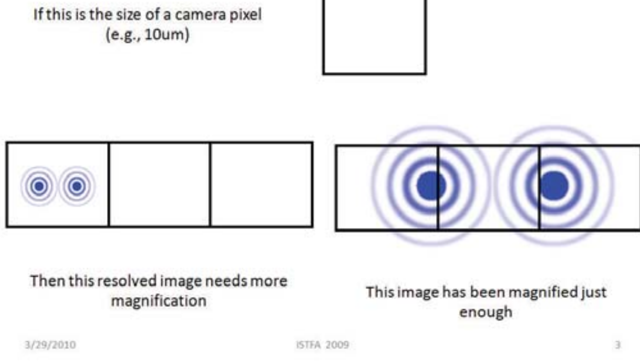


Figure 11: Point sources just resolved by Rayleigh and Sparrow Criteria.

For example, consider a pair of objects resolved per the Sparrow criteria at ¼ micron separation. The job of magnification is to enlarge that just resolved dimension so that it fills two eye resolution elements.

$$M = \frac{eye_resolution}{Sparrow_resolution}$$

Where M is the required magnification. For our example:

$$M = \frac{2\ (76\ \text{microns})}{0.25\ \text{microns}} \sim 600x$$

In this example we need about 600x magnification to just resolve the objects with our eye. 600x is readily achieved with a 60x objective and a 10x eyepiece. It is common to use a 100x objective and 10x eyepiece for a total magnification of 1000x to make the job of the eye a little easier, particularly for those of us who are older. If the eyepiece is removed and replaced with a camera with a 12-micron pixel pitch, the required magnification is:

$$M = \frac{2\ (12\ \text{microns})}{0.25\ \text{microns}} \sim 100x$$

There is a minimum magnification that is appropriate to the N.A. of the objective. Any magnification above the minimum is known as ‘empty magnification’ because it doesn’t improve the resolution of the instrument, it renders the blurred detail a little larger, but not less blurred. It is customary to exceed this magnification a bit to make the job of the user easier, but it is pointless and we will show that it is deleterious to exceed the required magnification by too much.

Magnification & Pixel Pitch for UV or IR Applications
The Sparrow Criteria may be used effectively to choose appropriate magnifications and camera pixel pitch for ultraviolet and infrared applications. For example, consider a hypothetical air immersed microscope objective working at 200 nm in the deep ultraviolet. Assuming the numerical aperture is at the theoretical limit of N.A.=1, and the desired camera has 6 micron pixel spacing, what should the objective magnification be? From the Sparrow Criteria:

$$D = \frac{0.5\ (.200\ \text{microns})}{\text{NA:1}} = 0.100\ \text{microns}$$

Sparrow Criteria for Deep UV example

and using this result to compute the required magnification:

$$M = \frac{2\ (6\ \text{microns})}{0.100\ \text{microns}} \sim 120x$$

For this deep UV example it can be seen that a 100x lens just will not be enough.

For an infrared example, the author has used a thermal infrared microscope working at 3 microns wavelength to measure junction temperatures. The infrared detector pixels were spaced 24 microns apart and the objective had an N.A. of 0.5 to allow ample working distance to microprobe the circuits. The Sparrow Resolution in this case will be [0.5*3 microns / 0.5 = 3 microns]. So for this infrared example the limiting resolution is 3 microns. Applying the magnification formula, [2*24microns / 3 microns = 16x]. We can see from these two examples that microscope magnification must be increased for short UV wavelengths, but may be decreased for longer infrared wavelengths.

Downside to Excess Magnification
Most of us have noticed that microscopes used at high magnification must be protected from vibration or shock. Any lateral vibration that is present will also be magnified by the objective so that activity on the lab bench supporting the microscope can appear as an earthquake in the eyepieces. If the user increases the magnification beyond the useful resolution of the objective, the image will not become any clearer but any ambient shock or vibration will be magnified perfectly, possibly rendering the instrument useless. A very common example is when a microscope is coupled to a functional test head. Without mitigation, the fans in the test head will badly blur images from the high magnification lenses. In a later section we will also show that the throughput or signal at the camera plane is inversely proportional to the magnification squared. That is why the user must commonly increase the illumination when they switch to higher magnifications. A 100x objective tends to just match or exceed the capabilities of current tungsten halogen bulbs to illuminate the specimen. So loss of illumination is yet another penalty for excess magnification.

Field and Aperture Planes
A field plane is a location anywhere in the microscope where a focused image is located, whether virtual or real. The specimen plane, the camera focal plane and the retina of the eye are all field planes. There is also a field plane inside the tube at the focus of the eyepiece. There are additional field planes in the illumination path. These field planes are copies of one another, excepting in the illumination path, and are said to be joined or “conjugates”. Some field planes are special, e.g., the eyepiece field plane is an excellent location for a cross hair reticle, since it will be in focus and superimposed on the sample image. One of the field planes in the illuminator often has a ‘field diaphragm’ so the user can adjust the size of the illuminated spot to match the field of the lens (to reduce stray light and glare). When a field Diaphragm is adjusted the user can see the field of view shrink or grow, usually with the edges of the diaphragm vanes clearly visible.

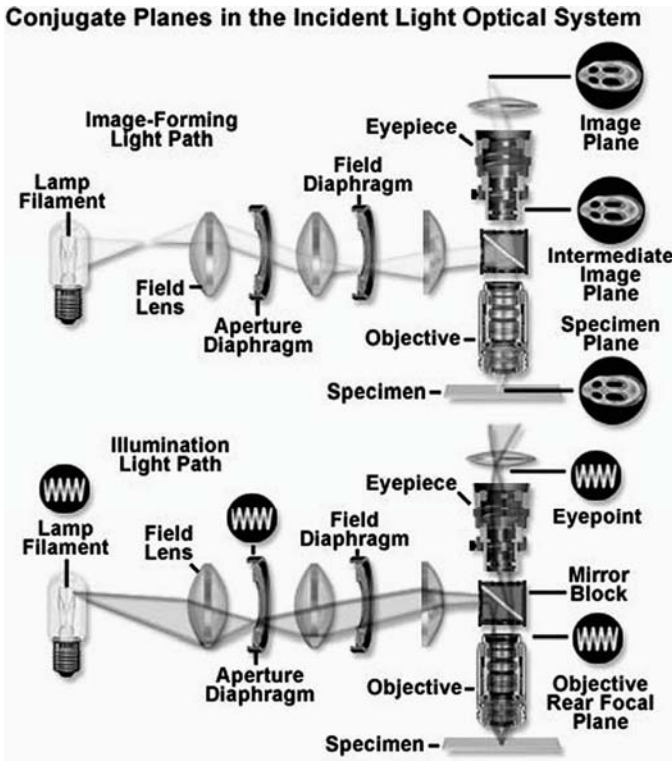


Figure 12: Locations of Field (image) and Aperture Planes in an epi-illuminated microscope are designated with images of a specimen or lamp filament respectively. Used by permission from www.molecularexpressions.com.

There is another set of conjugate planes in a microscope known as aperture planes. The aperture planes limit the cone angle of light and occur at the iris of the eye, the back of the objective, the illuminator diaphragm and often at the lamp filament. The illuminator filament doesn't affect the cone angle but it is deliberately placed at an aperture plane to facilitate Koehler Illumination. Any object located at a field plane is imaged in sharp focus at the other field planes. Less obvious but likewise true, any object located at an aperture plane is imaged in sharp focus at the other aperture planes. Objects in focus in either the field or aperture set of planes are perfectly defocused, actually collimated, at the other set of planes. To illustrate, if you could somehow place a bit of paper at the back of the objective you would see a sharp image of the bulb filament. (If the microscope is the type with a ground glass diffuser, then an image of the diffuser will be there instead.) There is another sharp image of the lamp filament just at the iris of the user's eye but the filament is perfectly diffuse and unfocused at the retina. Likewise in the set of conjugate field planes there is a sharp image of the retina of your eye formed at focus of the eyepiece and at the specimen. In the design and use of microscopes these conjugate planes come into play in various surprising ways. For example there is a Fourier Transform of the specimen at the back of the objective and unusual and powerful filtering tricks can be applied there in microscopes that allow access to that point. In the next section on Koehler Illumination, we show how to make the filament perfectly diffused and uniform at the specimen plane by deliberately focusing it in the aperture plane.

Illumination – Koehler Illumination

In 1893 Professor August Koehler published his brilliant technique for microscope illumination, now almost universally fielded on top quality microscopes. Koehler's technique requires that the filament be focused perfectly in the aperture planes, so that it will be perfectly collimated and therefore very diffuse and very uniform at the specimen plane. There are several adjustments on a top quality microscope for setting up Koehler illumination. In a multi-user environment these are often badly adjusted. Poor Koehler adjustment leads to reduced resolution, uneven illumination and diffraction artifacts, it is worth learning how to adjust the illumination. For an epi-illuminated system, the most important aperture adjustment is the field diaphragm located closest to the microscope body. When first adjusting the microscope and after each lens change, the user should adjust the field diaphragm so the vanes just barely disappear from the field of view. If the field diaphragm is opened too far light spills uselessly inside the microscope tube and creates a veiling glare that reduces contrast and makes viewing difficult.

The aperture diaphragm (if equipped) controls the cone angle of the illumination light. It should typically be left wide open for epi-illumination. In principle, restricting the aperture diaphragm can lead to reduced resolution but the highly reflective surfaces of circuits tend to mitigate the effect. Nevertheless, it is prudent to leave the aperture diaphragm wide open.

In biological microscopes the illumination usually comes from below, to transilluminate the glass specimen slide. This requires a separate condenser below the sample stage that must be adjusted up and down on a stage to suit each objective lens. A good (expensive) substage condenser will also have a field and aperture diaphragm. For transillumination the adjustment of the aperture diaphragm is critical for achieving high performance. It should be opened wide for the high N.A. lenses (the high magnification lenses), to match the illumination cone to the lens N.A. cone. Failure to do so directly sacrifices the resolving capability of the high N.A. lens. This is easily demonstrated with biological microscopes and in theory it is true as well in our failure analysis labs for epi-illuminated microscopes. However in practice the author finds that scattering due to reflection from epi-illuminated samples often seems to compensate for badly adjusted (or missing) aperture diaphragms. As with the field diaphragm, an opening too large may cause excess light to spill and reflect about in the microscope tube. This creates glare and reduces contrast. Some microscopes permit (and therefore require) adjustment of the filament position. If there is no ground glass an image of the filament can be seen by removing the eyepiece and looking down into the barrel. This is difficult without a special tool called a Bertrand Lens. The image will be too bright and too small to see clearly without the Bertrand Lens. Nevertheless the filament should be centered and focused in the tube.

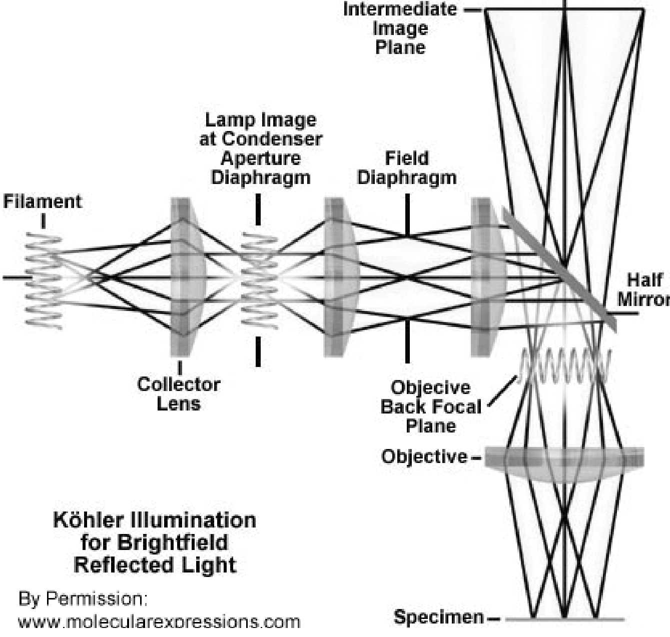


Figure 13: Adjust the field diaphragm at each change of magnification so the vanes just disappear outside the illuminated field. Adjust the aperture diaphragm (if equipped) wider for high N.A lenses, and narrower for low N.A. lenses to optimize resolution and contrast.

Adjusting Binocular Eyepieces

Binocular eyepieces commonly provide two adjustments, and remarkably, casual users have difficulty with these. The first of these adjusts the interpupillary distance. Typically the user

need only push or pull the eyepieces closer together or further apart until they are comfortably spaced. On some poorly designed microscopes this adjustment will change the distance from eyepiece to the objective, requiring some refocus.

A binocular head also provides a separate focus adjustment for each eye, since few of us have the same requirement for each eye. The adjustment is usually a ring surrounding the eyepiece that rotates, slightly raising or lowering that specific lens. In some cases only one lens is adjustable, for which case the eye without adjustment uses the main focus adjustment on the body of the microscope. To adjust, close the eye over the adjustable eyepiece and using the other eye, focus the microscope normally with the general focus knob. Then close this eye, and open the one over the adjusting ring. Then with only the adjusting ring, focus for the remaining eye. If both eyepieces are adjustable adjust one eyepiece midway between the stops, to the mid-adjustment position and then treat it as though it has a fixed focus, performing the adjustments with only the other eyepiece.

Lens Aberrations / Correction for Backside Imaging through Si

The Airy disk and Sparrow criteria indicate the diffraction limit to resolution, based on the behavior of light when constrained by a finite aperture. Well-designed and typically expensive lenses will preserve the diffraction limit. Lesser quality microscopes may exhibit some of the following typical aberrations which will degrade the resolution to less than the diffraction limit.

Field Flatness

A simple lens naturally focuses an image to a curved surface and a correction must be applied to make the image suitably flat for CCDs or film. Inexpensive objectives are not corrected for field flatness and should be avoided for failure analysis. If the user notices that she can focus the center, or the edges of the field, but not both at the same time, that lens lacks field flatness. Better objectives with field flattening are usually designated so with the prefix ‘plan’, e.g., ‘*plan apochromat’*.

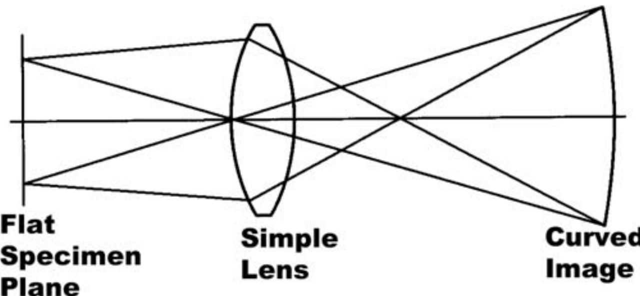


Figure 14: A simple lens naturally focuses light to a curved surface. If the detector is flat, e.g., a camera, this results in poor focus at the corners and edges of the image. A ‘Plan’ type objective is corrected to project a flat image.

Chromatic Aberration

The index of refraction of glass varies slightly with the wavelength. That is why the classic equilateral prism can disperse light into a rainbow; each wavelength is bent by a differing degree. That creates a natural problem for imaging optics. Chromatic aberration occurs because a simple lens is unable to bring light of all colors to a common focus.

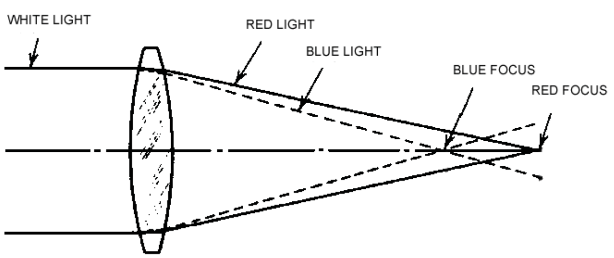


Figure 15: Chromatic Aberration blurs the image by causing light of different wavelength or color to focus at different locations.

Fortunately it can be corrected by a careful optical design, by pairing lenses with slightly differing color dispersion. A greater degree of correction usually costs more, so the industry has created a labeling system, often stamped on the lenses, to identify the quality of correction. The most common designations are listed in the nearby table. The author suggests that for the exacting resolutions required for failure analysis work, the user should always insist on the plan-apochromat type lenses.

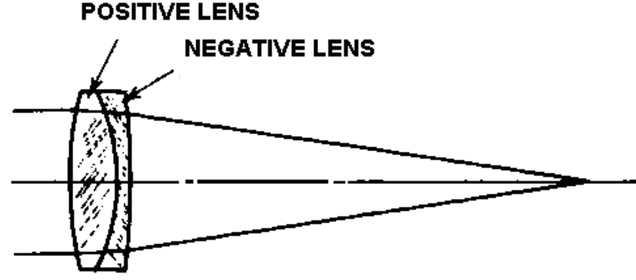


Figure 16: Simple achromat doublet corrects chromatic aberration by pairing lens powers and shapes so that each cancels the aberration of the other. Similar techniques are used to correct all aberrations.

Common Microscope Lens Descriptors	
Achromat	Least expensive and common in dental and physician offices and modest labs, they are corrected for 2 wavelengths, red and blue, slightly out of focus for green and other colors.
Fluorite or semi-apochromat	Corrected for 3 wavelengths, sometimes corrected for flatness.
Apochromat	Corrected for 4 or more wavelengths, effectively eliminating chromatic aberration.
Plan Apochromat	Corrected for 4 or more wavelengths and corrected for field flatness. This is an expensive, superior lens and the only type that should be employed for high-resolution failure analysis.

Coma, Astigmatism & Distortion
Coma is an unwanted variation of magnification related to where in the aperture the light enters. It causes a point source to look like a comet with a pointed tail. Astigmatism occurs when a lens has more focusing power in one direction, e.g., the horizontal, than another. Astigmatism correction often involves adding the equivalent of a cylindrical lens, which has focusing power in one direction only. Distortion is seen as pin-cushioning or barrel distortion where objects become misshapen at different parts of the field of view. Elimination of these aberrations and others require extra optical design effort and additional lenses in the interior of the objective, and additional manufacturing care. Higher quality, more expensive lenses make sharper images for a good reason.

Spherical Aberration & Backside Imaging through Silicon
Spherical aberration has special interest for Failure Analysis because it affects imaging through the backside of silicon. Spherical aberration causes energy from a point source to spread into a surrounding halo. Spherical aberration is from rays entering the edges of the aperture, which focus at a greater distance than rays entering the central part of the aperture.

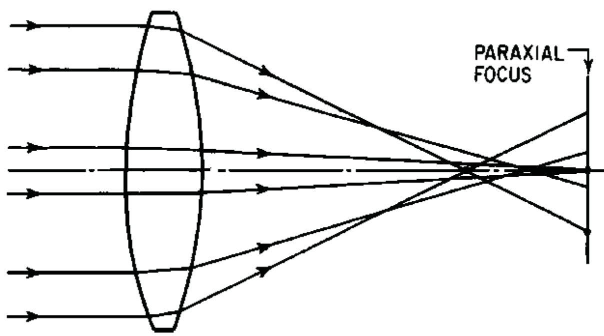


Figure 17: Sperical aberation blurs the image by causing rays entering the lens off center to focus away from the paraxial focus..

A skilled optical designer can correct spherical aberration provided he controls all the refractive media. However when inspecting circuits from the backside of the wafer, the intervening silicon becomes part of the optical path. The silicon introduces spherical aberration whose severity depends on the thickness of the silicon and the numerical aperture of the examining lens. The aberration can be corrected provided the silicon thickness is known, or sometimes the correction is user adjustable. Biological microscopes have long faced this issue since the biologist places a cover-slip of glass on his slide to keep his specimen from drying out. High N.A. biological objectives are often designed with a correcting ring. The user can dial in the required compensation for the thickness of the cover-slip. If the failure analyst is performing a backside inspection with a high numerical aperture objective, the image will be degraded unless the lens is corrected for the intervening silicon.

Immersion and Solid Immersion Lenses (SILs)
Equation 2 shows that the numerical aperture increases directly with the index of the medium between the lens and the sample. Since the best available resolution depends on numerical aperture we can improve resolution considerably by employing an objective that is designed to be immersed in a high index medium. Biologists have long used water or oil immersion lenses to get a 20% or 40% improvement in resolution, respectively.

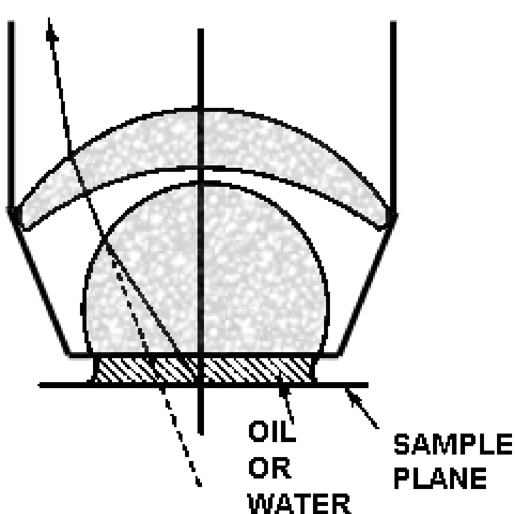


Figure 18: Specially designed fluid immersion lenses improve resolution by up to 40% over air immersed lens. The index of the immersion medium must match that of the 1st lens.

Immersing a circuit in oil or water is usually undesirable and commercial oil or water immersion lenses are poorly corrected for looking through silicon; the author's experiments with them have been disappointing. However it is possible to make direct contact to a silicon substrate with a near infrared silicon lens; that is, the immersion is in a solid material. This type of immersion lens is called a solid immersion lens and often abbreviated to SIL.

Solid immersion lenses have long been used for making certain types of detectors; Warren Smith described the technique in 1966 in Modern Optical Engineering [4]. Now they are being applied effectively to failure analysis.

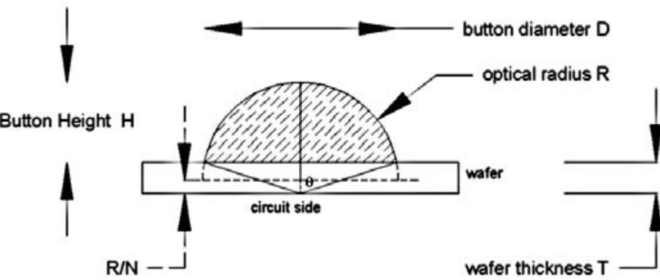


Figure 19: Solid Immersion Lens (SIL) in optical contact with the wafer gives a huge benefit in resolution for backside inspection. Note the hemisphere continues virtually, but not actually into the wafer.

Since the index of silicon is 3.5, the minimum resolution from the Sparrow Criteria (eq. 3) is

$$D = \frac{(0.5)(1.0)}{3.5}$$

$$SIL_resolution = 143nm$$

where we use 1 micron, the shortest wavelength that passes through silicon, and we assume that the NA half angle is operating at 90 degrees, the theoretical maximum. In this case the Sparrow Criteria yields 143 nanometers minimum resolution. The high refractive index of the silicon pays off, yielding resolution considerably better than that available from air immersed lenses viewing from the front side. More realistically, we should plan on a maximum practical half angle of 72 degrees instead of 90 degrees, and a wavelength minimum of 1.1 microns. This still results in a Sparrow Criteria resolution of 165 nanometers, which is very respectable.

For the SIL technique to work, both the contacting lens and the wafer must be smooth and flat with no air gaps. Any air gaps of wavelength dimension will degrade the resolution from the theoretical limit. Few papers, perhaps none, have reported achievement of theoretical resolution for contacting SIL lenses. But nevertheless they have reported excellent resolution.

Several types of SILS have been described for FA work. The most common implementation is a partial hemisphere made to contact the circuit substrate as in figure 19, except with considerable supporting hardware. But in 2003 Koyama, et. al. described a clever implementation where they machined the hemisphere directly into the substrate and in 2005 Zachariasse et. al. etched a Fresnel lens into the substrate with a focused ion beam tool.

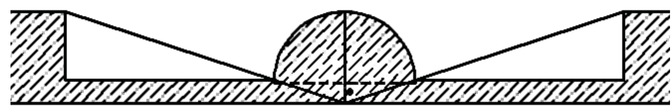


Figure 20: FOSSIL , Koyama et. al., IRPS 2003

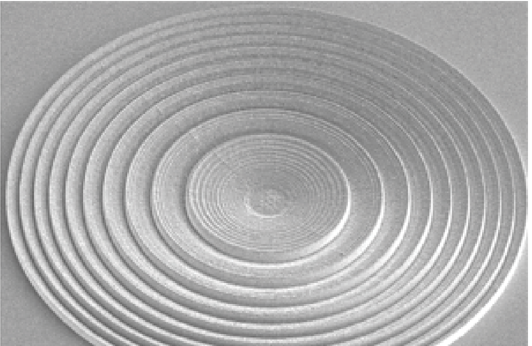


Figure 21: Zachariasse et. al. ISTFA 2005

In failure analysis practical SILSs are often built into a custom micro-objective assembly that mounts directly to a common microscope lens turret.

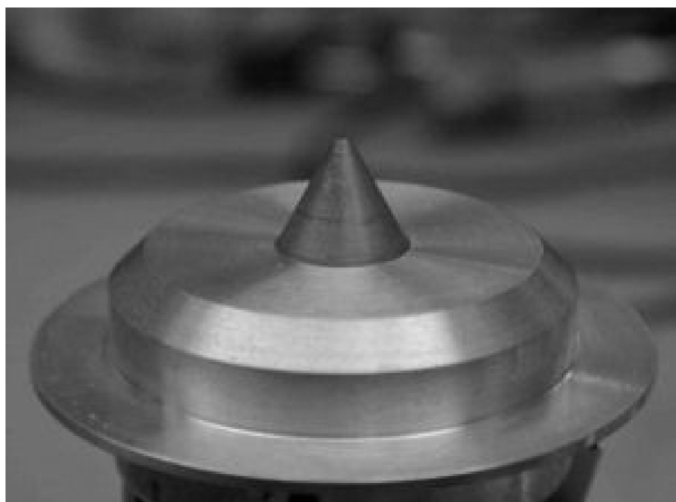


Figure 22: SIL Lens made by DCG. The tip of the Si cone contacts the substrate. The rest of the assembly contains collecting and magnifying optical elements.

The light emerging from the SIL must be collected by another lens assembly and for high performance SILs this must be located axially quite close to the SIL. It may be convenient to mount the SIL together with the collecting optics as seen above. Note that since the SIL resolution is very small, the overall magnification of the SIL together with the collecting objective must be large or the camera pixels will be found too coarse to preserve the newly obtained resolution. The magnification should be about 150x to 200x depending on the CCD pixel pitch. However the increased magnification will make the microscope much more sensitive to vibration, requiring more care with vibration isolation.

SIL focusing and navigating can be difficult. In normal microscope usage, the worker starts with a low magnification lens which is easy to focus and easy to find the area of interest. Then the worker progresses through greater magnification, at each lens re-adjusting the centering and focus in preparation to move the next higher magnification. When the solid immersion lens is finally employed it must touch the substrate and the contact must be hard to flatten the SIL to substrate contact. The lens or the sample must often be adjusted in 2 axes of tilt as well to make the contact flat it as well all of which requires a fair bit of fussing. Moreover, the field of view is very small because of the great magnification. So the user must hunt for the area of interest, all the while tapping or pushing the immersion lens to center it. Despite these difficulties, solid immersion lenses offer very improved resolution.

Brightfield / Darkfield

In brightfield microscopy, illumination light is introduced coaxially through the microscope objective. This is accomplished with a beam splitter, positioned just above the objective. This causes the field of view to be flooded with light and to appear bright, hence the name. Objects of interest at the specimen plane tend to reflect light out of the field of view, so that they appear dark in a bright field.

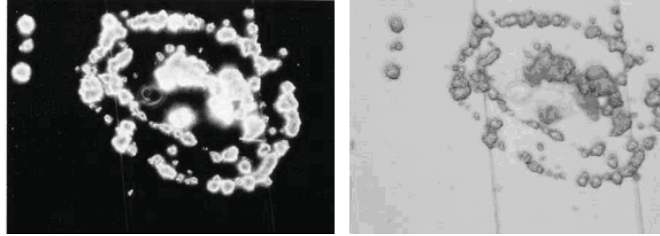


Figure 23: The left image of crystal defects is darkfield illuminated, the right image is brightfield. Light scattering objects show up well against the dark background.

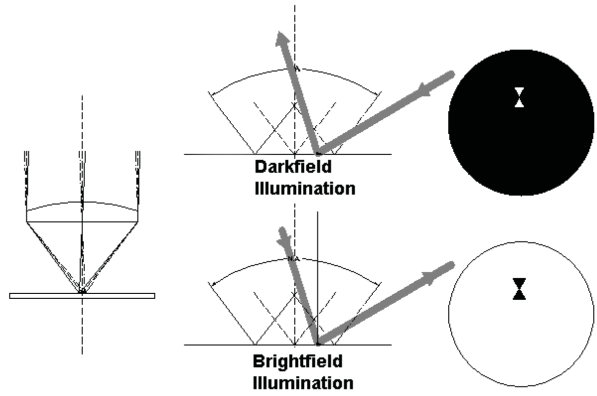


Figure 24: In darkfield, objects illuminated from outside the field reflect light into the field of view, appearing bright against a dark field. In brightfield the field is illuminated, and objects tend to reflect light out of the field, and so appear dark.

In darkfield the specimen plane is illuminated from light outside the numerical aperture cone of the objective. It will not enter the objective unless diverted by feature of interest, so the field appears dark and areas of interest appear brightly contrasted. The contrast effects are quite different and a feature can appear quite different by each method. A very rough surface may appear dark in brightfield illumination, since it reflects light away from the collecting NA of the objective. Thus a rough surface could be mistaken for a absorbing material or circuit contaminate. Brightfield illumination requires a beam splitter (figure 12) or it requires a condenser if illuminating from below the specimen. Darkfield illumination can be introduced with a fiber illuminator or other bright source aimed right at the working distance between the sample and objective. Many top grade emission microscopes have a darkfield ring illuminator on the macro lens. The working distance for most high magnification objectives is too close, and the light requirement too great to introduce darkfield light so casually. There are commercial darkfield objectives that deliver illumination from a surrounding housing reflected into the field from mirrors surrounding the final lens.

Polarized Light Microscopy and Liquid Crystal Hot Spot Detection

Most forms of natural and artificial illumination produce light waves whose electric field vectors vibrate in all perpendicular planes with respect to the direction of propagation. The electric field vectors may be restricted to a single plane with a filter. In that case the light is *polarized* with respect to the direction of propagation and all waves vibrate in the same plane. Polarized light microscopy employs 2 polarizing filters. The first, called a polarizer is placed inline near the illuminator. The second polarizing filter is materially identical to the first but is known as the analyzer. It is placed between the objective and the detector (human eye or camera). The analyzer (and sometimes the polarizer) may be rotated. Each filter blocks any light not polarized parallel to the axis of that filter. If the axes are arranged to be orthogonal, then all the light will be extinguished.

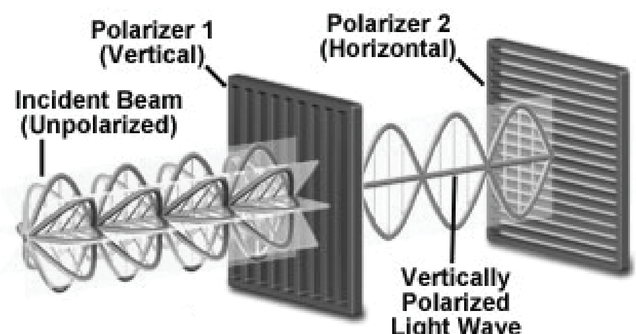


Figure 25: Crossed polarizers extinguish light vectors not parallel to the axis of the polarizer. Used by persmission from www.molecularexpressions.com.

Polarization techniques are effective for investigating materials that have properties that change the polarization of light. One such technique is liquid crystal detection of short circuits or other hot spots. The technique exploits a characteristic phase change in the liquid crystal solution. The phase change occurs at a known temperature, characteristic of the particular liquid crystal solution. Below the phase change temperature, the nematic phase, the liquid crystal changes incident light polarization. Above the characteristic temperature the liquid crystal loses this property. When placed between crossed polarizers, the circuit will be visible in the nematic phase, but gray or black above the characteristic temperature. The trick is to adjust the circuit temperature with a thermal stage while looking for the hot spot as the crystals in that area change phase. The technique is effective and relatively inexpensive. Unfortunately it requires contact with the heat-generating defect, so it is quite ineffective for backside inspection. Some labs are prohibiting the use of liquid crystal because the solvent, usually methylene chloride, may have properties harmful to humans.

UV Microscopy

The physics of microscopy favors short wavelengths for the most detailed resolution (equation 2). As of this writing 193nm microscopy dominates front-end photo-lithography precisely because shorter wavelength microscopy yields smaller feature sizes. However deep ultraviolet (UV) microscopy still isn't used much for back end inspection or for failure analysis. Below about 180 nm there is significant interaction between air and the UV photons, so vacuum must be employed between the objective and sample. There are significant technical difficulties with deep UV microscopy. There are only two glasses that can be used in the objectives, fused silica and calcium fluoride. Both glasses have nearly the same index of refraction so neither can be used to effectively correct chromatic aberration for the other. That is, at this time it is very hard to correct aberrations in deep UV micro-objectives. This is exacerbated of course, because the whole point to using UV is to obtain finer resolution, so we have less tolerance for aberrations than we would at longer wavelengths. In addition UV photons are so energetic that they tend to damage the glass and glass cements with protracted use. It is possible to use metal reflective optics but reflective optics often trade large field of view for numerical aperture or for suitable working distance. The result is that currently state of the art UV reflective objectives have too little N.A., or field of view and too much working distance and too little light throughput. These problems and the lack of glasses for refractive optics have so far severely limited the resolution that we hoped to achieve from deep UV. The light sources whether laser or plasma-generated are expensive and have short lifetime issues. Cameras for detecting UV are available but also immature. There just are no deep UV sensitive detector materials with high quantum efficiency. Deep UV microscopy is restricted to front side inspection since UV wavelengths do not penetrate silicon. We may yet see more UV in reliability and FA applications in the future but at this time deep UV has only barely touched semiconductor failure analysis. These technical problems are the same for photolithography so why is it useful for the front-end? In photolithography the end result, the desired pattern, is known in advance, unlike microscopy for inspection. The lithographer can employ phase shifting lithography masks to dramatically improve sharpness and resolution of the final patterned object and the resists can be adjusted for this as well. However it isn't possible to predict the nature of a flaw or contaminant in advance so unfortunately we cannot use a phase shifting technique for inspection.

Infrared Microscopy

For some FA labs, increasing numbers of metal layers and flip-chip packaging techniques are making front side circuit inspection impractical. However, silicon is transparent at infrared wavelengths longer than 1.1 microns so with infrared techniques, it is possible to examine the circuits from the back of the die or wafer. This has given rise to several infrared techniques for fault location and failure analysis. We use mid-wave infrared microscopes, from 2 to 4 microns to detect thermal infrared radiation, to locate short circuits by their joule heating. With careful calibration we can even measure junction temperatures from the front or back of the die. We

use short wave infrared, from 1 to 1.6 microns, to detect recombination photoemission sites that indicate a variety of failures. We use a special form of a laser-scanning microscope, equipped with 1.34-micron wavelength lasers to generate micro-local heating from the backside. This is known variously as TIVA, OBIRCH or XIVA fault detection techniques. If an infrared laser much closer in wavelength to the silicon bandgap is employed, such as 1.064 microns, it can generate photo-carriers at micro-locations. That enables OBIC, LIVA and PC-XIVA techniques for locating faults from the backside.

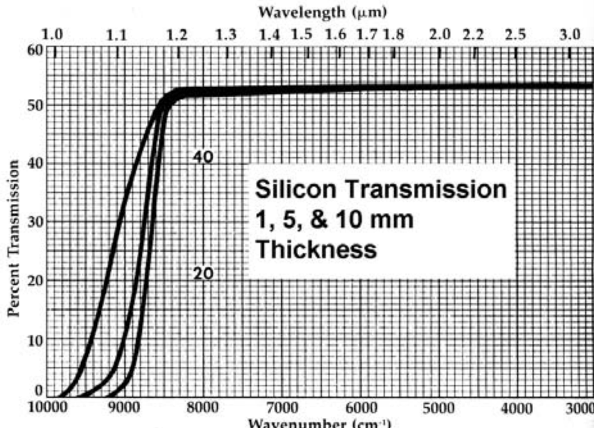


Figure 26:Undoped silicon transmits infrared well.

Silicon Transmission & Surface Reflection

The plot for undoped silicon shows the transmission rising at about 1 micron wavelength, and rises to and stays about 54% to the end of the plot. The transmission is actually quite good to about 9 microns wavelength where there is an oxygen absorption dip, and it persists to longer wavelengths. Silicon is commonly used for lens material for 3 – 5 micron military night vision application. Silicon's very high index of refraction causes light to travel 3.5 times more slowly in silicon than in air. That causes an impedance mismatch at the silicon air interface that generates a vigorous reflection. The maximum transmission in the graph is limited by reflections at the surface. With proper anti-reflective coatings it has a transmission of 98% or higher. The failure analyst can improve their backside microscopy results by applying an anti-reflection coating but high performance vapor-deposited anti-reflection coatings are not trivial and are outside the scope of this tutorial. There are some commercial spin-on coatings that provide some benefit. Likewise, in a pinch you can apply a drop of extra virgin olive oil for some relief from reflection, and also as a quick and dirty means of smoothing a rough or poorly prepared silicon surface.

Doped Silicon Backside Thinning

Heavily doped silicon is much less transparent than undoped silicon because of band-gap shifts and because of free carrier absorption and scattering. Following absorption measurements by Aw, et. al. [6], Aaron Falk [7] gives empirical methods for calculating absorption and required thinning for doped silicon. Using his formulas, the following figures show calculated optical transmission for a 600, 100 and 50 microns thickness, including reflection loss from one surface. Figure 28 shows transmission less than one percent for 100 microns of heavily P doped silicon, and the spectral bandpass is limited to a region near the bandgap at 1.1 micron.

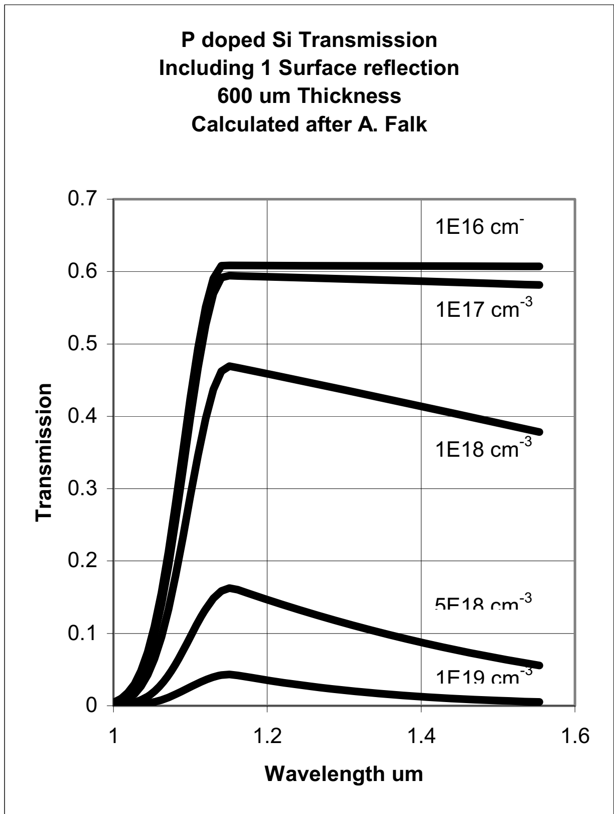


Figure 27:Transmission of P-Doped Silicon falls dramatically with increasing carrier concentration. Calculated from empirical formulas given by A. Falk [7].

In the case of heavily doped substrates, the silicon must be thinned to permit backside inspection. However many circuits employ lightly doped silicon substrates and no thinning at all is required on these. Calculating the opacity may not be necessary. It takes little time to examine a die for backside transparency. Substrates that must be thinned are typically reduced to about 100 microns thickness as a suitable compromise between transmission and thermal and mechanical fragility. Many labs are content to thin with a polishing disk. Others employ machines to open rectangular "pockets" in packages. These machines may be of the CNC Mill variety or much less expensive gravity feed grinders. The thinned surface ultimately becomes an optical surface. Take care to polish the surface at the end of the thinning process. Any remaining surface roughness will scatter light and degrade the image. For laser techniques, any residual waviness will cause interference artifacts that will impede the examination.

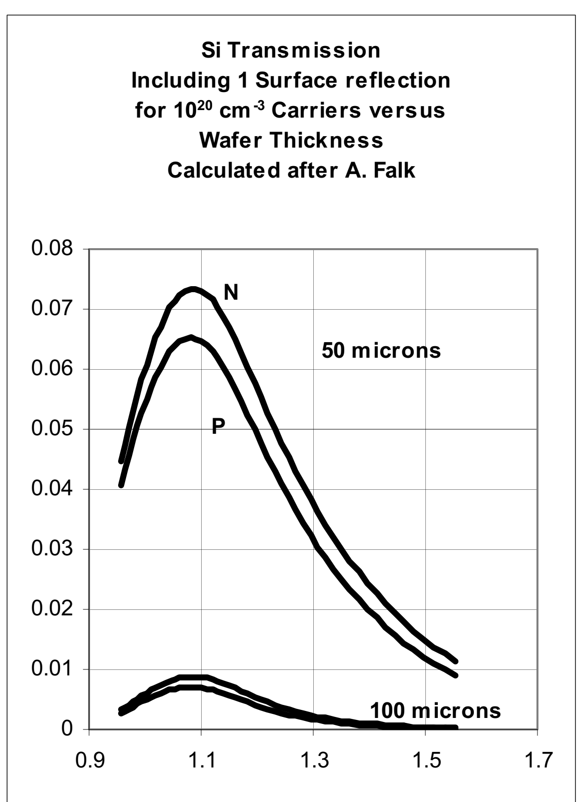


Figure 28:Transmission of Si improves with wafer thinning Note that transmission for 10^{20} cm^{-3} material through 100 microns is less than a percent . Calculated from empirical formulas given by A. Falk in [7].

Photoemission Infrared Microscopy

Many defects permit unwanted electron-hole recombination. An example would be a conducting path punched through an oxide by static damage. Electron-hole recombination is typically accompanied by photon emission, and these are known as recombination photons. Photon Emission Microscopy, sometimes known by the initials PEM, is a technique by which such failure sites are located by their photon emissions. The PEM is most often comprised of a very sensitive camera mounted on a microscope in a dark enclosure. Usually the camera employs a peltier or cryogenically cooled detector to reduce noise and improve sensitivity. The setup often includes a probe station or means to dock to a functional tester. Photoemission microscopy is so effective in locating failure sites that it is frequently the first failure analysis tool to be employed. It is reasonable to expect that Si emissions will occur at or near the 1.1um bandgap, but in junctions some emissions may be blue shifted due to exogenous energy from bias (hot electrons). Likewise some emissions may be red shifted due to phonon losses to the lattice. In the literature we see evidence for slight NIR shifting of emissions as in the following University of Singapore data, but we also see strongly red-shifted data as published by Rowlette, et. al.

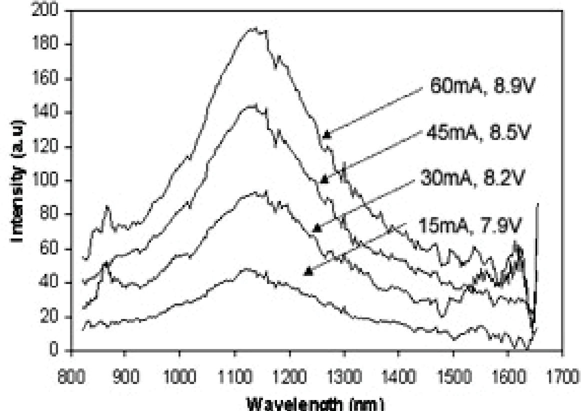


Figure 29. Emission spectra under reverse bias showing slight red shift near the 1.1um bandgap of silicon. Len WB, et. al., University of Singapore, ISTFA 2003.

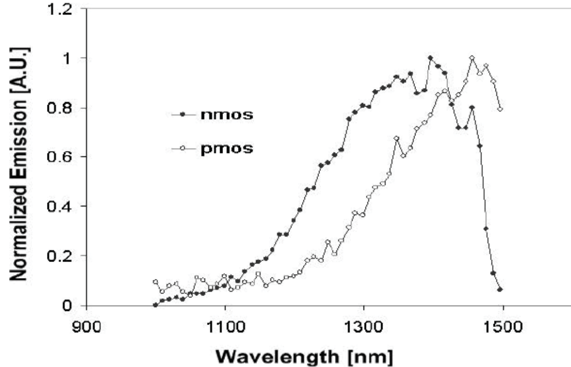


Figure 30:Rowlette reports red shift of PMOS recombination emissions from a 10 micron length transistor. Data given at LEOS, 2003 [8].

It is hard to reconcile these very different results. The author's experience is that the vast majority of samples defects can be found easily and even better with a silicon based scientific camera, which argues for the Singapore data. But indeed there are some emission samples that cannot be seen at all with a silicon detector and only a shortwave infrared (SWIR) detector such as InGaAs or HgCdTe will suit. The silicon camera solution tends to be much less expensive and easier to operate, but if your samples require the SWIR detector then that us what you must employ. It should be noted that the Rowlette camera employed a sharp 1.5 um cut-off filter so the data increases in some unknown way beyond 2.5um. Also, the device in the Rowlette data was a normally operating MOS transistor under time-resolved emmi investigation. There may have been very little electron-hole recombination in the device and perhaps more I^2R heating and indeed, thermal blackbody infrared radiation.

Photoemission High NA Macro Lenses

Modern NIR or SWIR photoemission microscopes must employ NIR (near infrared) objectives, rather than the more common visible objectives. The NIR lenses are far superior for long wave microscopy compared to visible lenses. These objectives were never intended to work at wavelengths as long as 1.6 microns but they seem to have adequate performance in that band and they are reasonably priced and available. Common optical glasses do not work for Mid-wave, 2-4 micron infrared microscopy. Mid-wave infrared requires custom objectives with silicon and germanium lenses. Future requirements for photoemission microscopy may encourage us to develop similar custom lenses for SWIR work, but for now, commercial NIR objectives suit.

Specially designed high N.A. macro lenses have considerable value for photoemission work and the user should insist on them. Macro lenses offer 1x or less magnification and they are needed for canvassing the entire die at once to find a defect that could be anywhere. If the PEM is not equipped with a macro, the user is forced to systematically scan a large die with higher magnification lenses, a tedious and time-consuming task. The macro must have a high NA to gather all available signal so as to reveal the emission sites. However most commercially available macro lenses compromise the NA terribly in order to make the lens mechanically compatible with the higher magnification objectives in the lens turret. The best macro objectives for photoemission work abandon mechanical compatibility in favor of extremely high signal throughput. The high NA macros are created by splitting the optical power between two equal or similar lens groups, one near the camera focal plane, and the other near the sample focal plane. A simple way to do this is to employ 2 matched 35mm camera lenses with equal or similar focal lengths and fast F/#s arranged per figure 31. A pair of lenses arranged this way has amazing light gathering power, and often the emission sites are most readily located with the macro. The technique is described in US patents 4,680,635, Khuruna, and 4,755,874 Esrig. The magnification of the paired lenses is given by the ratio of their focal lengths; hence if they are equal, the magnification will be 1x. If the lens nearest the camera has a focal length shorter than the lens nearest the sample, the magnification will be less than unity.

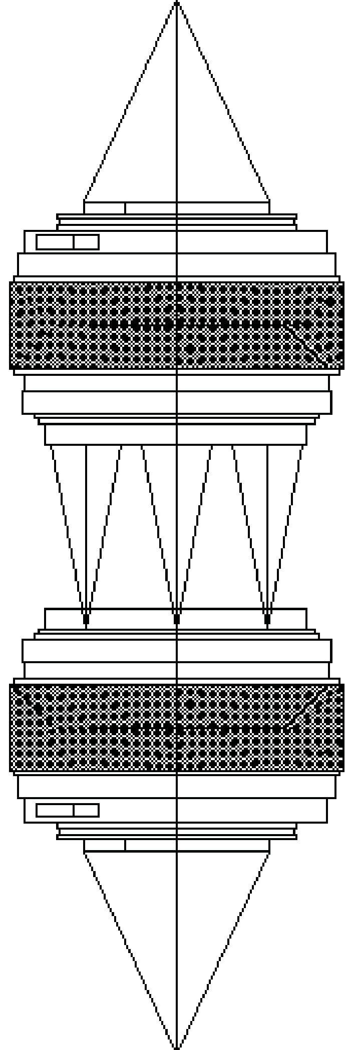


Figure 31:A pair of camera lenses used back to back makes an extraordinarily sensitive 1x macro for photo-emission detection.

Thermal Infrared Hot Spot Detection

In the past liquid crystal techniques served to find short circuits from the front surface. Liquid crystal techniques fail for backside inspection because the liquid crystals must themselves be warmed by the hot spot to work, and heat from the short diffuses too much to be to be detected from the backside. However all warm objects radiate some infrared photons according to the Planck blackbody law and this thermal photon emission may be used to detect short circuits from the front or back of the die.

Thermal infrared light from objects near room temperature radiates from the motion within and between molecules, from twisting, bending and coupled oscillations caused by thermal excitation. At higher temperatures the agitation is greater, increasing the radiation overall and blue shifting the peak radiation. At about 800 Kelvin the radiation peak wavelength is short enough to be visible to the eye as a kind of dull orange, seen in glowing coals in fireplaces.

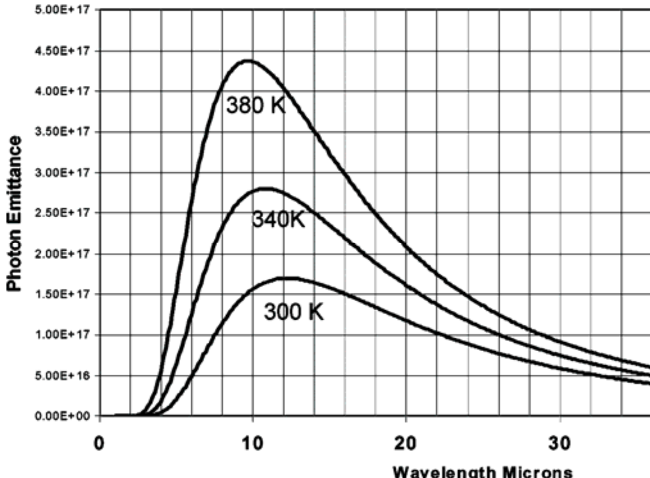


Figure 32:All objects warmer than absolute zero emit infrared radiation according to the Planck blackbody law. This property may be used to detect heat from short circuits via infrared microscopy.

$$Q(T,\lambda) = \varepsilon \cdot \frac{2\pi c}{\lambda^4} \cdot \frac{1}{e\left(\dfrac{ch}{\lambda kT}\right) - 1}$$

Equation 4:Planck's law yielding the curves above where:

Q = Photons per second per cm^2 per μm
ε = emissivity constant of the material ranging from 0 to 1
c = velocity of light $3x10^{10}$ cm/s
h = Planck's constant, $6.63x10^{-34}$ watts•$seconds^2$
k = Boltzmann's constant, $1.38x10^{-23}$ watts•seconds per Kelvin
λ = wavelength in centimeters
T = source temperature in degrees Kelvin

$$Q = \varepsilon \sigma' T^3$$

Equation 5; The integral of Planck's law over all wavelengths known as the Stefan-Boltzmann Law illustrates the cubic relationship of radiation to temperature, where:

Q = radiant photon emittance, photons per second per cm^2
ε = Emissivity constant of the material ranging from 0 to 1
σ' = $1.52x10^{11}$ per second per cm^2 per ºKelvin^3
T = Kelvin temperature of the source

Not all materials emit equally well according to the Planck laws. A perfect emitter is also a perfect absorber. A perfect absorber has no reflection so this type of source is called a blackbody source, and it was for a blackbody source that Planck formulated his laws. Any given material may be more or less efficient at emitting infrared radiation and the property that describes how closely a material approaches a perfect blackbody source is called emissivity, which ranges from zero to unity. Metals are poor emitters of infrared with emissivities near or below 10%. Semiconductor circuit materials tend to be about 40% to 60% emissive. Packaging materials, plastics, ceramics and epoxies usually have very high emissivities. Thermal infrared microscopes are very effective for locating current leaks from their heat signature, but less so if the leak is under metal or some low emissivity material. A thin layer of black paint applied to the surface of a circuit can enhance the emissivity and boost the infrared signature to reveal hard to find hot spots.

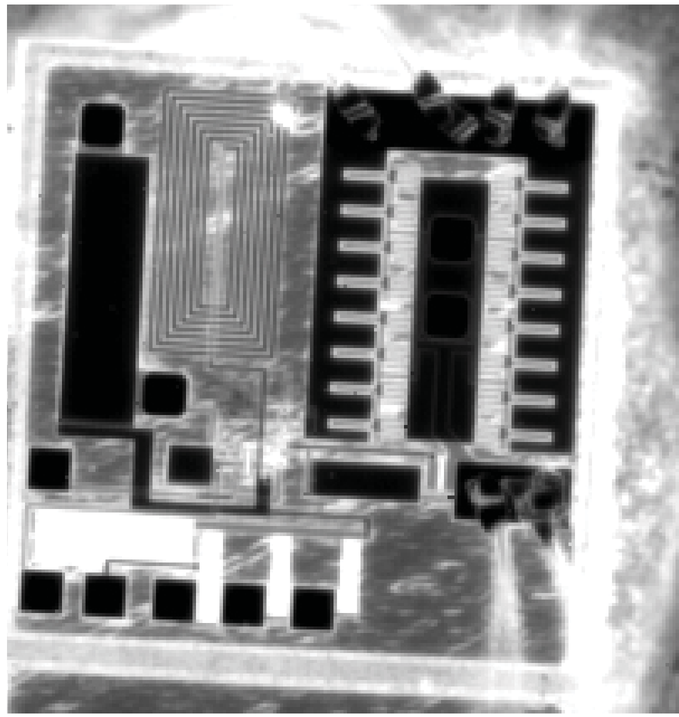


Figure 33: This is a thermal infrared image of a GaAs radio frequency amplifier. The circuit isn't powered and there is no external illumination. The detected infrared is all from the IC and all of the contrast is due to emissivity differences. The metal lines and pads emit much less infrared than does the junction areas or the packaging.

Thermal infrared has subtle but important differences from other sources of light. A dark box does little good for thermal microscopy, because the box walls themselves glow with infrared emissions. You cannot turn out the lights in the infrared, except by cooling the source. Normal glass lenses are opaque and will not work at thermal infrared wavelengths. Instead infrared optical workers use silicon or germanium for lens materials, or exotic compounds such as zinc-selenide, or magnesium fluoride. This makes thermal infrared objectives much more expensive than conventional objectives, and requires optical shops with special skills to grind the lenses. Thermal infrared is long wave, compared to recombination photoemissions, and so is limited by diffraction to relatively larger airy disks and comparatively coarse resolution.

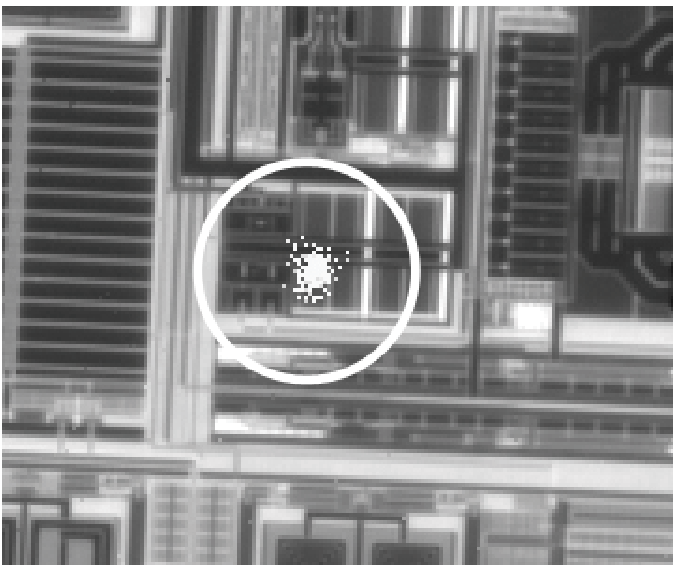
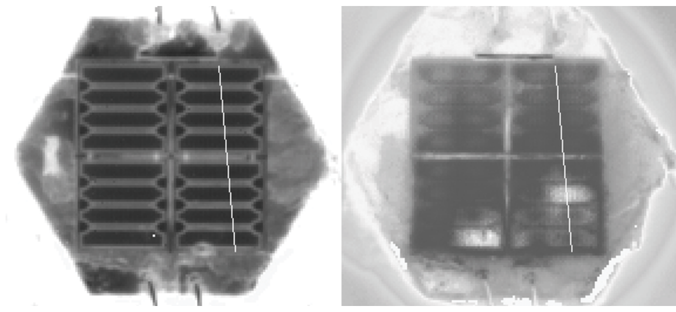


Figure 34: Thermal infrared detected hot spot on a silicon IC, overlayed on a thermal infrared reference image. Normally the hot spot is displayed in false color to contrast vigorously with the black and white reference image. Courtesey of Quantum Focus Instruments Corporation.

Inspection of the Planck curves in figure 32 show that the peak radiation is near 10 microns wavelength. However the Sparrow Criteria favors shorter wavelengths for best resolution. The manufacturers of thermal infrared microscopes balance these requirements and the overwhelming vast majority of fielded instruments operate between 2 to 4 microns wavelength. By contrast, recombination photons detected in emission microscopy are limited to 1.6 microns wavelength or less.



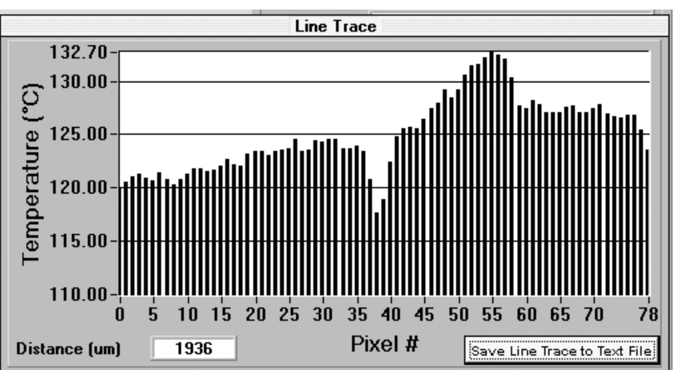


Figure 35: Certain thermal infrared microscopes are designed to accurately measure temperatures on ICs. Upper left, thermal infrared reference image of a 5 watt high brightness LED (for illumination). Upper right, false color temperature image of the HBLED showing excessively hot regions. Lower, values of temperature in degrees celsius along the line. Courtesy of Quantum Focus Instruments Corporation.

Laser Signal Injection Microscopy

Shorts, junction defects, problem VIAs and other integrated circuit defects can sometimes be located from the front or backside with Laser Signal Injection Microscopy (LSIM). An LSIM works by scanning a laser beam through a microscope lens, crossing an integrated circuit while monitoring the circuit input and output for laser-induced changes. Various laser wavelengths induce different effects. Short wavelength lasers produce electron-hole pairs, (photo-carriers) in the semiconductor. These photo-generated carriers reveal failure sites in transistors and PN junctions. Longer wavelength lasers induce micro-local heating to reveal resistive and ohmic problems. Both photo-carrier generation and microthermal heating may be performed from the front or backside of the wafer. Various techniques have evolved for semiconductor fault location by Laser Signal Injection. The pioneer was OBIC in 1988. Since then other techniques have appeared known by names such as LIVA, TIVA, OBIRCH, SEI and XIVA. Each of these techniques generates either photo-carriers or micro-local thermal heating to reveal defect locations. They differ in how the circuit under test is biased and connected to the measuring amplifier. The LSIM is actually an adaptation of a laser-scanning microscope (LSM). In the LSM, the injected laser light is scanned over a circuit in a raster pattern, and the reflection is detected and displayed. An LSM image appears much like any light microscope image.

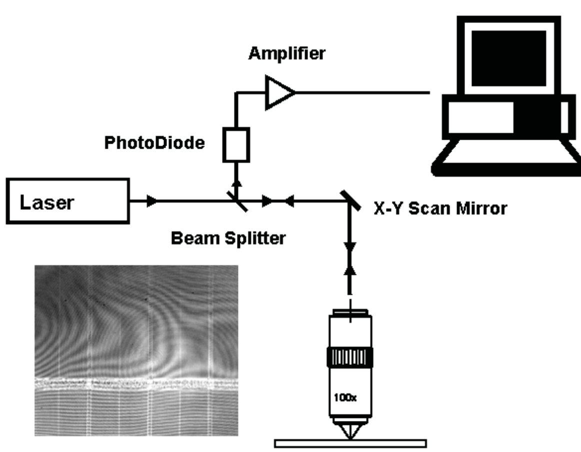


Figure 36: Laser Scanning Microscope (LSM) forms an image by detecting reflections from a raster scanned laser. The example is a backside image showing interference lines caused by laser light waves reflecting from both surfaces of the die, which causes destructive and constructive interference.

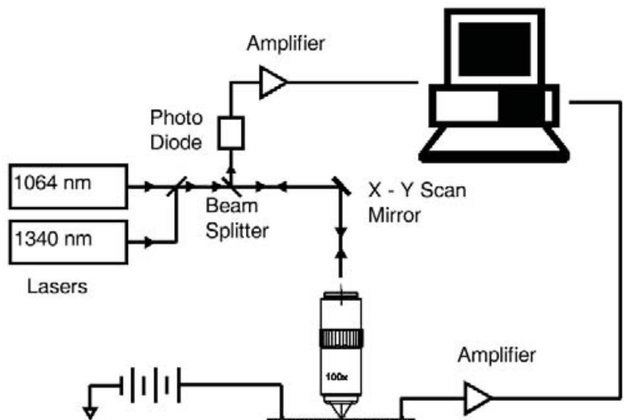


Figure 37: Laser Signal Injection Microscope (LSIM) forms an image by detecting laser induced changes in microcircuits. More commonly known by OBIRCH, TIVA, XIVA, etc. Note that an LSIM can simultaneously produce an image by reflection.

As can be seen in figure 36, images from laser microscopy from the backside often show interference patterns. These patterns result from the destructive and constructive interference of monochromatic light waves that are reflected from each surface of the die. The effect can be so pronounced that the circuit can be quite hard to see. It can be mitigated with antireflection coatings on the back surface of the die, or in some cases by regrinding the die back surface. Often the fringing is more evident at low magnifications.

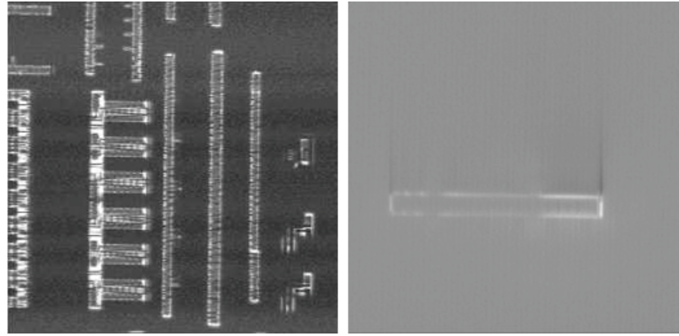


Figure 38: Left: OBIC image from laser induced photo-carriers, courtesey of Optometrix Inc. Right: Thermal XIVA image from microthermal heating showing current in a conductor. Scan direction on right image is vertical.

List of Laser Techniques:

Known as:	Full Name	Approx. Intro-duction	Physics	Application
OBIC	Optical Beam Induced Current	1988	Photo-Carrier	Junction Defects
LIVA	Light Induced Voltage Alteration	1994	Photo-Carrier	Open Junctions and Substrate Damage
OBIRCH	Optical Beam Induced Resistance Change	1998	Thermal	Locations of Shorts and defective VIAs
SEI	Seebeck Effect Imaging	1999	Seebeck Effect	Location of Opens
TIVA	Thermal Induced Voltage Alteration	1999	Thermal	Locations of Shorts and defective VIAs
XIVA	External Induced Voltage Alteration.	2001	Thermal or Photo-Carrier per wavelength	Junctions, substrates, shorts.

Acknowledgements

Special thanks to Michael W. Davidson of the National High Magnetic Field Laboratory at Florida State University for generous permission to use many of the excellent illustrations from www.molecularexpressions.com. The author encourages the reader to visit the Molecular Expressions website for an amazing wealth of information on microscopy.

Thanks also to Aaron Falk for providing his absorption calculations in the form of a spreadsheet, adapted here to generate the doped silicon transmission curves.

Additional Resources

Online: www.molecularexpressions.com. Excellent and voluminous microscopy tutorial from the Florida State University and the National High Magnetic Field Laboratory.

General Optics:
Optics, Eugene Hecht et, al. Excellent and well illustrated survey of optics.

References

1. E. Spitta, *Microscopy: The Construction, Theory and Use of the Microscope*, 3rd edition, E.P. Dutton and Company (1920) (Out of print).
2. Y. Asoma, U.S. Patent 4,505,553
3. S. B. Ippolito, A. K. Swan, B. B. Goldberg, and M. S. Ünlü, "High Resolution Subsurface Microscopy Technique," Proceedings of IEEE Lasers and Electro-Optics Society 2000 Annual Meeting, Vol. 2, 13-16 November 2000, pp. 430-431
4. W. Smith, *Modern Optical Engineering, 2nd edition*, p. 261 & p. 413, McGraw-Hill Inc., New York (1990)
5. T. Koyama, E. Yoshida, J. Komori, Y. Mashiko, T. Nakasuji and H. Katoh: "*High Resolution Backside Fault Isolation Technique Using Directly Forming Si Substrate into Solid Immersion Lens*," *Proc. Int. Rel. Phys. Symp.* (IRPS), 2003, pp. 529-35.
6. S. W. Aw, H. S. Tan, et al. "*Optical Absorption Measurements of Band-Gap Shrinkage in Moderately and Heavily Doped Silicon*." J. Phys.: Condensed Matter 3: 8213-8223, 1991.
7. R. A. Falk, "*Near IR Absorption in Heavily Doped Silicon - An Empirical Approach*", Proceedings of the 26th ISTFA, 2000.
8. J. Rowlette, E. Varner, S. Seidel, 'Hot Carrier Emission from 50 nm n and p-Channel MOSFET Devices" *Conference on Lasers & Electro-Optics*, (LEOS) 2003.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Scanning Electron Microscopy

W. Vanderlinde
IARPA
Washington, DC

Introduction

The scanning electron microscope (SEM) remains the most versatile instrument in the failure analysis lab for high resolution imaging. Other tools such as TEM or AFM may have higher ultimate resolution, but the SEM is more generally useful because of simple sample preparation, high depth of field, and ease of navigation. For comparison, TEM requires thinning a sample to a 100 nm or less and then only a very small area of the sample can be imaged. AFM requires a sharp tip to be scanned across a sample surface, thus only a very small area can be imaged and the apparent shape of the sample surface can be distorted by tip convolution effects. The SEM requires little or no sample preparation and large samples can be scanned. Under the right circumstances SEM can achieve 1 nm resolution which is very close to TEM and AFM. It is also relatively simple and inexpensive to add an energy dispersive x-ray detector (EDX) to a SEM which then provides a very useful chemical micro-analysis capability (see Chapter 15).

This paper provides an overview of how to use the SEM for imaging integrated circuits. A minimum of SEM theory is covered while most of the article describes practical methods for getting a good image. Specialized SEM techniques for defect localization such as voltage contrast, electron beam testing, and charge induced voltage alternation (CIVA) are covered in Chapter 11.

Principles of Operation

SEM / optical comparison
It is always a good idea to inspect a new sample in an optical microscope before putting it into a SEM. The SEM has much higher resolution that an optical microscope. However, the best SEM images are taken in slow-scan mode and TV rate SEM images tend to have poor resolution and poor contrast. Thus it can be very difficult to navigate and find objects in the SEM. Furthermore, many defects that show strong contrast in an optical microscope may show little or no contrast in the SEM. The optical microscope also has the advantage that dielectric layers are transparent, and the shallow depth of field can actually be an advantage since it allows one to "optically section" a part, i.e. tell what features are in the same horizontal layer. Since it may take several minutes to vent and pump down the SEM to transfer a sample, it is well worth the effort to first map out a plan of action using a visual inspection in the optical microscope. A comparison between SEM and optical microscopy is summarized in Table 1.

Table 1: Comparison of SEM and Optical Microscopy.

Technique	SEM	Optical Microscope
Resolution	Few nm	$\lambda/2 \sim 250$ nm
Depth of field	Large (up to mm)	Very shallow (~λ) at high magnification
Contrast	Material composition, topography	Reflectivity, color, phase contrast, bright field/dark field, polarization
Sample loading	Slow – requires vacuum	Fast – no vacuum
Sample preparation	May require sectioning or delayering, Insulating samples may need sputter coating	Usually none - dielectrics are transparent, No problem with insulating samples
Navigation	Difficult – poor contrast and resolution at TV rates	Easy – real time optical image
Defect localization	Defects often show little contrast, Special localization techniques may be needed	Defects are often easily visible, especially in dark field or phase contrast
Sample Damage	May damage active devices	Non-destructive

Scanning principles
The scanning electron microscope consists of an electron column, a sample stage, and one or more detectors, see Fig. 1. The electron column will have an electron source, a series of magnetic and electrostatic lenses, and raster coils for scanning the beam. Of course the range of electrons in air is very limited so a vacuum system is also required.

The scanning electron microscope operates by scanning a finely focused electron beam across a sample surface in a raster pattern. Secondary electrons emitted from the sample surface are collected and amplified, and this signal is then used to modulate the intensity on a TV monitor that is rastered in synchronization with the electron beam. It is important to note that the electron column is used *only* to create a small spot on the sample surface. This is quite different from a TEM where image forming lenses are used to create the image. The

SEM's secondary electron detector is not a camera and it has no ability to detect where an electron comes from, it only counts electrons. All the image information comes from the fact that the electron beam raster and the image raster are synchronized.

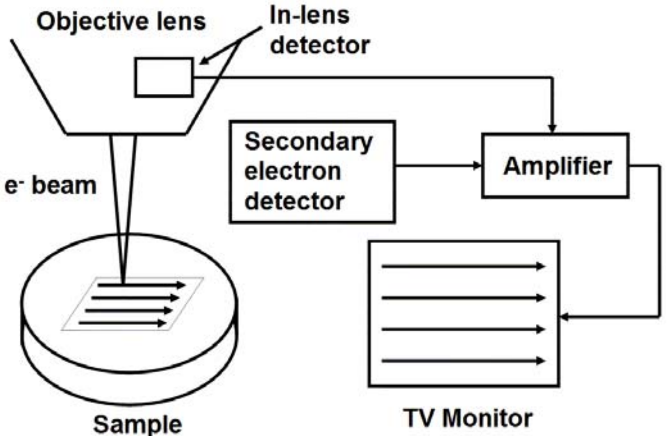


Figure 1: Block diagram of a scanning electron microscope.

The scanning method may seem like a rather artificial way to produce an image, and yet the SEM readily produced very lifelike, easily interpreted, three dimensional images, see Fig. 2. Although images like this are easy to interpret, the artificial way they are constructed can lead to some confusion. First, the perspective on the sample is as if one is looking down the column of the microscope, not viewing the sample from the detector. Remember, the detector is not a camera and it simply counts electrons. What the detector does is to provide an apparent "source of illumination" for the image, i.e. secondary electrons can more easily get to the detector when they are emitted from a part of the sample that faces the detector.

Therefore it is vital to know which direction the image is oriented relative to the secondary electron detector. The human brain is wired to expect illumination from above, and will misinterpret images that are illuminated from below. In Figures 3(a) and 3(b), the same object is seen differently when illuminated from above and below. The correct image is Figure 3(a), a hill on a bumpy surface.

Typically, a SEM will be set up with the secondary detector at the top of the image, but some users may rotate the raster for convenient viewing of cross-sections or other unusual situations, so it is wise to check.

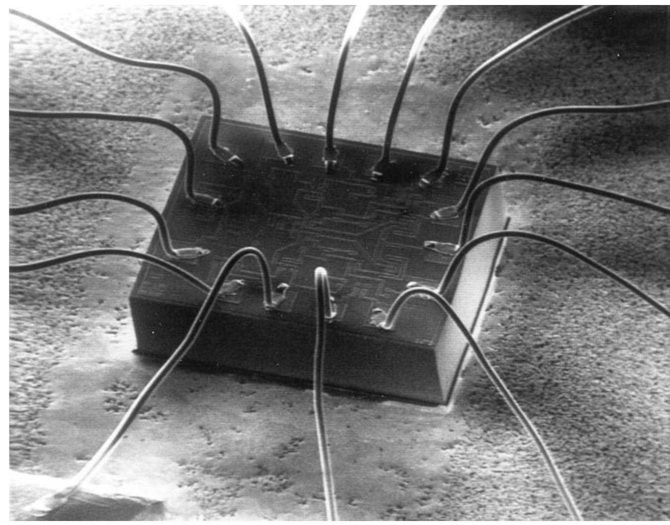
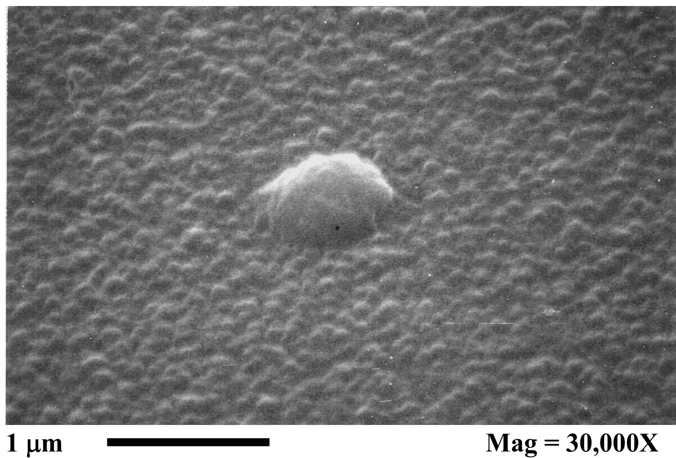
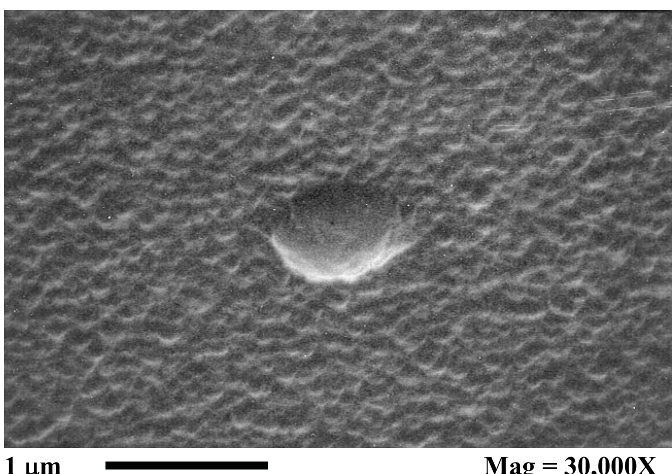


Figure 2: Low magnification SEM image of an integrated circuit.



1 µm **Mag = 30,000X**

Fig. 3(a): SEM image of a hill on a bumpy surface. Secondary detector is toward the top of the image (top illumination).



1 µm **Mag = 30,000X**

Fig 3(b): SEM image of the same feature as Figure 3(a), but using incorrect bottom illumination.

In-lens detection
Some newer microscopes will have an additional secondary electron detector located inside the final lens, called an "in-lens detector." Because the detector collects electrons that are emitted normal to the sample surface, images taken with this detector will show much less surface topography that those taken with the usual in-chamber detector. However, it is easier to collect electron from deep holes in the sample which would otherwise be deeply shadowed. Sometimes sample charging effects can be minimized by using in-lens detection or some combination of in-lens and chamber secondary signal.

Magnification
Another consequence of the scanning process is that the image magnification is purely the geometric ratio of the raster size on the TV monitor to the corresponding field size on the sample surface. As shown in Figure 4, magnification M is give by:

$$M = L / l$$

where:
L = the length of an object on the TV monitor
l = the length of an object on the sample surface

To *increase* the magnification, one simply *decreases* the raster size on the sample. Thus high magnification is achieved by using a very small raster area on the sample. Some new SEM users may think it is a good idea to "turn down the magnification" when the SEM is idle, but in fact it is low magnification that uses maximum current in the rater coils and it is unwise to leave a SEM at lowest magnification for long periods of time.

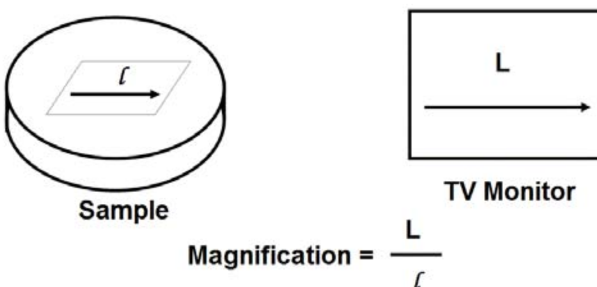


Figure 4: Magnification is the geometric ratio of the raster size on the monitor to the field size on the sample.

The maximum useful magnification will be achieved when the spot size on the sample corresponds to the pixel size on the monitor. Any greater magnification of the image will be "empty magnification" that simply magnifies a blur and does not improve the image. As shown in Figure 5, if the electron beam spot size on the sample is given by **d**, then the spot size translated to the TV monitor will be d * M. The human eye cannot focus on objects smaller than about 0.2 mm and that is typically the size of one picture element or "pixel" on a digital computer screen. If the electron spot size translated to the monitor is less than one pixel diameter then the image will be in sharp focus, but if the spot size overlaps two or more pixels then the image will appear fuzzy. Thus the maximum useful magnification will be given by:

$$M_{max} = 0.2 \text{ mm} / d$$

For a spot size of d = 5 nm, the maximum useful magnification will be 40,000x. Newer field emission SEMs may have a spot size as small as 1 nm which corresponds to a maximum useful magnification of 200,000x. At very high magnifications above 100,000x factors involving beam-sample interaction will limit the useful magnification and special technique may be required to produce good images.

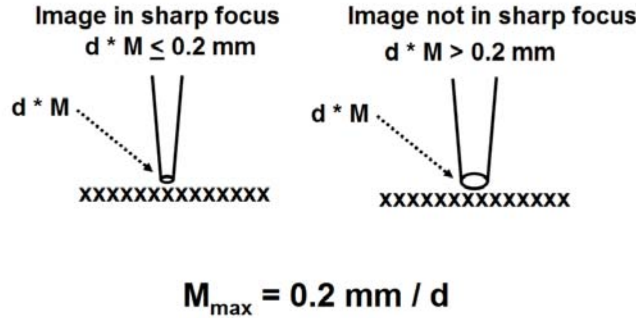


Figure 5: Illustration of maximum useful magnification.

It is worth mentioning that more than one "magnification convention" has been used with SEMs. Historically, most SEMs were equipped with Polaroid cameras for taking permanent images. These Polaroid cameras almost universally produced 3.5" by 4.5" photos. The TV monitors on the SEMs were about twice that size. Since magnification is quite literally the ratio of the image size to the field size on the sample, the actual magnification on the TV screen was about twice that on the Polaroid photos. Since the same magnification number appeared on both, one of them had to be wrong. In fact, by convention the magnification number printed on the photo was correct and the screen magnification was actually about twice that. This was known as the "Polaroid magnification convention." Modern SEMs are generally run by computer control and images are saved to computer files like TIFF or JPEG. These files will display on a computer monitor or in a printed document at many different physical sizes depending on the users choice. For this reason, many SEM manufacturers have now taken to displaying the true screen magnification in their images and the Polaroid convention is obsolete. It is important when comparing images between older and newer microscopes that the magnification numbers may be off by a factor of two when viewing images at the same true magnification. In any case, it is essential to always include a scale bar or "micron marker" with every image. The magnification number has become somewhat meaningless, but the scale marker never lies.

Brightness
From the discussion of magnification, the highest resolution is achieved when the spot size on the sample surface is small. However, there are two additional qualities one needs for good imaging: a large current and a small convergence angle. Large current is needed to create a strong signal for the detector. A small convergence angle is needed to produce a large depth of field and to reduce lens aberrations such as spherical aberration that can ruin an image. Given that we wish to have a large current into a small spot and a small convergence angle, we can define a quantity called Brightness β which can be calculated from:

$$\beta = \text{current} / \text{area} * \text{solid angle} = 4\,I / \pi^2 * D^2 * \alpha^2$$

Where:
I = current
π = 3.14
D = beam diameter
α = convergence angle

The brightness is a fundamental figure of merit for an electron optical system and depends on the type of electron source. Devices such as magnetic lenses and apertures can trade off spot size and aperture angle for current, but the brightness is conserved and the brightness at the source is all that will ever be available anywhere in the column.

SEM cathodes
Over the years the quest for brighter sources has resulted in substantially improved electron microscopes. The fundamental problem is how to get a large number of electrons out of a material so that they form a tightly focused beam. Typically electrons are trapped in materials by an energy barrier called the work function. A material can be heated until the electrons have enough energy to jump over the work function barrier. This is known as thermionic emission. Alternately, a strong electric field can be applied that causes the electrons to tunnel through the work function barrier, and this is called field emission. Electron sources are also known as cathodes since they are at a negative potential relative to ground. The terms "filament" and "tip" are also sometimes used.

The oldest electron source is the tungsten hairpin thermionic emitter. This source works by heating a thin tungsten wire until electrons boil off and are directed down the column by a high voltage. Since only a relatively poor vacuum is needed, this tungsten source is relatively cheap and is still often used on less expensive SEMs. The primary limitation is that the tungsten wires have the relatively short service life of roughly two weeks.

An artificial crystal composed of Lanthanum hexaboride (LaB_6) was developed to which has a lower work function and thus achieves higher brightness at a lower operating temperature. LaB_6 sources may last six months or more and the higher brightness allows for more current into a smaller spot. However, the LaB_6 tip is very sensitive to surface contamination which means that a higher vacuum is required, which makes the microscope more expensive.

The highest performance electron microscopes use field emission sources. These will have tungsten tips which are sharpened to a very fine point which concentrates an applied electric field such that electrons tunnel through the work function barrier. Field emission sources have a very high brightness which allows for high resolution even at low beam voltage. This makes them useful for looking at insulators which tend to charge up at high beam voltage. However, they require the highest vacuum and are the most expensive. Cold field emission sources are very sensitive to stray gas atoms landing on the tip so they tend to have unstable current output which makes them unsuitable for x-ray mapping. However, the low energy spread of the cold field emission sources reduces chromatic aberration which allows for very good performance at low beam voltage. The Schottky field emission electron sources uses thermal assistance which reduces performance slightly but results in a more stable beam. Most high performance SEMs sold today have Schottky field emission sources. Table 2 shows a comparison of the common electron sources.

Table 2: Comparison of electron sources.

Source	W Hairpin	LaB_6	Cold field emission	Schottky field emission
Vacuum (Torr)	10^{-5}	10^{-7}	10^{-10}	10^{-8}
Brightness (A/cm2-sr)	10^{+5}	10^{+6}	10^{+8}	10^{+8}
Resolution	10 nm	5 nm	1 nm	1 nm
Lifetime (hours)	40-100	200-1,000	> 1,000	> 1,000

Electron optics
The electron column consists of a series of lenses and apertures which are designed to produce a small electron spot and convergence angle while maximizing the beam current (see Figure 6.) The electron source or cathode is held at a high negative potential V_o called the beam voltage of the system. A metal cap called the Wehnelt is held at a slightly more negative potential which tends to repel electrons, but the electrons are attracted to an anode plate beyond the Wehnelt which is held at ground potential. Small holes in the Wehnelt and anode plate permit electrons to pass through, and the overall effect of the electrical field configuration produced by these elements is to create an electrostatic lens which focuses the electrons to a point known as first crossover. As discussed above, the brightness at the tip of the cathode is the maximum brightness that can be achieved anywhere in the column. The lenses and apertures can only make trade-offs between current, spot size, and convergence angle, while the brightness is conserved. A magnetic lens will decrease spot size while increasing convergence angle. An aperture decreases convergence angle but also decreases current. We wish to

have a small spot size and small convergence angle, but the cost will be the loss of much of the beam current. Typically, cathodes will emit about one hundred microamps while the final spot may contain ten picoamps. This means that less than one electron in a million successfully makes the trip down the column, as most of the lost electrons are absorbed by the anode plate and the beam limiting apertures.

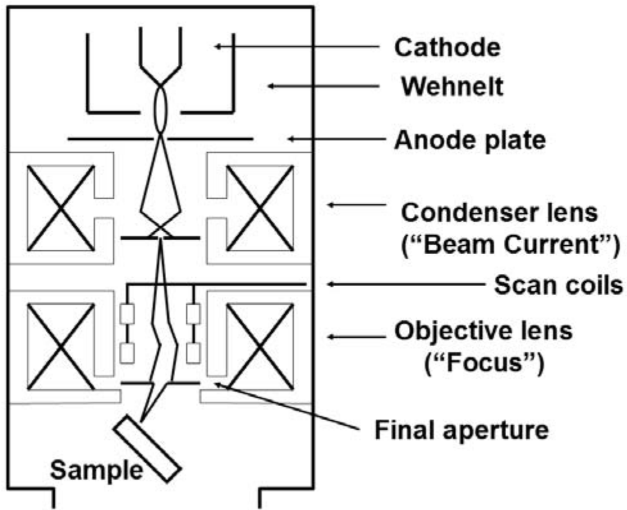


Figure 6: Diagram of an electron column (after Goldstein, et al. [1])

Magnetic lenses will cause a diverging electron beam to converge to a new cross-over, and will reduce the spot size at the new cross-over. Scanning electron microscopes generally have two or more magnetic lenses, each of which can be used to reduce or "demagnify" the beam spot diameter. The first lens is usually called the condenser lens and its most obvious effect is to influence the amount of current that gets through the first aperture. For this reason the condenser lens control is sometimes labeled the "beam current" knob. But it is important to remember that adjusting the condenser lens will trade-off beam current for beam spot size, i.e. adjusting the lens for maximum magnetic field will reduce the beam spot size and also reduce the amount of current that reaches the sample, since loss in the aperture is increased. The condenser lens will also produce a small second-order change in the focus. Some microscopes may have two or more condenser lenses.

The final lens, always called the objective lens, brings the beam to a sharp point known as the final cross-over. This point must intercept the sample surface in order for the image to be in focus. Since small adjustments to the objective lens will have the effect of focusing the image, the objective lens control is sometimes labeled the "focus knob." However, it is vital to realize that the final spot size and hence the resolution of the image will depend on the field strength in the objective lens. The objective lens is most strongly energized when the final cross-over is brought close to the lens, i.e. at "short working distance." However, one cannot arbitrarily set the objective lens for small spot size since the sample must be in focus. Consequently, it is important to place the sample at short working distance in order to achieve small spot size and have good images at very high magnification. On some microscopes, the required working distance for optimum resolution may be as small as 3 to 5 mm. Achieving such short working distance may require careful mounting of the sample to avoid collisions between the lens and the sample holder, but this is vital for obtaining very high resolution images.

Inside the objective lens are two pairs of scan coils for creating the X and Y raster on the sample surface. Two pairs are used to create a double deflection so that the beam exits the final lens near the center of the electron optical axis in order to minimize spherical aberration.

Some modern SEMs have radically new designs such as columns with no intermediate crossovers and compound electrostatic and electromagnetic final lenses. The basic principles of electron optics still apply to these systems though performance may be improved, especially at low voltage. [2]

Beam-sample interaction

When an electron beam enters a sample, the electrons are scattered laterally and gradually lose energy due to collisions with atoms in the sample. The overall result is that the electron energy is deposited in a tear-drop shaped figure known as the "excitation volume" (see Figure 7.) Signals that can be collected from this volume include secondary electrons, backscattered electrons, and characteristic x-rays. X-ray analysis in the SEM is discussed elsewhere and will not be further considered here.

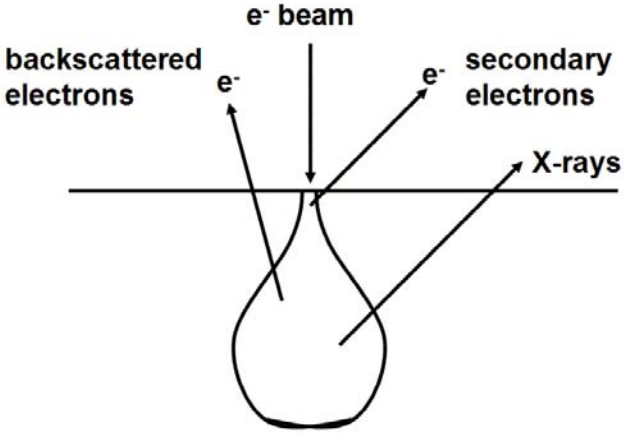


Figure 7: Electron beam – sample interaction volume and interaction products.

Secondary electrons are low energy electrons with very short range, so the secondary electrons that can escape from the sample must originate very close to the sample surface. Since the excitation volume is very narrow near the surface, the

secondary electrons are strongly localized to the entry point of the beam and are useful for high resolution imaging. Secondary electron images are also strongly surface sensitive and show surface topography quite well. Consider Figure 8, as the tilt angle of the incident electron beam increases relative to the local sample surface, the number of secondary electrons will increase. Thus as an electron beam scans across a bumpy surface, the local surface tilt changes and one will see contrast in the secondary electron image which corresponds to the surface topography. Since electrons emitted directly toward the detector will be collected more easily, the surface will be shadowed which further brings out the surface topography. The fact that this process is similar to light scattering off a surface is what causes SEM images to be easily interpreted by the human eye.

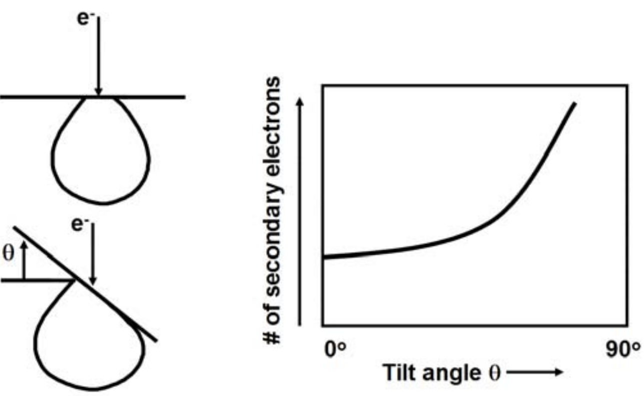


Figure 8: Variation in secondary electron emission leading to topography contrast.

A significant fraction of the incident electrons may be scattered backwards and leave the sample with high energy, in some cases close to the original beam energy. These backscattered electrons tend to originate deeper in the sample and thus are less surface sensitive than the secondary electrons. As a result they may undergo substantial lateral scattering and are generally not good for high resolution imaging. However, backscattered imaging is useful for picking out areas of high atomic number material on a sample since the backscattered electron signal is strongly dependent on the atomic number of the sample. Figure 9 shows a plot of the backscattered electron coefficient for some common semiconductor materials. Note that although Al and Si show very little contrast to each other, Ti, Cu and especially W do show very strong contrast and can be distinguished by backscattered electron imaging, see Figures 10(a) and 10(b). Although backscattered images are generally collected using a special solid state detector placed directly under the final lens, it should be noted that when backscattered electrons exit a sample surface they will create some secondary electrons which are collected by the secondary electron detector. So it is not uncommon that secondary electron images will show some atomic number contrast, and under some conditions as much as 50% of the contrast in a secondary electron image may be due to backscattered electrons.

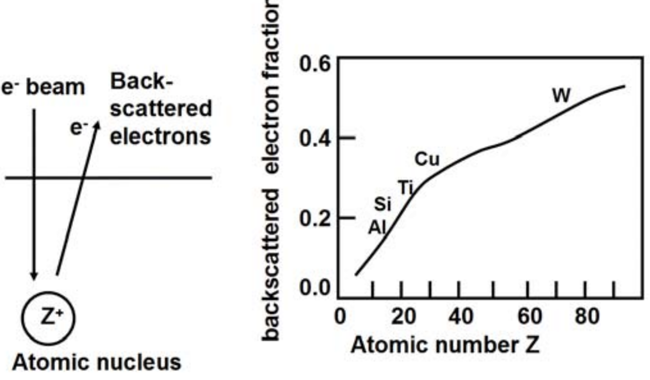


Figure 9: Variation in backscatter emission with atomic number Z.

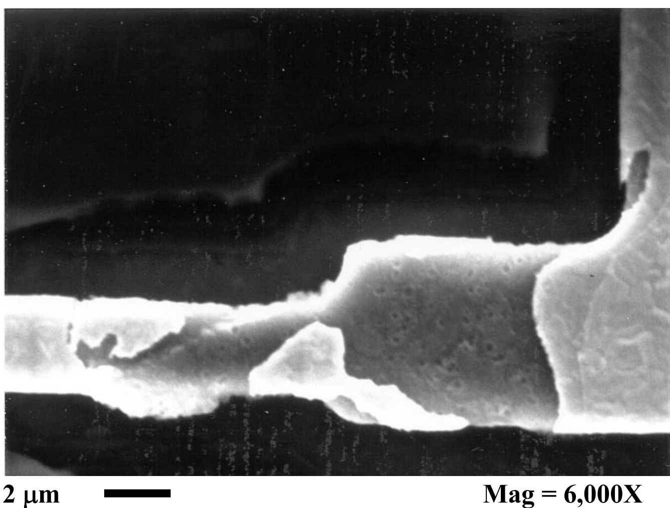


Figure 10(a): Secondary electron image of an aluminum line with a tungsten cap layer damaged by electromigration.

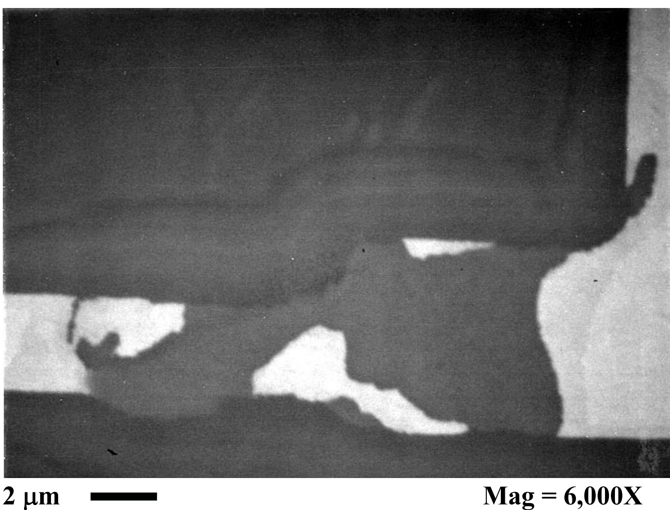


Figure 10(b): Backscattered electron image of same area as Figure 10(a). Note that surface detail is lost but it is very easy to identify the tungsten (bright areas).

Image distortions

The SEM imaging process assumes that the electron beam will raster in a linear fashion on the sample surface. At very large scan angles (i.e. low magnification), the electron beam raster may become non-linear resulting in distortions such as the pincushion distortion shown in Figure 11. The straight grid lines appears to bow inward at the edge of the image because the electron beam is unable to raster as far into the corner of the image field as it should. The opposite effect, known as barrel distortion, can also occur in which straight lines appear to bow outward. These effects are generally negligible at high magnification. Very low magnification is typically used for navigating on a sample and the distortions are not critical in that case. If there is a need to minimize distortions at low magnification one can increase the working distance which will produce a lower beam deflection angle for a given magnification.

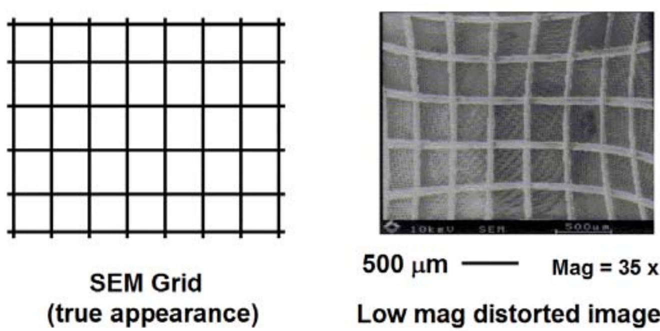


Figure 11: Pincushion distortion in a low magnification SEM image.

Sample charging

One of the greatest challenges in scanning electron microscopy is how to deal with charging problems on insulating samples. If an electron beam enters a properly grounded conducting sample then the excess charge is simply carried away to ground. If a sample is insulating (most commonly glass, polymer, or ceramic) then charge may accumulate in the sample surface. This can seriously degrade the image quality because of electrical field effects on the emitted secondary electrons. Although one might assume that sending electrons into a sample would always charge the sample negatively, in fact positive charging can occur as well, see Figure 12. At very low incident electron energy few electrons are emitted and negative charge accumulates. At intermediate energy, more secondary electrons and backscattered electrons may be emitted than enter the sample so positive charging may occur. At high energy the beam penetrates deeper into the sample and fewer electrons are emitted, so high voltage almost always results in negative charging. There will be two beam energies where the charge is neutralized, E_1 and E_2. If one can find the correct E_2 value then it may be possible to image an insulating sample without seeing any charging. The E_2 value varies from 0.4 keV to 4 keV, depending on the material and on other factors such as surface texture and sample tilt. In general the better insulators have lower E_2 values and show more severe charging effects. E_1 is typically only a few hundred electron volts which is too low in energy to be useful in ordinary electron microscopes.

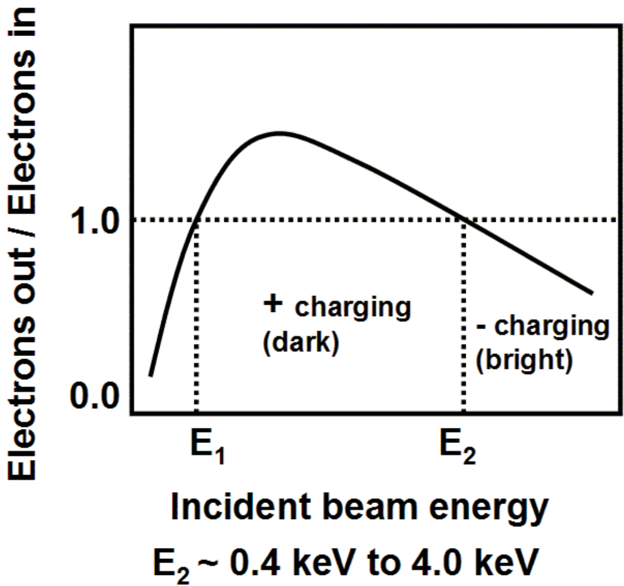


Figure 12: Electron beam charging dependence on beam energy.

Since SEMs are often operated above 5 KeV it is most common to see samples charge negatively. An example of negative charging is shown in Figure 13. The negative charge tends to increase the emission of secondary electrons towards the detector, resulting in a bright, and often completely saturated, image. Even where the image is not saturated, the image quality is poor since the SEM is really imaging the electric field above the sample surface, not the sample itself.

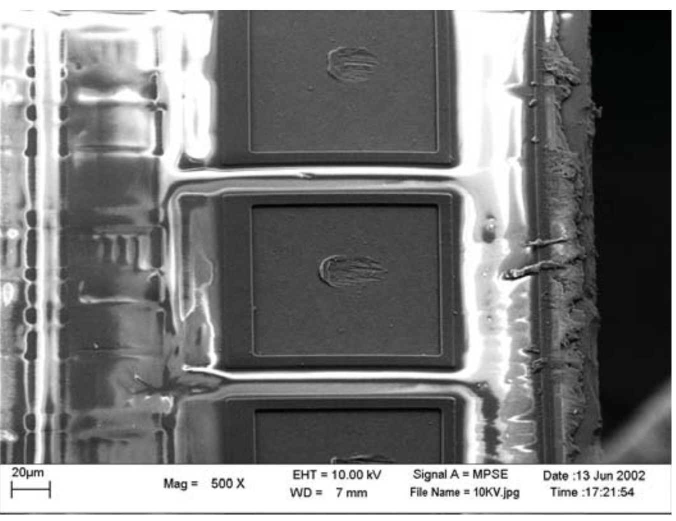
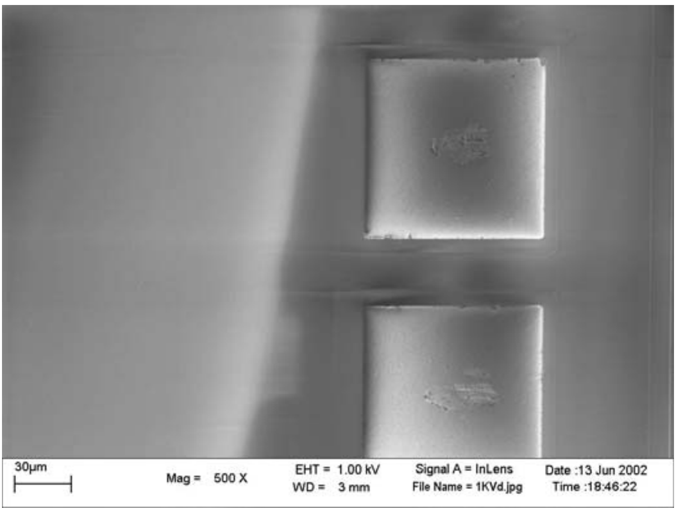


Figure 13: 10 keV image of circuit with glass passivation showing negative (bright) charging. Only the metal bond pads are grounded and thus charge-free.

Positive charging occurs at low voltage and is generally not as severe as negative charging. If the sample surface charges to more than 15 volts, then most of the secondary electrons will be unable to escape the sample surface, and further positive charging will be impossible. Negative charging, on the other hand, can increase to either the electron beam voltage or the sample's surface breakdown potential, whichever comes first. Positive charging produces a dark area on the sample surface since at equilibrium fewer secondary electrons will be able to escape from the sample surface (see Figure 14.)

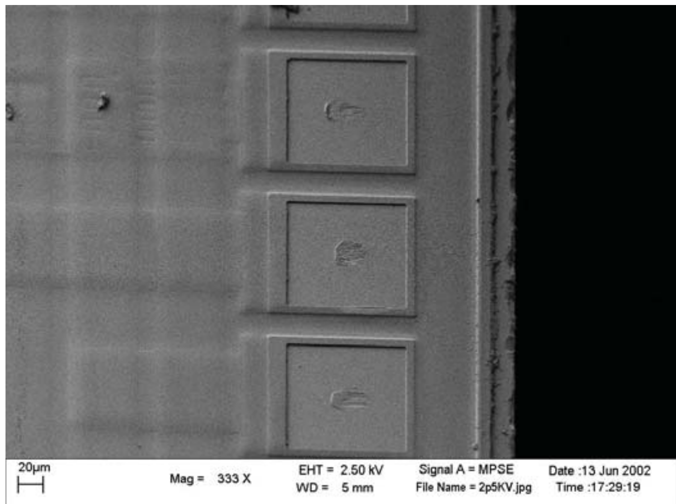


30 µm ▬ Mag = 500X

Figure 14: 1.0 keV image of circuit with glass passivation showing positive (dark) charging. Positive charging tends to be less severe than negative charging and the effects are more subtle.

If the appropriate beam energy is chosen to neutralize the charge, the image will be sharp and of good contrast even on highly insulating samples, see Figure 15. Other methods for decreasing or eliminating charging include increasing the sample tilt to increase secondary electron emission, using an in-lens or backscatter detector, using variable pressure, adjusting beam current, raster area, or raster rate, or sputter coating the sample. Adjusting beam current, raster area, and raster rate may reduce the amount of charging but will not eliminate it. Sputter coating a sample should completely eliminate charging but this has other significant effects on the image and should be done only after serious deliberation. Sputter coating is discussed in more detail in the sample preparation section of this article.

It should be pointed out that charging is not necessarily a bad thing and may produce useful contrast in an image. A technique called "passive voltage contrast" may reveal shorts or floating conductors in integrated circuits. Passive voltage contrast does not require any special equipment, it simply consists of taking an ordinary SEM image and adjusting the beam current, raster size, or raster rate so that open and shorted circuit structures show different charging contrasts.



20 µm ▬ Mag = 500X

Figure 15: Same sample as Figures 13 and 14 using 2.5 keV beam energy, which neutralizes the charge on the glass passivation.

Beam damage

High energy electrons may cause significant damage to delicate samples. This damage may be of several different types. "Raster burn" is often just sample charging which may vanish if the sample is vented to atmosphere. If enough current is put in a small area (for example during a high resolution x-ray map) then carbon may be deposited by the beam, especially if the sample is dirty or the chamber vacuum is poor. A large current applied to a delicate insulator may cause melting or electrostatic discharge. For electronic devices, beam penetration to the gate of a CMOS transistor may cause a threshold voltage shift. This shift can often be annealed out at 150 ^{o}C for one hour.

Practical SEM Techniques

The remainder of the paper will describe practical techniques for getting good images in the SEM.

Sample mounting

The ideal sample mounting method will be quick and easy, mechanically rigid (to prevent vibration that would degrade high resolution imaging), conductive to prevent charging, and will not damage the sample. No single method fulfills all these requirements, but three common methods are listed below:

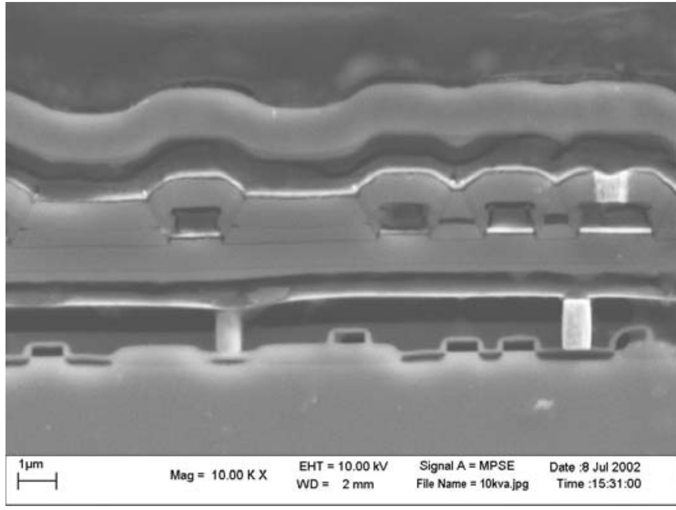
- Carbon paint or silver paint is a common method that is rigid and conductive, but it is slow since the paint must dry. Although it generally doesn't damage samples it is possible for the paint to wick onto the sample and render it useless.

- Spring clips and screw mount holders are fast and easy to use, and are rigid, but they may damage some samples. The author favors this method for many types of samples.

- Metal tape and double-sticky carbon dots are fast and easy, but they lack the rigidity of other methods and are not suitable for high resolution imaging. Double-sticky carbon dots may be so sticky that it is difficult to remove samples without damaging them.

Sample preparation

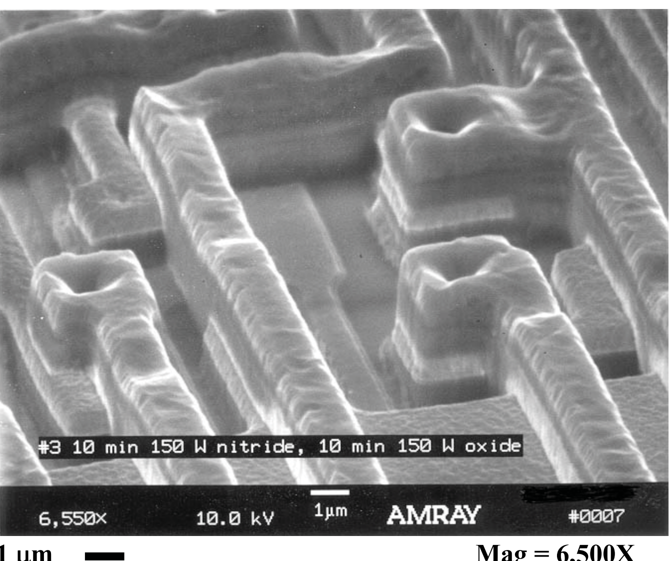
Solid materials are generally opaque to electron beams so SEM images of an intact chip surface such the one shown in Figure 15 are generally not very interesting to a failure analyst. In order to see defects, one must generally take the chip apart in some way. Cross-sections are very popular, both for viewing defects and for seeing how the fabrication process performed. Process engineers often request large numbers of integrated circuit cross-sections, particularly for processes that are new or that have fallen out of specification. A typical cross-section is shown in Figure 16. If the analyst is interesting in viewing the doped areas, a junction stain might be used such as 10:3:1 Acetic:Nitric:Hydrofluoric which is applied for a few seconds under bright light. (A small lamp or even a flashlight will be adequate.) [3] For delineating oxides, a wet or plasma oxide etch might be used which has a slightly different etch rate for different oxide layers. This differential etch will produce surface topography which is easily imaged in the SEM.



1 µm **Mag = 10,000X**

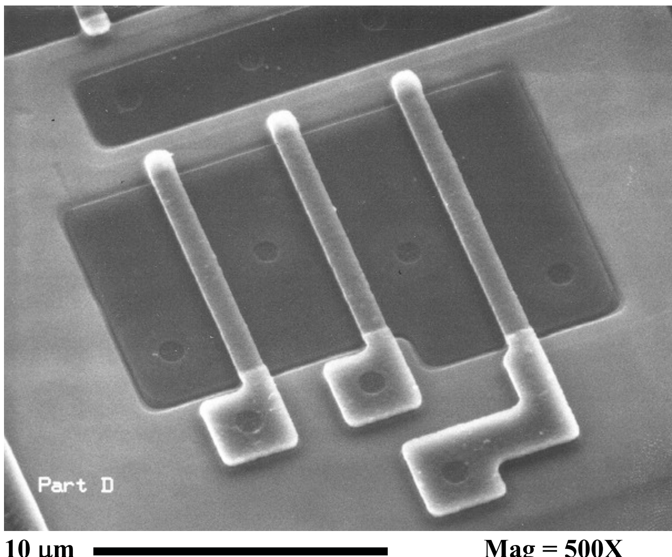
Figure 16: SEM cross-section of an integrated circuit following a junction stain.

Alternately, a circuit can be delayered from the top-down by plasma etch, wet etch, or planar polishing. Examples of RIE deprocessing and wet chemical deprocessing are shown in Figures 17 and 18. Wet chemical etching can be very fast and it provides the best selectivity to different materials, but the etch rate can be difficult to control and will undercut layers because wet etching is isotropic. Plasma etch can be selective to metals or dielectrics, can be directional or isotropic, and is easy to control, but the etching can be slow and may leave residue or "grass" artifacts. Mechanical polishing is relatively insensitive to material composition and thus can provide a planar delayering process, but the method is slow and the resulting flat surface may show little contrast in secondary electron images. In many cases plasma, wet etch, and mechanical polish can be combined to produce the desire result. For example, commonly found silicon nitride passivation layers are impervious to most wet etches. But after the passivation is removed by plasma or mechanical polishing, the circuit can be wet etched down to the poly gate layer as was done for the chip shown in Figure 18.



1 µm **Mag = 6,500X**

Figure 17: SEM of an integrated circuit after plasma delayering. M2, M1, poly and field oxide layers are visible.



10 µm **Mag = 500X**

Figure 18: SEM of an integrated circuit after wet etching down to the poly gate layer.

Sputter coating

Novice electron microscopists may feel a strong urge to sputter coat all insulating sample so that charging will be avoided. However, it is best to avoid sputter coating if

possible. A small amount of charging may actually add some contrast such as passive voltage contrast that is useful in understanding the sample. Sputter coating may interfere with x-ray analysis and image preparation methods such as plasma etching or chemical staining. At very high resolution, sputter coating may obscure important surface detail and introduce artifacts such as contrast from sputter grains. The high magnification secondary electron image of an aluminum line shown in Figure 19 contains these artifacts. The only "real" sample features are the large crease in the upper right corner and the small particles in the lower left corner. The small wavy lines are sputter coating artifacts which have nothing to do with the original sample surface.

Most SEM labs will have access to a small sputter coater that can deposit a few nm of a conducting material on a sample surface. Au/Pd has a small grain size and is very popular, but Cr, Pt, Ir, and C are also used. Sputtered Ir has become popular recently because it has a very smooth surface. Carbon is generally less effective than the other materials but has the advantage that is doesn't interfere with x-ray analysis and can be removed with an oxygen plasma. About 10 nm of metal is required to produce a continuous coating, but as little as 2 or 3 nm may produce significant changes in the image.

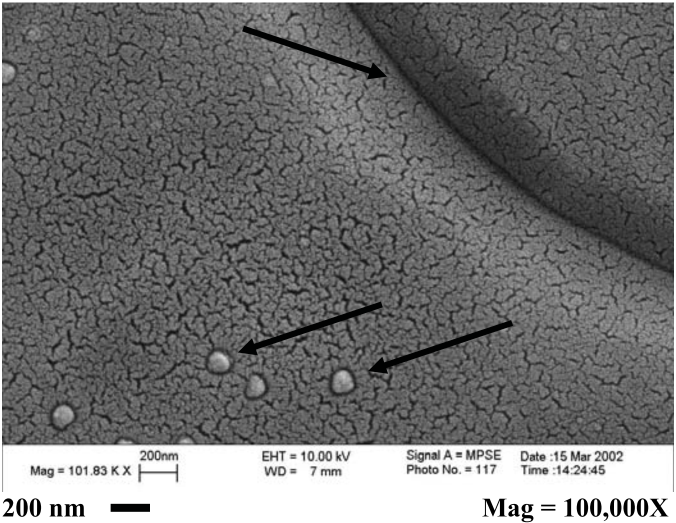


Figure 19: SEM image of an aluminum line with 10 nm of Au/Pd sputter coating. All the small dark wavy lines are sputter grain artifacts.

Sputter coating actually does three things which benefit the image. First, it does eliminate sample charging. Second, it will increase the secondary electron emission from the sample surface which increases the signal and improves signal-to-noise in the image. Third, it decreases the range of secondary electrons which increases the surface sensitivity of the image. The latter two effects will often improve the image of both charging and non-charging samples, even if a very thin coating (< 5 nm) is used. Thus thin sputter coatings are often applied to improve imaging, but it is the coating that is being imaged, not the sample.

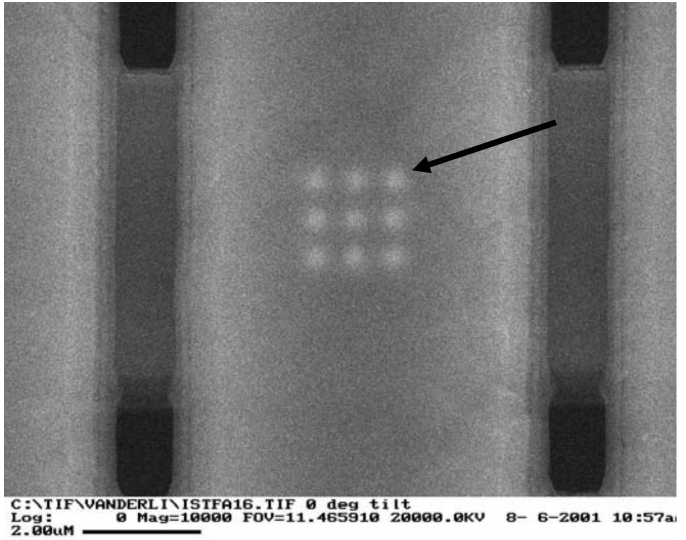
Beam voltage

A critical decision in the operation of an electron microscope is what beam voltage to select for a given sample. Novice users often read the instrument specification sheet and observe that the electron beam spot size is a minimum at high beam voltage. They therefore may choose a very high beam voltage. In fact, low voltage often produces better images. Although high voltage does improve the spot size, it also greatly increases the range of the electrons in the sample and subsequent scattering (see Table 3.) At 1.0 kV beam voltage the very small range of the electrons results in a secondary electron signal that is very surface sensitive and well localized to the electron probe. Furthermore, low voltage may reduce charging problems. In general it is best to use the lowest possible beam voltage that gives a good image.

Table 3: Beam voltage vs. spot size and electron range in aluminum. Range is calculated from the Kanaya-Okayama formula and spot size is for a LEO 1550 FE SEM.

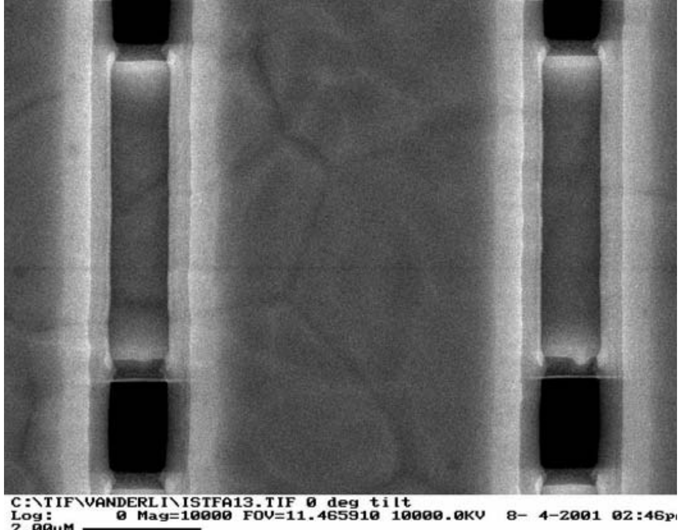
Beam Voltage	Spot size (nm)	Range in Al (μm)
1	2.4	0.028
3.5	1.5	0.22
5	1.3	0.41
10	1.1	1.32
20	1.0	4.19
30	1.0	8.24

High beam voltage can produce a sharp image but one that lacks surface detail (see Figure 20(a).) At 20 kV the image lacks surface detail, but tungsten plugs buried one micron deep under the aluminum metal become visible (although the plug images are fuzzy due to electron scattering in the sample.) On the other hand, Figure 20(b) at 10 kV shows better surface detail while maintaining high resolution. Figure 20(c) at 1 kV has less resolution but does show some additional surface detail. In particular, black contamination marks can be seen which were not visible at higher energy. These black marks are likely a very thin layer of carbon contamination.



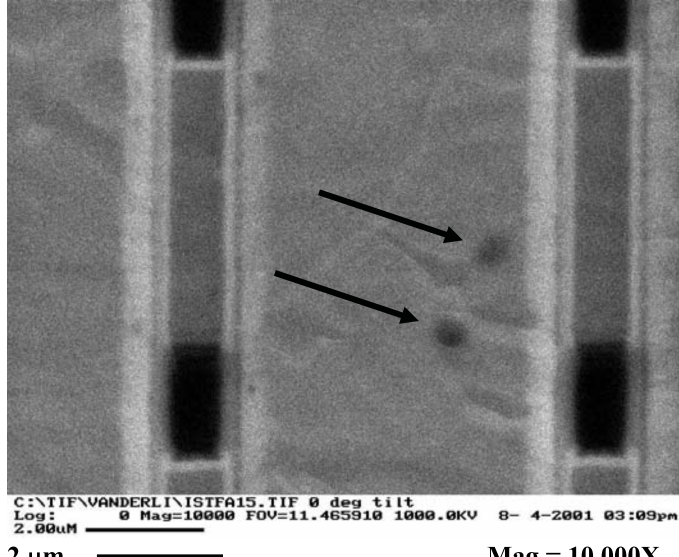
2 µm **Mag = 10,000X**

Figure 20(a): SEM image at 20 kV beam voltage of tungsten plugs underneath an aluminum line.



2 µm **Mag = 10,000X**

Figure 20(b): SEM image at 10 kV beam voltage of an aluminum line.

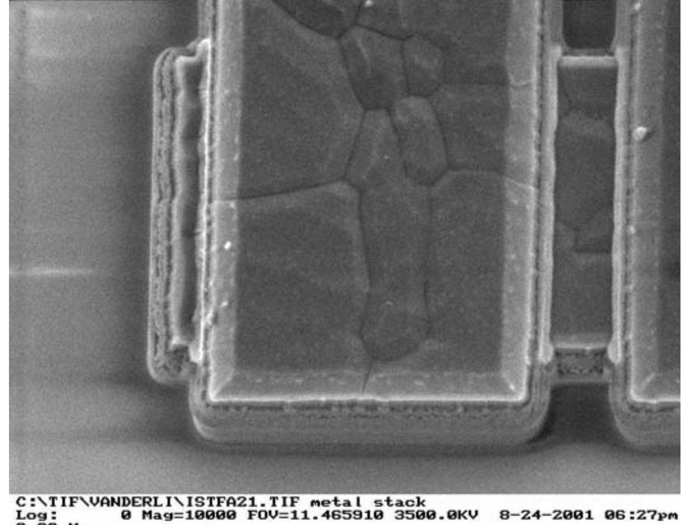


2 µm **Mag = 10,000X**

Figure 20(c): SEM image at 1 kV beam voltage of an aluminum line.

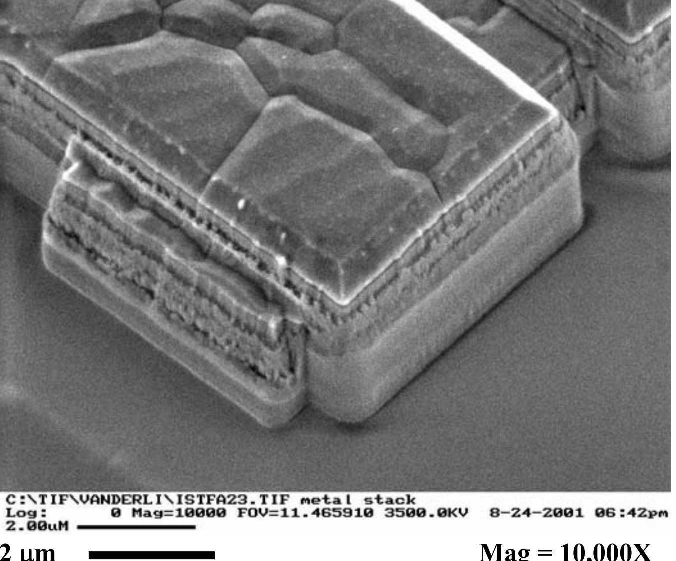
Sample tilt and image composition

SEM sample stages are designed to tilt the sample towards the electron detector and in most cases it is best to take photos with at least 15 degrees of tilt. Tilting the sample increases the secondary electron emission and also increases the number of secondary electrons captured by the detector. Furthermore, tilting the sample makes it easier to interpret surface topography and three-dimensional structures. Consider Figure 21 with no tilt, and Figure 22 with 30 degrees tilt. One disadvantage to tilting the sample is that the vertical axis is foreshortened so it becomes more difficult to make metrology measurements.



2 µm **Mag = 10,000X**

Figure 21: SEM image with no tilt.



2 μm **Mag = 10,000X**

Figure 22: SEM image at 30 degrees tilt and 45 degrees rotation.

Focus and astigmatism correction

The adjustment of focus and astigmatism corrections (or “stigmation”) are the most difficult skills for a novice electron microscopist to develop, but they are essential for obtaining the best high resolution images. Some tips for developing these skills are provided below.

In should be noted that focus and stigmation techniques apply to both SEMs and focused ion beam (FIB) systems, so skilled electron microscopists are often adept at getting good images in the FIB as well. The main difference is that the FIB is constantly destroying the object being imaged so it is essential to move quickly in focusing and stigmating a FIB image. It has been observed that a FIB beam optimized for imaging is not necessarily the best beam for cleanly cutting a surface. Optimizing a FIB beam by focusing an image will narrow the spot size that contains the majority of the current, but it may increase the beam tails which can cause rounding at the edge of a FIB cut. Cleaner cutting may be achieved with a de-focused ion beam that has a broader central peak but smaller beam tails.

Ideally, the magnetic lenses in a SEM will be rotationally symmetric and will bring all electrons back to the optical axis at the same point in space, the final crossover. Astigmatism, the rotational asymmetry of a lens, causes electrons in different planes to converge at different focusing points along the optical axis. Astigmatism may be due to small defects in the lens, dirt on the apertures, or stray electrical fields in the chamber. Astigmatism is corrected by applying small additional magnetic fields which are controlled by knobs typically labeled X and Y “stig”. (One occasionally sees stigmation controls labeled as theta and magnitude, this is simply the difference between applying the correction in polar vs. Cartesian coordinates.) Unfortunately, the proper amount of astigmatism correction varies with beam voltage, lens settings, and working distance, so changing those parameters will require the image to be stigmated again. If the sample is highly charging it may produce a very large amount of astigmatism that cannot be corrected by the stigmators.

Figure 23 illustrates astigmatism, where electrons in the vertical plane will converge at one point called “under focus”, while electrons in a horizontal plane will converge to a second point called “over focus”. The “true focus” in the absence of astigmatism would be roughly midway between these two points. Sweeping the objective lens through the under focus or over focus condition will appear to bring the sample into focus, but it will not be a proper focus. The image will lack symmetry and sharp edges will appear sharp in only one direction, see Figures 24(a) and 24(b). The ability to distinguish under and over focus from true focus is critical to the stigmation process.

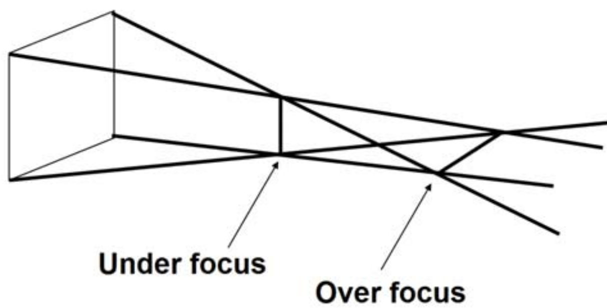
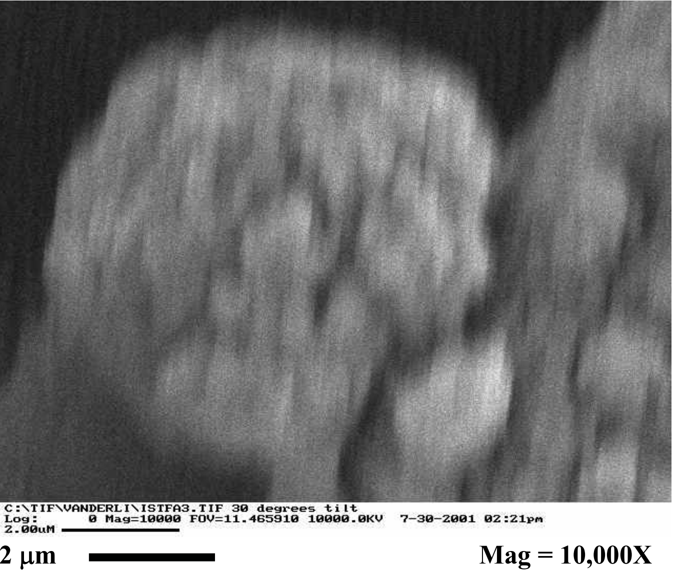


Figure 23: Illustration of the effects of astigmatism on the final focus of an electron beam.



2 μm **Mag = 10,000X**

Figure 24(a): Illustration of under focus, with sharp edges only in the vertical direction.

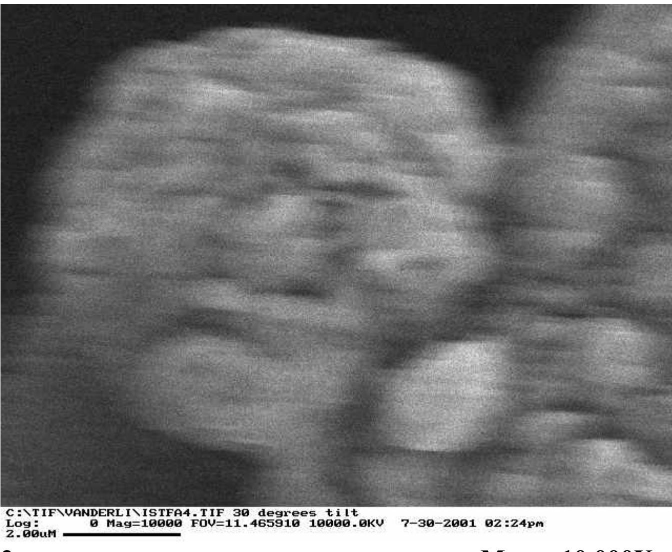


Figure 24(b): Illustration of over focus, with sharp edges only in the horizontal direction.

The procedure for adjustment of fine focus and astigmatism is to sweep the focus back and forth and attempt to identify the under focus and over focus positions of the objective lens control (i.e. the “focus knob”.) Once those two points are identified, the lens control should be placed exactly in between those two positions. The image will be fuzzy in an overall symmetrical fashion, see Figure 24(c). Now the X and Y stigmators are adjusted to obtain the best image. The result will be a much sharper image, see Figure 24(d). At this point, it may be necessary to further adjust the focus and again re-stigmate the image. Once a good image is obtained for this magnification, the magnification can be increased and the focus and stigmation procedure repeated.

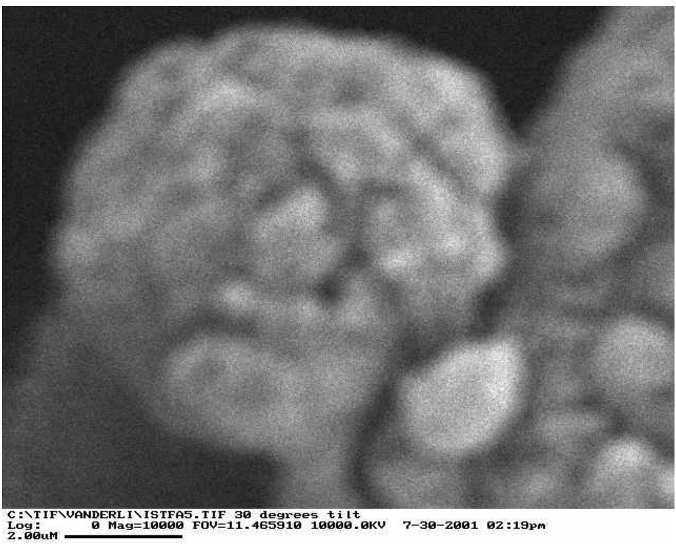


Figure 24(c): True focus, prior to astigmatism correction.

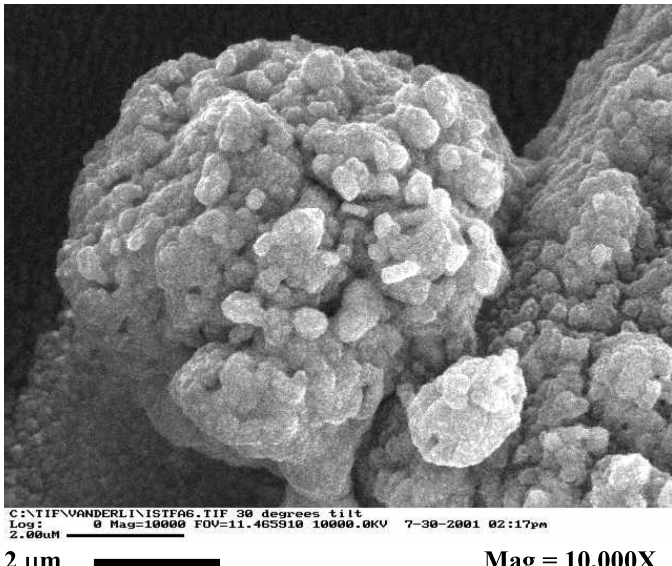


Figure 24(d): True focus, after astigmatism correction.

One should also align the electron source, apertures, and other column components as the magnification is increased. Since the alignment procedure varies with each instrument, the reader is referred to his or her owner’s manual for more details. In general, alignment is optimized by “wobbling” the beam, which is just an automated feature for sweeping the focus back and forth. Since magnetic lenses have the property of rotating the electrons around the optical axis, this sweep will appear to rotate the image. If the column is in good alignment, the rotation will be centered at the center of the image. If the alignment is incorrect, the rotation will be off-center, and a given feature will appear to displace back and forth when viewed at high magnification. Alignment should be adjusted until this displacement is minimized.

The focus, alignment, and stigmation can be performed at progressively higher magnification until the maximum useful magnification is reached. One often focuses and stigmates at much higher magnification than the final desired image so that the effects can be easily seen. For example, it may be necessary to focus, align, and stigmate at 300,000x magnification in order to obtain a good 100,000x photo. Selecting a good object to focus on is absolutely critical. The ideal object should be circular. If nothing circular is available, one can use a square or rectangular object, or the intersection of two orthogonal lines. Integrated circuits often have long straight lines (see Figure 25.) One should avoid focusing on a straight line since it will be possible to identify only under or over focus. If one attempts to stigmate using under or over focus rather than the true focus, it will ***impossible*** to properly stigmate the image. It can be tempting to focus on a dust particle, but many dust particles are insulators and may charge up. Any charged feature should be avoided since charging will itself add astigmatism as mentioned above.

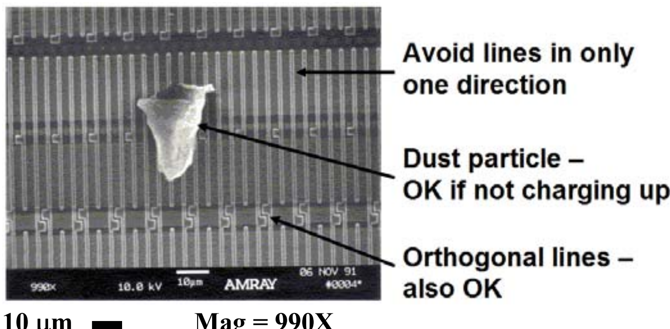


Figure 25: A good focus object should be circular or square and should not be charged.

Once a good image is obtained at high magnification, then the beam spot has been optimized at the sample surface and the image will be in focus at any lower magnification. One should therefore focus at high magnification and *not* adjust the focus again when taking a lower magnification photo. If one wishes to move around the sample to search for features of interest, it may be necessary to make small only corrections to the focus, and the alignment and astigmatism should not require significant adjustment unless the beam voltage, apertures, or working distance are changed.

Typically, the SEM stage will tilt toward the secondary electron detector and the image raster will be aligned such that the detector is at the top of the image. Thus the tilt axis will coincide with the x-axis of the stage. Therefore if the sample is tilted, the top of the image will be slightly further away and the bottom of the image will be slightly closer to the objective lens than the center of the image. What this means is that if you move the sample along the Y-axis the proper focus changes rapidly, but there will be little or no change in focus when moving the sample along the X-axis. In many cases, there may not be any good focusing objects in the field of view desired for the final image, or one may be concerned that raster burn from the high magnification focusing may ruin the final image. In this case, one can select a different focus object which is at some distance along the stage X-axis, and then translate the stage back in X for the final image.

Dynamic focus and image correction

Although the SEM has good depth of focus, at high magnification it may be difficult to keep a highly tilted sample completely in focus. For flat samples, it is possible to apply a “dynamic focus” feature in which the SEM automatically sweeps the beam at a angle to match the tilt of the sample surface. To apply dynamic focus, use a small area raster box to focus the center of the image and then move the box to the upper or lower edge of the image and focus that edge using the dynamic focus correction.

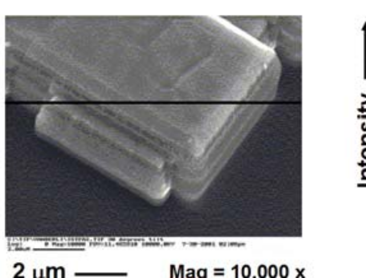
When the sample is tilted, the vertical measurements of the image of a flat sample will be foreshortened due to simple geometry by a factor of Cos θ. Therefore when making critical dimension measurements, one should always align the measurement with the tilt axis of the stage which should be in the horizontal direction. It is possible to correct for the foreshortening by applying “image correction” which simply expands the vertical axis of the image by 1/Cos θ. One can also do the same thing to a digital image off-line with a graphics program. Note that this image correction is only valid for a flat sample and any three-dimensional objects will be distorted. For sample, a sphere on the sample surface will appear elongated if image correction is applied to it.

Raster alignment

SEM users will notice that there are two distinct ways to make objects in the image appear to rotate: they can rotate the sample, or they can rotate the beam raster. In general it is best to keep the beam raster aligned to the stage axes. When working distance is changed the objective lens will introduce some image rotation which needs to be corrected. The correct procedure is to sweep the stage motion back and forth in the X direction and then adjust the raster rotation so that objects on the sample surface will move back and forth in the same horizontal plane in the image. Once that is done, one can then rotate the sample to produce a visually appealing image. If instead the raster rotation is used to compose the image and the raster is not aligned to the stage, the result may be strangely distorted images in which rectangular objects appear to be parallelograms.

Brightness and contrast

Novice users will note that there are two controls that affect brightness of the image, “brightness” and “contrast”. At first glance these may appear to have similar effects on the image, i.e. both will increase or decrease how bright the screen is. Adjusting these controls is made much easier if the microscope has a feature called “line scan” (see Figure 26.) The line scan is simply a plot of the image intensity on a line across the sample. Ideally, the image intensity will vary across the full dynamic range of the monitor, with no white or black areas indicating that the image is saturated at either high or low intensity. “Contrast” will expand or decrease the amount of variation in intensity. Too little contrast will result in a washed-out image with little so see, whereas too much contrast will result in image saturation at both high and low levels, with large areas that are all white or all black. “Brightness” increases or decreases the overall signal level. In general, one wishes to maximize the amount of contrast so that the full range of intensities are used, but without saturating the image at either the high or low end. The brightness is then adjusted so that the overall image intensity is correct.



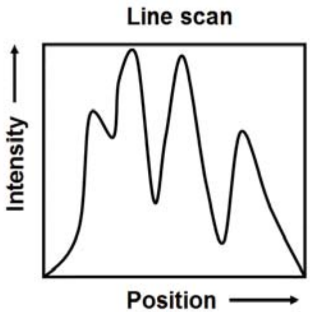


Figure 26: Illustration of how a line scan can be used to help adjust brightness and contrast.

Some images have a very large intensity variation, but there may be little variation on a feature of interest which lies near the black or white end of the intensity scale. In this case one can use Gamma, which is a non-linear contrast enhancement effect. Differential enhancement can be added to either the dark or bright end of the intensity scale to enhance these features.

The reader should note that there is no such thing as the "correct" choice of contrast and brightness. It really depends on what features the user wishes to display. It may be perfectly acceptable to have much of the image saturated if that allows a critical defect to be shown in good contrast. Like other forms of photography, scanning electron microscopy is an art form and each user may make different choices in how to image a sample.

Scan speed and image quality

The quality of an image depends on both the beam spot size and also the signal to noise ratio of the image, the latter depending on how many electrons were collected for the image. A very small beam that has little current may have a very good nominal resolution, but the image will be too noisy to contain useful information. A typical full-screen scan at TV rates will have very poor image quality which can make navigation difficult. The signal to noise can be improved either by (1) decreasing the raster area (i.e. a small area raster box) or (2) increasing the scan time (i.e. "slow scan"). Small area raster boxes are generally used for focus, alignment, and stigmation, while slow scan is generally used to capture the best archival images. Slow scan frame times can vary from a few seconds to more than a minute, but long scan times may be limited by stage drift or charging on the sample surface.

STEM-in-SEM

Modern field emission SEMs have a nominal spot size as small as 1 nm. However, it is very rare that 1 nm resolution is realized in everyday imaging. Typically a SEM service engineer will demonstrate the microscope resolution by using very high beam voltage such as 30 kV to image a special "resolution standard" which consists of gold islands on a carbon substrate, commonly referred to as a "gold-on-carbon" sample. See Figure 27 which is at extremely high 500,000X magnification. Imaging a more typical sample such as an electronic device will generally produce an image with much lower resolution that may not meet the needs for modern failure analysis and metrology.

It is sometimes possible to achieve very high resolution by sputter coating a very thin layer of metals such as Au, Au/Pd, Cr, or Pt. Typically, such coatings result in a better image of the sample surface, not only because they eliminate charging but because they also increase the surface emissivity of the sample and reduce the secondary electron range. About 10 nm of metal is required to produce a continuous sputtered film, but this amount of metal may obscure surface detail and cause image artifacts from the grain structure of the sputtered metal. It is therefore unacceptable to sputter coat for imaging at 1 nm resolution (see Figure 28.)

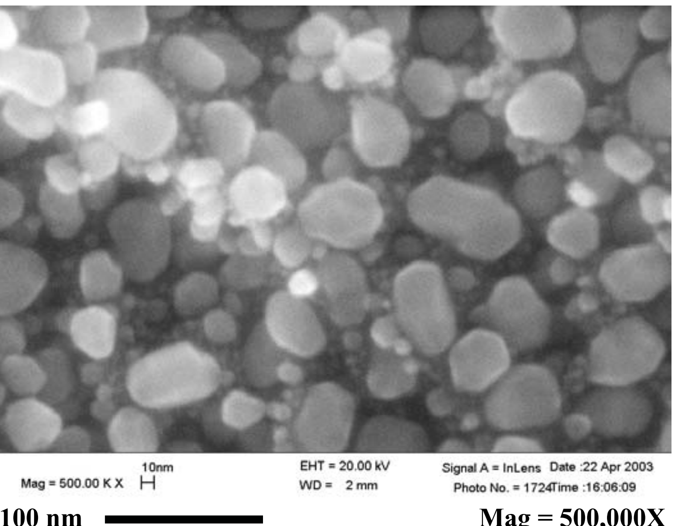
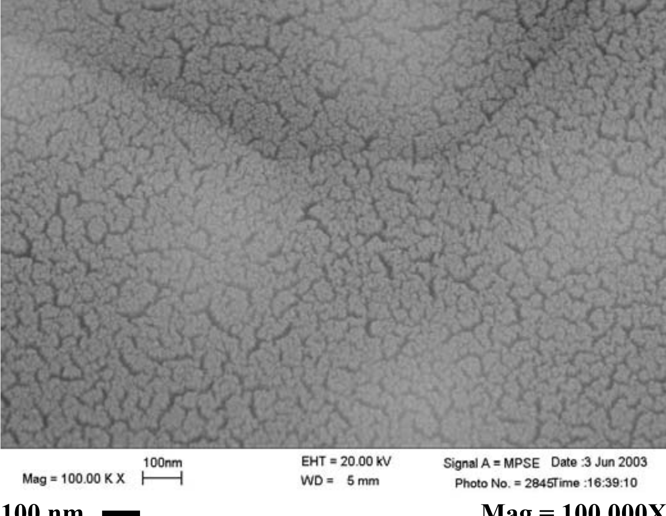


Figure 27: Secondary electron image of gold islands on a carbon substrate shows that very high resolution SEM images are possible under the right conditions.



100 nm **Mag = 100,000X**

Figure 28: Secondary electron image of aluminum metal sputter coated with 10 nm of Au/Pt. The image has high resolution but the original surface detail is obscured and the sputter coating grains introduce artifacts seen as jagged dark lines.

In order to achieve ultra-high resolution in the SEM, three conditions must be satisfied: [4]

1. The electron beam must be finely focused to a small spot at the surface of the sample.

2. The electron beam current must be high enough to produce sufficient signal so that visible contrast is achieved in the image.

3. The signal used for imaging must originate very close to the impact area of the beam on the sample.

Conditions 1 and 2 are satisfied when the electron source is operating at high brightness, which is maximized when the electron source is at high beam voltage. However, high energy electrons have a large range in most materials. When the electrons are backscattered from deep in the sample towards the surface, they may exit the sample at a significant distance from the probe location (see Table 3.) Thus condition 3 is not satisfied at high voltage since the backscattered electrons are not closely correlated with the probe. These electrons also produce secondary electrons as they exit and this further reduces the spatial correlation between the secondary electrons and the probe. If the beam voltage is reduced so as to decrease the range of the electrons, then the brightness of the electron source is reduced and it is not possible to achieve a small spot with sufficient beam current (conditions 1 and 2). For the special gold-on-carbon sample all three conditions are satisfied at high beam voltage because gold has a very high backscatter coefficient and carbon has a very low backscatter coefficient. Thus incident electrons are usually either reflected from the gold surface or lost in the carbon substrate. However, it is very difficult to find ordinary samples that will allow the three conditions to be met.

Table 3: Electron beam spot size (per the manufacturer [8]) and penetration range (from the Kanaya-Okayama formula [9]) as a function of electron beam energy.

Beam energy (keV)	Spot size (nm)	Electron penetration range in aluminum (microns)
1	2.4	0.028
3.5	1.5	0.22
5	1.3	0.41
10	1.1	1.32
20	1.0	4.19
30	1.0	8.24

One solution to satisfying the three conditions for high resolution imaging is to create a very thin sample ($\leq$ 100 nm thick) and image using the transmitted high energy electrons. Electrons will scatter very little as they pass through a thin film. Thus the detected signal is highly localized, even at high beam voltage where the electron optics produce the best spot size. This approach enables high resolution SEM images that look very similar to TEM images. FIB techniques enable the creation of “TEM samples” less than 100 nm thick much easier and faster than in years past.

A typical “STEM-in-SEM” holder is shown in schematic in Figure 29 and in an optical photo in Figure 30. [5,6] The thin sample is mounted on top of a graphite block. After passing through the thin film, the electron beam travels down a narrow hole drilled in the graphite block. By absorbing the most highly scattered electrons, the graphite block collimates the beam which improves the image resolution. After passing through the collimator, the electrons strike a pair of gold reflectors that create secondary electrons which are collected by the ordinary in-chamber secondary electron detector. Alternately, one can invest in a dedicated STEM-in-SEM detector which can be added to some SEM models. [7]

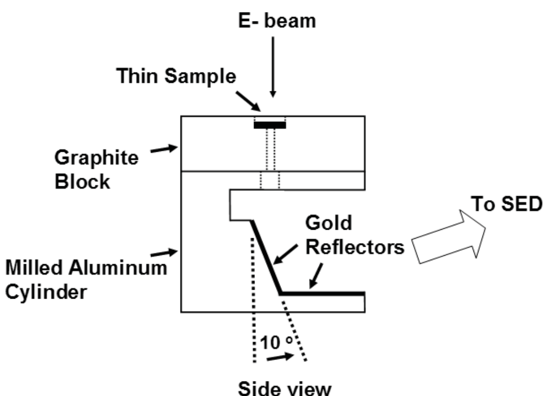


Figure 29: Design for a "STEM-in-SEM" sample holder. The overall height is 2.5 cm.

STEM-in-SEM samples are typically mounted onto a standard 3 mm diameter TEM disk, which is then placed into a cylindrical hole in the top of the STEM-in-SEM holder and held in place by a retaining ring. However, the STEM-in-SEM holder can be easily modified for a variety of sample shapes and lengths. When thin samples are prepared by FIB, the 3 mm limitation on sample size imposed by standard TEMs can significantly increase the difficulty of the preparation. With STEM-in-SEM, the sample can be of virtually any size. It is, however, crucial that the sample be mounted in a clean and mechanically rigid fashion. Metallic tape is often used to mount SEM samples, but the tape will allows unacceptable sample drift. Carbon paint is also used to mount SEM samples, but with STEM-in-SEM the paint may transfer carbon contamination to the thin section. Therefore, a mechanical mounting method such as spring clips or retaining rings is recommended.

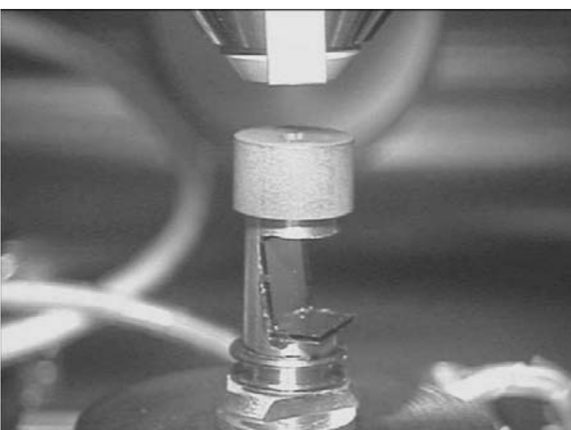


Figure 30: Optical photo of the STEM-in-SEM sample holder mounted to the SEM stage. The pole piece of the objective lens is visible above the sample holder.

Transmission electron microscopes typically operate at 200 kV or higher, so it may be somewhat surprising that an SEM's electron beam voltage of 30 kV or less can be used to view a sample in transmission. While it is often said that TEM samples must be thinned until they are "electron transparent," this is a misleading statement. The electron range of a 30 keV electron in aluminum is more than 8 microns. TEM samples are typically thinned to less than 100 nm thick so that most of the electrons undergo no more than one scattering event. Accordingly, TEM imaging is typically done with electrons that have not undergone any scattering events (bright field) or those that have been scattered exactly once by the lattice (dark field). At 200 keV the mean free path for elastic scattering is large (~350 nm) compared to the thickness of a 100 nm thin film. For STEM-in-SEM at 30 keV the elastic mean free path in aluminum is 17 nm for a scattering angle of 2 degrees or greater, and an electron will typically be scattered several times when passing through the sample. This results in substantially degraded lateral spatial resolution.

To estimate the resolution degradation, a Monte Carlo simulation of electron trajectories was conducted using the commercially available "Electron Flight Simulator" program [10] as shown in Figure 5. The Monte Carlo simulation shows that 30 keV electrons will be scattered into a cone of half-angle 8.5 degrees, resulting in a beam diameter of about 30 nm as the beam exits a 100 nm thin film. In order for a raster beam system to produce a sharp image, electrons should collect information only from the area very close to where the beam enters the sample. If the beam scatters 30 nm laterally within the sample, then contrast will be collected from materials that are far from the electron beam position, and images will be seriously degraded. Thus in one design [5] a graphite collimator was placed in the STEM-in-SEM holder below the sample to absorb the most highly scattered electrons. The design of the collimator necessitated a trade-off of large transmitted signal and ease of alignment with the need to eliminate the most highly scattered electrons that will degrade image resolution.

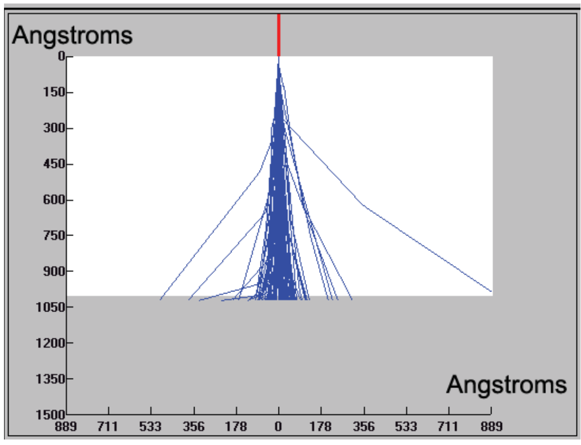
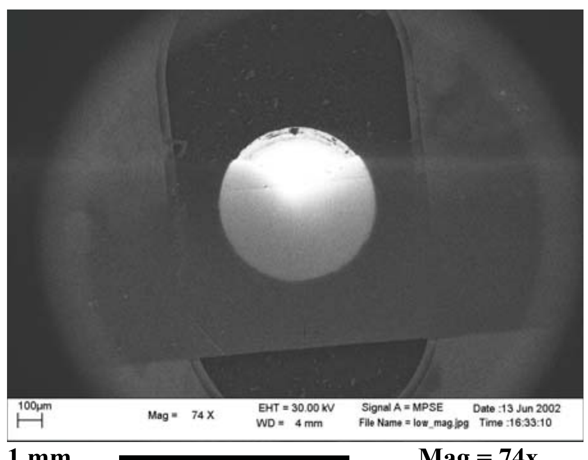


Figure 31: Monte-Carlo simulation of the trajectories of 30 keV electrons passing through a 100 nm aluminum thin film.

Procedure

Figure 32 is a low magnification STEM-in-SEM image. The image shows a conventional TEM sample, consisting of a thinned strip of silicon attached to a 3 mm copper grid with an oval hole. The bright circle in the center of the picture corresponds to the hole in the graphite collimator which

allows electrons to penetrate to the gold reflectors. If necessary, the holder can be tilted to bring the area of interest to the center of the bright circle, and the entire holder is then rotated to maximize the collection of secondary electrons from the gold reflectors to the secondary electron detector. Once the holder is aligned, high resolution STEM-in-SEM images can be acquired by selecting a thin area of the sample and simply increasing the magnification, and then focusing and stigmating in the usual fashion.

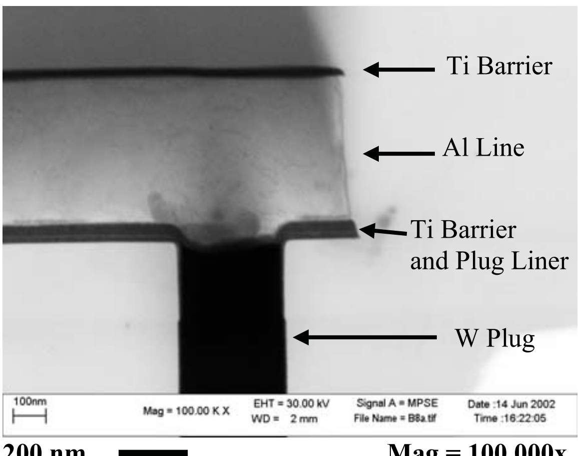


1 mm **Mag = 74x**

Figure 32: Low magnification STEM-in-SEM image showing the TEM sample.

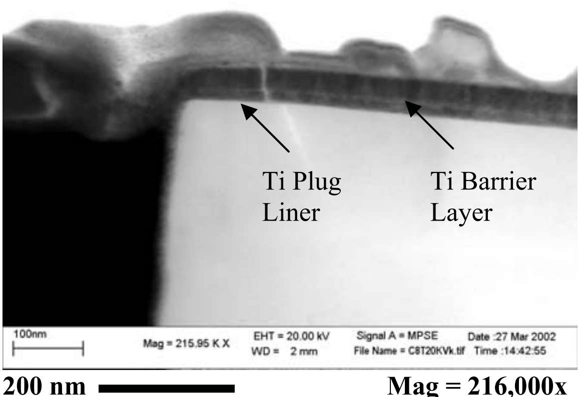
Semiconductor cross-sections

Figures 33, 34, 35, and 36 are STEM-in-SEM images of typical CMOS structures such as aluminum lines, tungsten plugs, and poly lines with sidewall spacers. These images were taken at 20 kV or 30 kV beam voltage and 2 mm working distance. The STEM-in-SEM images have a similar appearance and contrast to TEM images. They show internal grain structure on the poly-silicon, the aluminum, and the barrier metals. Close examination of the STEM-in-SEM images demonstrates a resolution of approximately 2 nm which is far superior to that achieved by traditional SEM images of these structures. The STEM-in-SEM image contrast is based on differences in atomic number, density, and crystallographic orientation, all of which influence how electrons scatter away from the detector. [11]



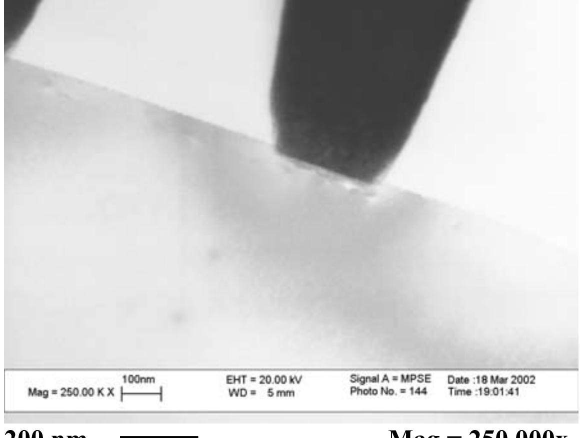
200 nm **Mag = 100,000x**

Figure 33: STEM-in-SEM image of an aluminum line connected to a tungsten plug contact. The barrier layer and the plug liner are visible.



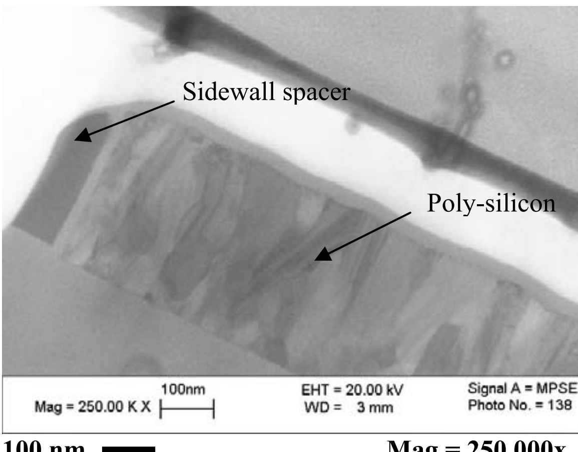
200 nm **Mag = 216,000x**

Figure 34: Higher magnification STEM-in-SEM image resolving a titanium plug liner and barrier layer.



200 nm **Mag = 250,000x**

Figure 35: STEM-in-SEM image of a tungsten plug contacting silicon.

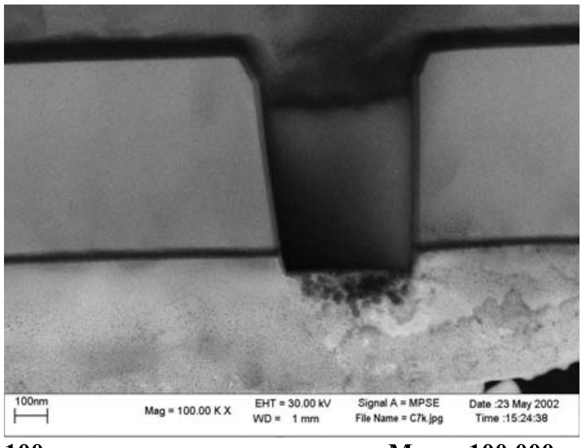


100 nm **Mag = 250,000x**

Figure 36: STEM-in-SEM of broad poly-silicon line with sidewall spacer.

The STEM-in-SEM holder can be tilted so that the imaged area is outside the bright circle that corresponds to the entrance of the collimator, see Figure 37. Under these conditions, only the scattered electrons are able to pass through the collimator and reach the gold reflectors, in effect creating a "dark field" image. Note that the tungsten which is dark in Figures 33 and 34 now shows some bright areas. The overall image has degraded resolution, but the contrast on the material just below the tungsten-silicon interface is enhanced. This type of dark field imaging could be used to enhance the contrast from highly scattering materials such as inter-metallic grains at material interfaces.

STEM-in-SEM also allows for very high spatial resolution x-ray maps. See the Chapter on Energy Dispersive X-ray Spectroscopy for details.



100 nm **Mag = 100,000x**

Figure 37: "Dark Field" STEM-in-SEM image.

Environmental SEM (ESEM)

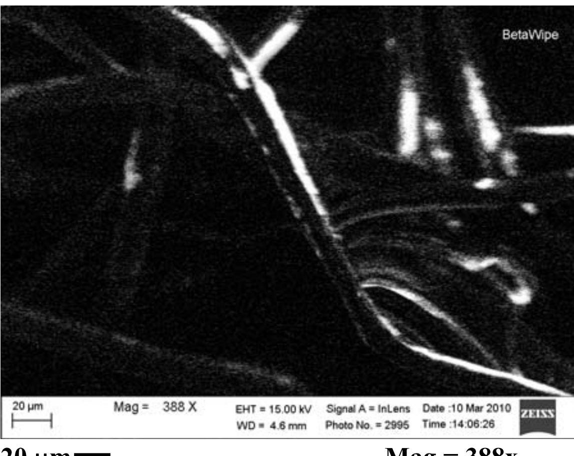
In most SEM instruments the sample chamber is held at high vacuum, typically 10^{-4} Pa (~ 10^{-6} Torr) or better to minimize scattering of the electron beam and to prevent contamination of the sample by beam-deposited materials. However, in some cases it is desirable to raise the chamber pressure. This can be achieved by "differential pumping" so that the electron source remains at low pressure while the sample chamber is held at much higher pressure. [12] A true "environmental SEM" can achieve chamber pressures greater than the vapor pressure of water at room temperature (~20 Torr or 2,700 Pa). This is particularly useful for looking at biological and mineral samples in a fully hydrated state. However, lower pressures have found applications to imaging insulating materials such as glass and plastic since the gas tends to dissipate charge and reduce sample charging problems. As little as 1 to 100 Pa (~ 10^{-2} to 1 Torr) is useful in such applications and some variable pressure SEMs are designed to achieve only that pressure range. SEM images of a highly insulating cleanroom textile with and without variable pressure are shown in Figures 38 and 39.

The traditional Everhart-Thornley detector is not suited to variable pressure since the high voltage in the detector will ionize the gas resulting in electrical breakdown and discharge. ESEM was originally performed using solid state backscatter detectors, but that limited the image quality. ESEM really came of age with the development of Gaseous Detection Devices capable of indirectly detecting secondary electrons at high pressure. The secondary electrons cause ionization in the gas and the gas ion are then collected and counted to produce the image signal.

As the gas pressure is raised in the sample chamber, more of the electrons in the beam will be scattered. Since most of the scattering occurs through a fairly large angle, there is relatively little image degradation as the "halo" produced by the electron scattering primarily increases noise and reduces contrast. However, the scattering does prevent the highest resolution so ESEM is generally used at fairly low magnification. Furthermore, the gas does limit the working distance and the energy at which the beam can be used. Working distance should be no more than 5 to 10 mm which minimized the distance the electrons pas through the gas. ESEM is generally not possible below 5 kV and 15 kV is a more typical beam voltage. Most insulating materials are best imaged at low voltage from 1 to 5 kV (see Fig. 12.) When choosing to use variable pressure to reducing sample charging, there is a trade-off between the charge reduction of the ionized gas vs. the loss of low voltage capability. Depending on the material, low voltage may be more useful than variable pressure for some samples. ESEM users would be well advised to test their own samples in a variable pressure SEM before purchasing.

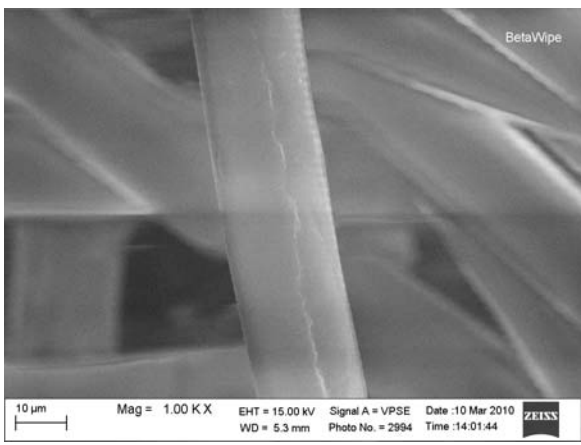
Energy dispersive x-ray (EDS) analysis is problematic in the ESEM because electron beam scattering into the beam skirt greatly degrades the lateral spatial resolution of the technique. Specimen current methods like EBIC (electron beam induced conductivity), RCI (resistive contrast imaging), and CIVA (charge induced voltage alteration) are also difficult at

variable pressure because the gas ionized by the electron beam alters the absorbed sample current.



20 µm **Mag = 388x**

Figure 38: Clean room textile at 15 kV without variable pressure.



10 µm **Mag = 1,000x**

Figure 38: Clean room textile at 15 kV and 100 Pa variable pressure.

Acknowledgments

The author would like to thank Mr. Dan Hinkel of LPS for taking the ESEM images.

References

1. Joseph I. Goldstein, Dale E. Newbury, Patrick Echlin, David C. Joy, A. D. Romig, Jr., Charles E. Lyman, Charles Fiori, and Eric Lifshin, *Scanning Electron Microscopy and X-ray Microanalysis, A Textbook for Biologists, Materials Scientists, and Geologists*, 2nd Edition, Plenum, New York, 1992, p. 22.
2. P. Gnauck, Proceedings of ISTFA 2003, pp. 132 (2003).
3. T.W. Lee, "A review of wet etch formulas for silicon semiconductor failure analysis", *Microelectronics Failure Analysis Desk Reference,* 4th Edition, ASM International, Materials Park, OH, p. 589-601 (1999).
4. Goldstein, p. 219.
5. W. E. Vanderlinde, Proceedings of ISTFA 2002 pp. 77-85 (2002).
6. E. Coyne, Proceedings of ISTFA 2002 pp. 93-99 (2002).
7. B. Tracy, Proceedings of ISTFA 2002, pp. 69-76 (2002).
8. Private communication, Peter Gnauck, LEO Electron Microscopy GmbH.
9. Goldstein, et al., p. 89.
10. Available from Small World LLC, 2226 Chestertown Drive, Vienna, VA 22182.
11. Goldstein, et al., p. 269.
12. G.D. Danilotos, SCANNING, Vol. 4, p. 9 (1981).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Ultra-high Resolution in the Scanning Electron Microscope

W. Vanderlinde
Laboratory for Physical Sciences
College Park, Maryland, USA

Introduction

Semiconductor manufacturers are now shipping product with sub-100 nm features. Measuring features and defects in this size range has become a challenge for the scanning electron microscope. For metrology below 200 nm, some manufacturers have begun routinely using TEM (transmission electron microscopy) which is tedious and expensive. [1] This is quite disappointing since the SEM should, in principle, be capable of providing more than enough resolution for these features.

In this article two innovative methods are described for significantly improving resolution in SEM imaging: STEM-in-SEM (scanning transmission electron microscopy in a scanning electron microscope) [2,3,4], and FSEI (forward scattered electron imaging.) [5] Both methods can be implemented in any SEM by using special sample holders and do not require any modification to the SEM.

Resolution limits in the SEM

Modern field emission SEMs have a nominal spot size as small as 1 nm. However, it is very rare that 1 nm resolution is realized in everyday imaging. Typically a SEM service engineer will demonstrate the microscope resolution by using very high beam voltage such as 30 kV to image a special "resolution standard" which consists of gold islands on a carbon substrate, commonly referred to as a "gold-on-carbon" sample. See Figure 1 which is at extremely high 500,000X magnification. Imaging a more typical sample such as an electronic device will generally produce an image with much lower resolution that may not meet the needs for modern failure analysis and metrology.

It is sometimes possible to achieve very high resolution by sputter coating a very thin layer of metals such as Au, Au/Pd, Cr, or Pt. Typically, such coatings result in a better image of the sample surface, not only because they eliminate charging but because they also increase the surface emissivity of the sample and reduce the secondary electron range. About 10 nm of metal is required to produce a continuous sputtered film, but this amount of metal may obscure surface detail and cause image artifacts from the grain structure of the sputtered metal. It is therefore unacceptable to sputter coat for imaging at 1 nm resolution (see Figure 2.)

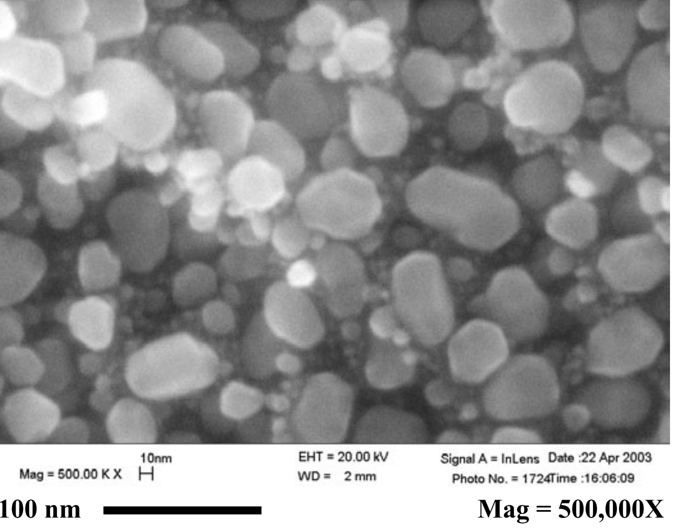


Figure 1: Secondary electron image of gold islands on a carbon substrate shows that very high resolution SEM images are possible under the right conditions.

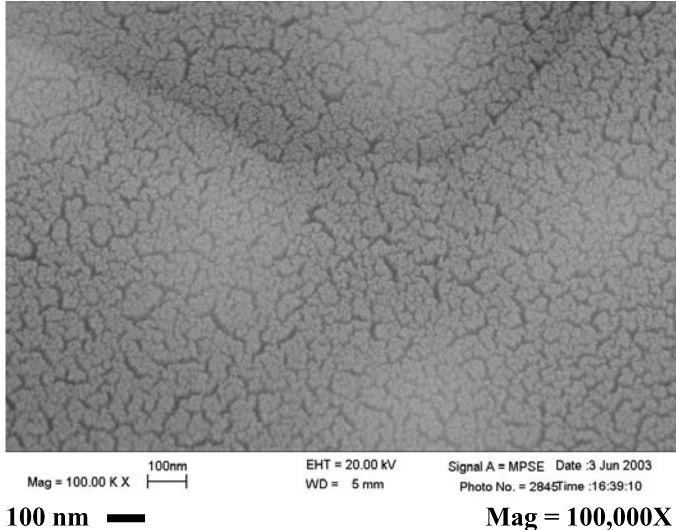


Figure 2: Secondary electron image of aluminum metal sputter coated with 10 nm of Au/Pt. The image has high resolution but the original surface detail is obscured and the sputter coating grains introduce artifacts seen as jagged dark lines.

In order to achieve ultra-high resolution in the SEM, three conditions must be satisfied: [6]

1. The electron beam must be finely focused to a small spot at the surface of the sample.

2. The electron beam current must be high enough to produce sufficient signal so that visible contrast is achieved in the image.

3. The signal used for imaging must originate very close to the impact area of the beam on the sample.

Conditions 1 and 2 are satisfied when the electron source is operating at high brightness, which is maximized when the electron source is at high beam voltage. However, high energy electrons have a large range in most materials. When the electrons are backscattered from deep in the sample towards the surface, they may exit the sample at a significant distance from the probe location (see Table 1.) Thus condition 3 is not satisfied at high voltage since the backscattered electrons are not closely correlated with the probe. These electrons also produce secondary electrons as they exit and this further reduces the spatial correlation between the secondary electrons and the probe. If the beam voltage is reduced so as to decrease the range of the electrons, then the brightness of the electron source is reduced and it is not possible to achieve a small spot with sufficient beam current (conditions 1 and 2). For the special gold-on-carbon sample all three conditions are satisfied at high beam voltage because gold has a very high backscatter coefficient and carbon has a very low backscatter coefficient. Thus incident electrons are usually either reflected from the gold surface or lost in the carbon substrate. However, it is very difficult to find ordinary samples that will allow the three conditions to be met.

Table 1: Electron beam spot size (per the manufacturer [7]) and penetration range (from the Kanaya-Okayama formula [8]) as a function of electron beam energy.

Beam energy (keV)	Spot size (nm)	Electron penetration range in aluminum (microns)
1	2.4	0.028
3.5	1.5	0.22
5	1.3	0.41
10	1.1	1.32
20	1.0	4.19
30	1.0	8.24

STEM-in-SEM

One solution to satisfying the three conditions for high resolution imaging is to create a very thin sample ($\leq$ 100 nm thick) and image using the transmitted high energy electrons. Electrons will scatter very little as they pass through a thin film. Thus the detected signal is highly localized, even at high beam voltage where the electron optics produce the best spot size. This approach enables high resolution SEM images that look very similar to TEM images. FIB techniques enable the creation of "TEM samples" less than 100 nm thick much easier and faster than in years past.

A typical "STEM-in-SEM" holder is shown in schematic in Figure 3 and in an optical photo in Figure 4. [2] The thin sample is mounted on top of a graphite block. After passing through the thin film, the electron beam travels down a narrow hole drilled in the graphite block. By absorbing the most highly scattered electrons, the graphite block collimates the beam which improves the image resolution. After passing through the collimator, the electrons strike a pair of gold reflectors that create secondary electrons which are collected by the ordinary in-chamber secondary electron detector. Alternately, one can invest in a dedicated STEM-in-SEM detector which can be added to some SEM models. [4]

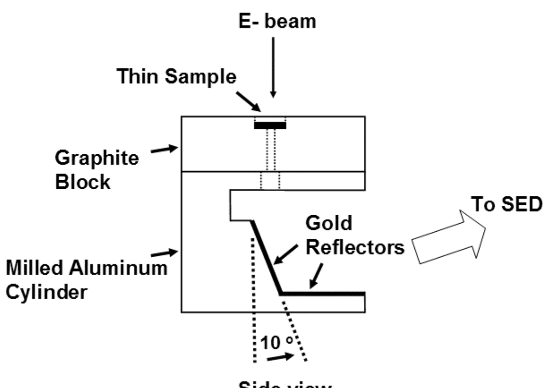


Figure 3: Design for a "STEM-in-SEM" sample holder. The overall height is 2.5 cm.

STEM-in-SEM samples are typically mounted onto a standard 3 mm diameter TEM disk, which is then placed into a cylindrical hole in the top of the STEM-in-SEM holder and held in place by a retaining ring. However, the STEM-in-SEM holder can be easily modified for a variety of sample shapes and lengths. When thin samples are prepared by FIB, the 3 mm limitation on sample size imposed by standard TEMs can significantly increase the difficulty of the preparation. With STEM-in-SEM, the sample can be of virtually any size. It is, however, crucial that the sample be mounted in a clean and mechanically rigid fashion. Metallic tape is often used to mount SEM samples, but the tape will allows unacceptable sample drift. Carbon paint is also used to mount SEM samples, but with STEM-in-SEM the paint may transfer carbon contamination to the thin section. Therefore, a

mechanical mounting method such as spring clips or retaining rings is recommended.

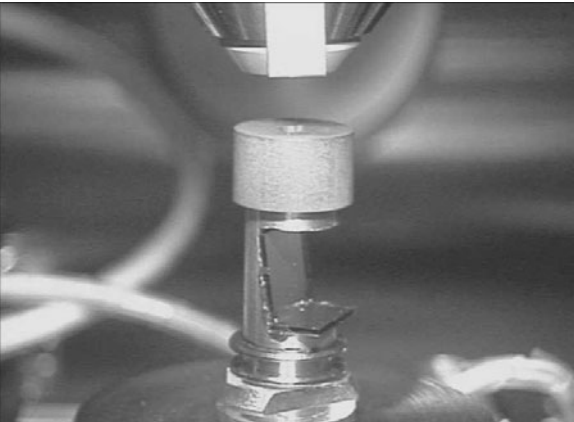


Figure 4: Optical photo of the STEM-in-SEM sample holder mounted to the SEM stage. The pole piece of the objective lens is visible above the sample holder.

Transmission electron microscopes typically operate at 200 kV or higher, so it may be somewhat surprising that an SEM's electron beam voltage of 30 kV or less can be used to view a sample in transmission. While it is often said that TEM samples must be thinned until they are "electron transparent," this is a misleading statement. The electron range of a 30 keV electron in aluminum is more than 8 microns. TEM samples are typically thinned to less than 100 nm thick so that most of the electrons undergo no more than one scattering event. Accordingly, TEM imaging is typically done with electrons that have not undergone any scattering events (bright field) or those that have been scattered exactly once by the lattice (dark field). At 200 keV the mean free path for elastic scattering is large (~350 nm) compared to the thickness of a 100 nm thin film. For STEM-in-SEM at 30 keV the elastic mean free path in aluminum is 17 nm for a scattering angle of 2 degrees or greater, and an electron will typically be scattered several times when passing through the sample. This results in substantially degraded lateral spatial resolution.

To estimate the resolution degradation, a Monte Carlo simulation of electron trajectories was conducted using the commercially available "Electron Flight Simulator" program [9] as shown in Figure 5. The Monte Carlo simulation shows that 30 keV electrons will be scattered into a cone of half-angle 8.5 degrees, resulting in a beam diameter of about 30 nm as the beam exits a 100 nm thin film. In order for a raster beam system to produce a sharp image, electrons should collect information only from the area very close to where the beam enters the sample. If the beam scatters 30 nm laterally within the sample, then contrast will be collected from materials that are far from the electron beam position, and images will be seriously degraded. Thus in one design [2] a graphite collimator was placed in the STEM-in-SEM holder below the sample to absorb the most highly scattered electrons. The design of the collimator necessitated a trade-off of large transmitted signal and ease of alignment with the need to eliminate the most highly scattered electrons that will degrade image resolution.

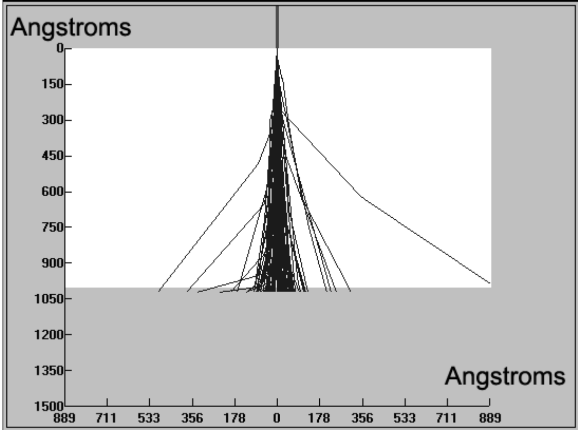


Figure 5: Monte-Carlo simulation of the trajectories of 30 keV electrons passing through a 100 nm aluminum thin film.

Procedure

Figure 6 is a low magnification STEM-in-SEM image. The image shows a conventional TEM sample, consisting of a thinned strip of silicon attached to a 3 mm copper grid with an oval hole. The bright circle in the center of the picture corresponds to the hole in the graphite collimator which allows electrons to penetrate to the gold reflectors. If necessary, the holder can be tilted to bring the area of interest to the center of the bright circle, and the entire holder is then rotated to maximize the collection of secondary electrons from the gold reflectors to the secondary electron detector. Once the holder is aligned, high resolution STEM-in-SEM images can be acquired by selecting a thin area of the sample and simply increasing the magnification, and then focusing and stigmating in the usual fashion.

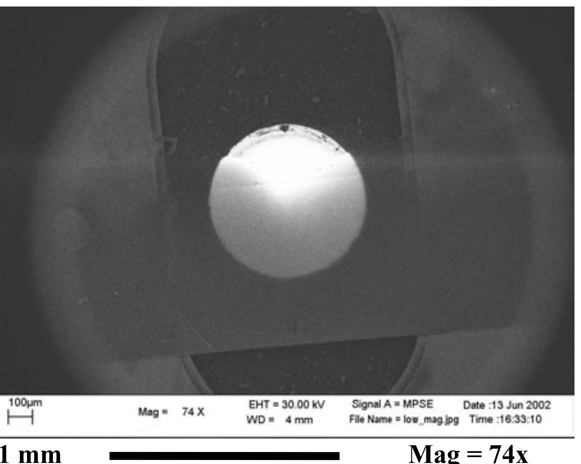
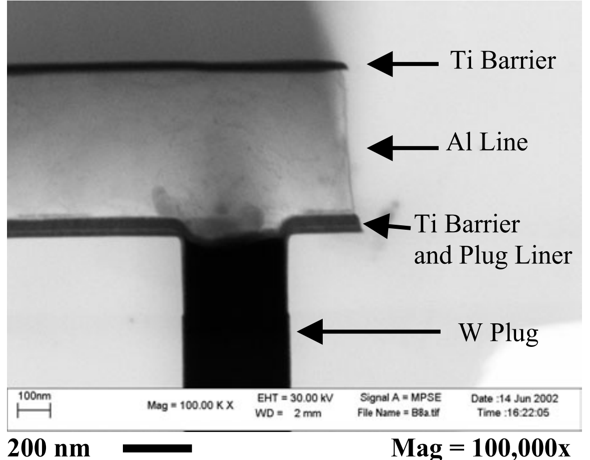


Figure 6: Low magnification STEM-in-SEM image showing the TEM sample.

Semiconductor cross-sections

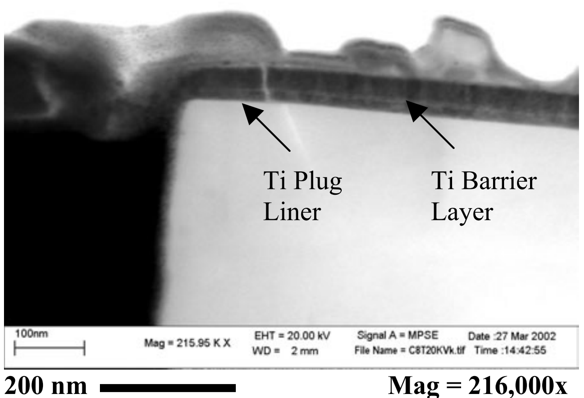
Figures 7, 8, 9, and 10 are STEM-in-SEM images of typical CMOS structures such as aluminum lines, tungsten plugs, and

poly lines with sidewall spacers. These images were taken at 20 kV or 30 kV beam voltage and 2 mm working distance. The STEM-in-SEM images have a similar appearance and contrast to TEM images. They show internal grain structure on the poly-silicon, the aluminum, and the barrier metals. Close examination of the STEM-in-SEM images demonstrates a resolution of approximately 2 nm which is far superior to that achieved by traditional SEM images of these structures. The STEM-in-SEM image contrast is based on differences in atomic number, density, and crystallographic orientation, all of which influence how electrons scatter away from the detector. [10]



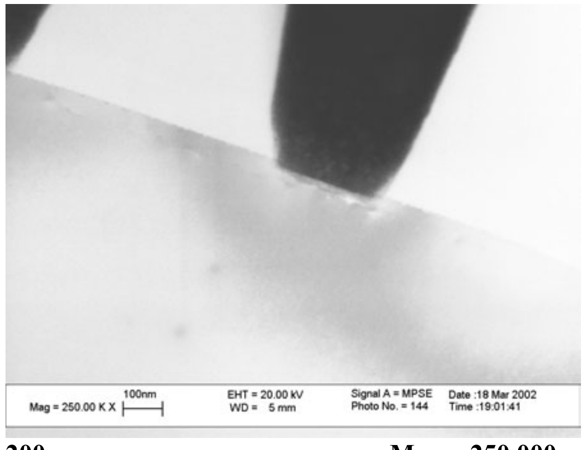
200 nm **Mag = 100,000x**

Figure 7: STEM-in-SEM image of an aluminum line connected to a tungsten plug contact. The barrier layer and the plug liner are visible.



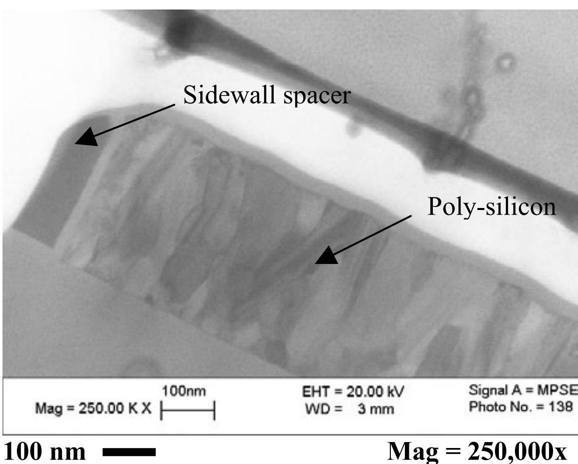
200 nm **Mag = 216,000x**

Figure 8: Higher magnification STEM-in-SEM image resolving a titanium plug liner and barrier layer.



200 nm **Mag = 250,000x**

Figure 9: STEM-in-SEM image of a tungsten plug contacting silicon.

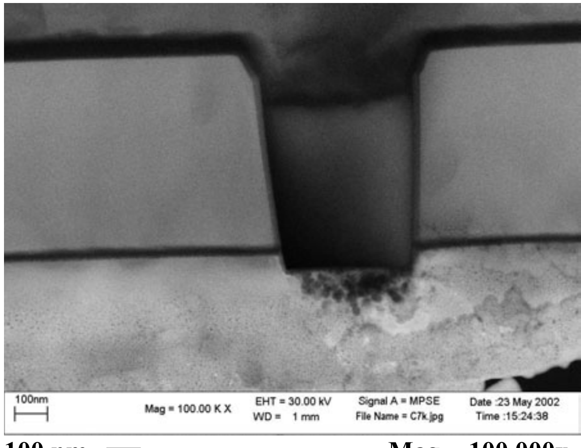


100 nm **Mag = 250,000x**

Figure 10: STEM-in-SEM of broad poly-silicon line with sidewall spacer.

The STEM-in-SEM holder can be tilted so that the imaged area is outside the bright circle that corresponds to the entrance of the collimator, see Figure 11. Under these conditions, only the scattered electrons are able to pass through the collimator and reach the gold reflectors, in effect creating a “dark field” image. Note that the tungsten which is dark in Figures 7 and 8 now shows some bright areas. The overall image has degraded resolution, but the contrast on the material just below the tungsten-silicon interface is enhanced. This type of dark field imaging could be used to enhance the contrast from highly scattering materials such as inter-metallic grains at material interfaces.

STEM-in-SEM also allows for very high spatial resolution x-ray maps. See the Chapter on Energy Dispersive X-ray Spectroscopy for details.



100 nm **Mag = 100,000x**

Figure 11: "Dark Field" STEM-in-SEM image.

Forward Scattered Electron Imaging

A second method for satisfying the three conditions for ultra-high resolution SEM is to image using low-loss electrons. When a high energy electron beam enters a sample at very high tilt angle, a significant fraction of the incident electrons will be forward scattered from the sample surface with relatively little energy loss. The electrons with the smallest energy loss will have experienced the smallest path length through the material and thus will be highly localized to the beam impact area. The low-loss signal is typically detected by a large acceptance angle (~45 degrees) energy-filtered detector placed below the sample, as developed by Wells at IBM. [11,12,13] However, this requires the expense and complication of adding a special detector to the microscope chamber. Interestingly, some very high spatial resolution "low-loss" images were obtained by Broers using a very different set-up, [14] including a very small acceptance angle (2.6 degrees semi-angle) detector in the forward scattered position. Although it was equipped with an energy filter, Broers left the filter turned off while obtaining very high resolution images! This result showed that a narrow collection angle is a more efficient method of selecting the high resolution low-loss electrons than energy filtering. This inspired the method now known as forward scattered electron imaging (FSEI) which can be performed in an unmodified SEM using a special sample holder. [5]

Sample holder design

The FSEI holder (see Figs. 12, 13 and 14) places the sample at high tilt angle to the beam. The forward scattered high-energy electrons will strike a gold reflector and create secondary electrons that are collected by the SEM's regular secondary electron detector. The holder is designed with the sample surface pointed away from the secondary detector so that only the forward scattered electrons contribute to the image. [5]

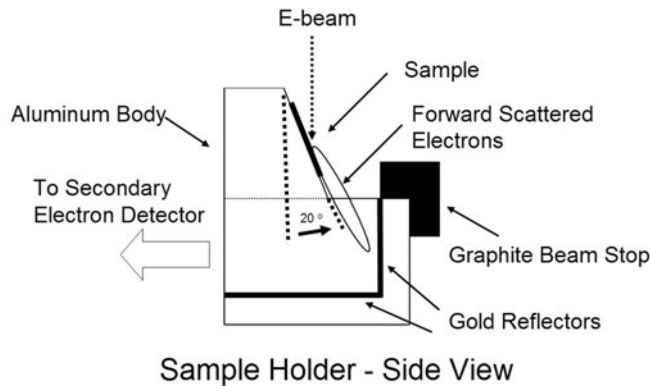


Figure 12: Design for sample holder with integrated detector for forward scattered electrons.

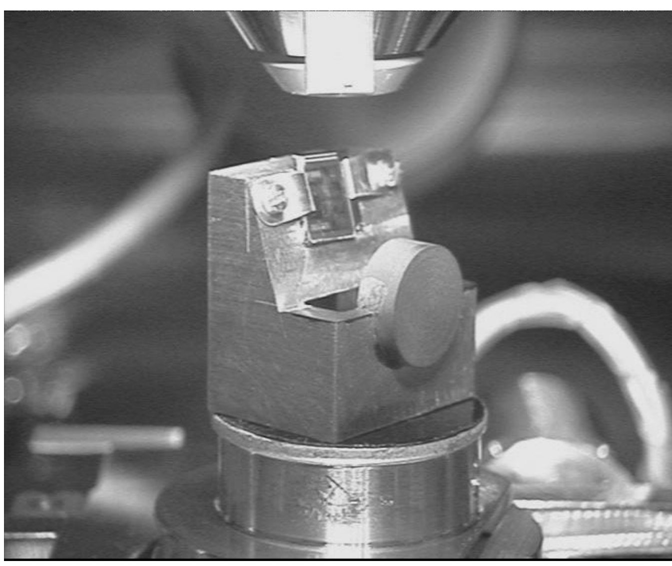


Figure 13: An optical image of the front of the FSEI holder.

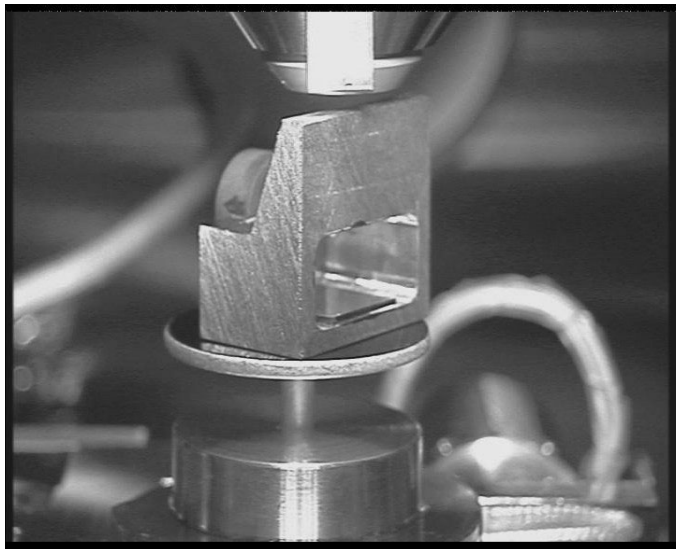


Figure 14: An optical image of the back of the FSEI holder.

This technique does not provide any energy filtering of the forward scattered electrons, so it will be referred to as "forward scattered electron imaging" rather than "low loss electron imaging." However, it is found that the best resolution is achieved when imaging only with the forward scattered electrons that travel the minimum distance through the sample. Since these electrons will have the least lateral scattering, a relatively small gold target is chosen and the remaining surfaces of the aluminum holder body are painted with electron absorbing carbon paint. A graphite "beam stop" is also added behind the holder. In this design, the gold target is 4 mm wide which only allowed electrons with a horizontal or vertical deflection of less than 10 degrees to hit the gold target.

The major drawback to this technique is that the sample normal is at very high tilt angle (70°) to the electron beam. Thus, the image is highly foreshortened along the vertical axis, by about a factor of three ($1/\cos 70^o = 2.9$), and the lateral spatial resolution of the image will be accordingly better than the vertical spatial resolution. However, this is not a major disadvantage for certain types of metrology. If structures such as sub-0.1 micron poly lines are oriented orthogonal to the tilt axis, the line width and edge profile can be measured with very high precision. Since the forward-scattered high energy beam is relatively unaffected by sample charging, it is also possible to image insulating materials like photo-resist or spin-on-glass at high beam voltage.

Procedure

The sample is mounted to a highly tilted surface near the top of the holder and held in place by metal clips. The sample surface is pointed away from the chamber secondary electron detector so that only the forward scattered electrons contribute to the image. Since most SEMs have their best resolution at very short working distance, it is suggested that the feature of interest be mounted near the top of the holder. The maximum sample size for this holder design is about 1 cm by 1 cm, although larger holders can be built. To use the sample holder in FSEI mode the stage tilt must be zero and the stage must be rotated so that the back of the holder points towards the chamber secondary electron detector. The sample cannot be re-oriented except by removing it from the vacuum chamber and remounting it.

Polysilicon gate metrology

0.7 micron polysilicon gate samples were prepared from a finished loose die by plasma etching to remove the passivation, followed by an HF dip to remove the metal and dielectric layers. A normal SEM image was first taken with the die face pointed towards the secondary electron detector (SED). The holder was then rotated 180° so that the die surface was pointed away from the SED, and the only signal reaching the detector was from secondary electrons produced by the forward scattered electrons striking the gold reflectors.

Images taken using high beam voltage on uncoated samples typically have very poor contrast and little surface detail as seen in Figure 15. However, when exactly the same beam conditions (30 kV, 4 mm working distance, 30 micron aperture, 70° tilt) were used in FSEI mode, the resulting image was far superior, see Figure 16. The image contrast and surface detail are both greatly improved. Note that this is an *uncoated* sample imaged at 30 kV. In many cases coating a sample is undesirable since the coating may obscure fine detail or interfere with microanalysis by energy dispersive x-ray or scanning Auger spectroscopy. Therefore, FSEI is particularly useful as a very high resolution technique that does not require sputter coating the sample.

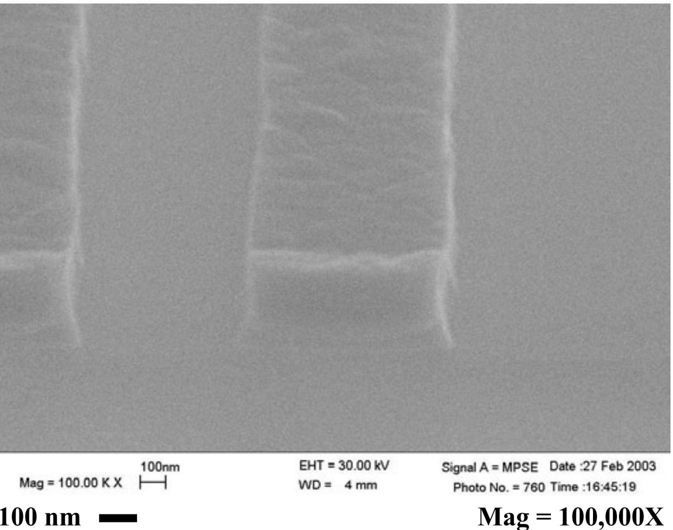


Figure 15: 700 nm poly lines on silicon, uncoated, 30 kV beam energy, 70 degrees tilt, secondary electron imaging.

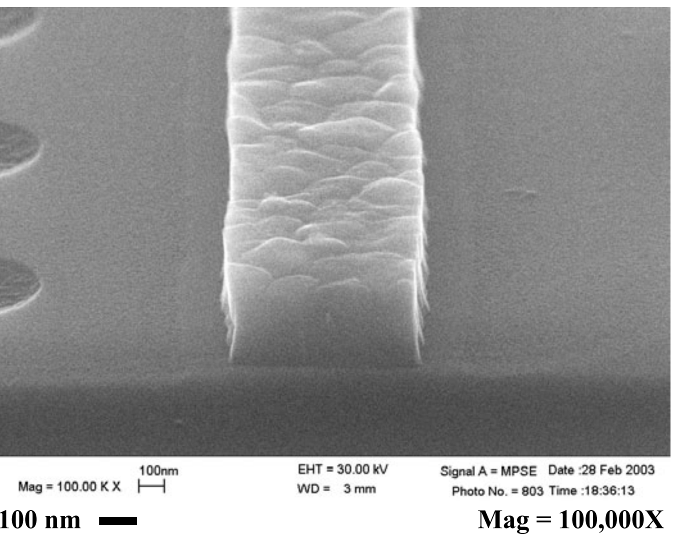
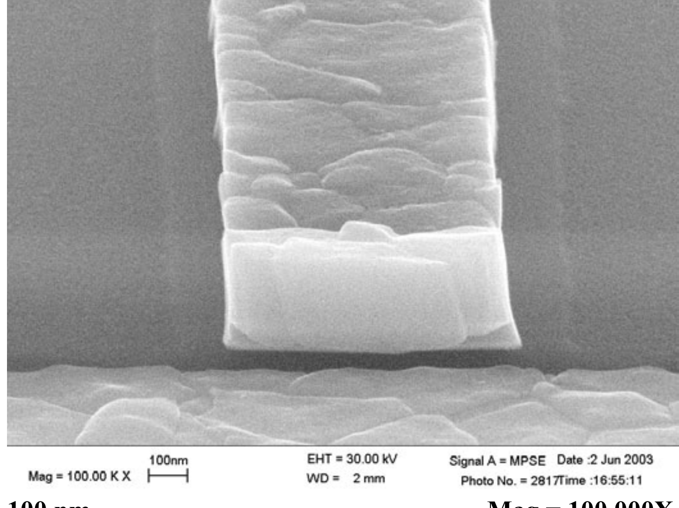


Figure 16: 700 nm poly lines on silicon, uncoated, 30 kV beam energy, 70 degrees tilt, forward scattered electron imaging.

Dynamic focus

One disadvantage of using high tilt angle is the difficulty in keeping the entire sample in focus due to limitations in the depth of field, particularly at the short working distances required for optimum resolution. Thus in Figure 16 the top part of the image lacks the sharp focus seen in the bottom of

the image. However, many electron microscopes have a feature called dynamic focus which allows the electron beam focus to be automatically adjusted as it rasters across a surface. This works quite well if the surface is flat as in Figure 17. Most of the FSEI images in this article were acquired using dynamic focus. The resolution of this image was measured to be about 1 to 2 nm.



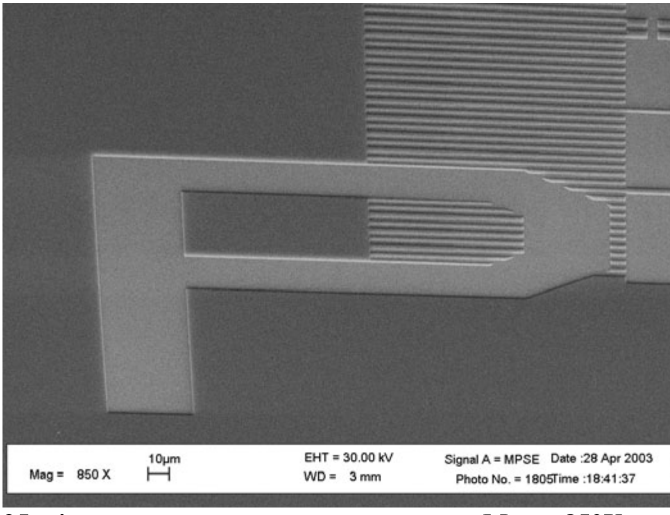
100 nm **Mag = 100,000X**

Figure 17: 700 nm poly lines on silicon, uncoated, 30 kV beam energy, 70 degrees tilt, forward scattered electron imaging. Dynamic Focus was used to keep the entire sample in focus despite the high tilt angle.

Photoresist metrology

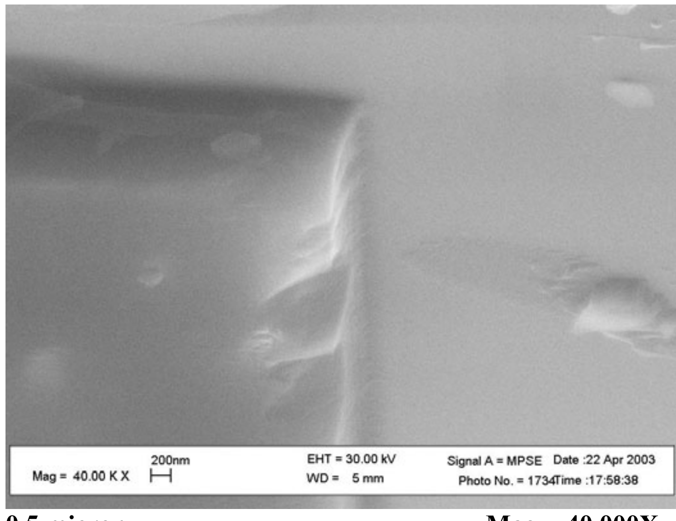
Photo-resist is a highly insulating material that is typically imaged in the SEM either with a metal sputter coating or at low beam voltage (~ 1 kV). However, since the forward scattered low loss electrons with close to 30 keV energy are much less sensitive to sample charging than are the low energy secondary electrons with less than 50 eV energy, it is possible to image uncoated resist with a 30 kV beam using the forward scattering technique.

A photo-resist layer 1 micron thick on silicon dioxide was patterned and developed with a test structure, and then imaged with a 30 kV electron beam using the forward scattered technique, see Figs. 18 and 19. No charging or sample damage was observed. Since the resist has a lower density than silicon or metals the incident beam should tend to penetrate deeper into the sample before being scattered towards the detector. Thus it was expected that the spatial resolution and surface sensitivity would be reduced when imaging resist compared to imaging other semiconductor materials. However, Figure 19 at 40,000X is surprisingly sharp and surface sensitive. This physical mechanism for this result is not yet fully understood.



25 microns **Mag = 850X**

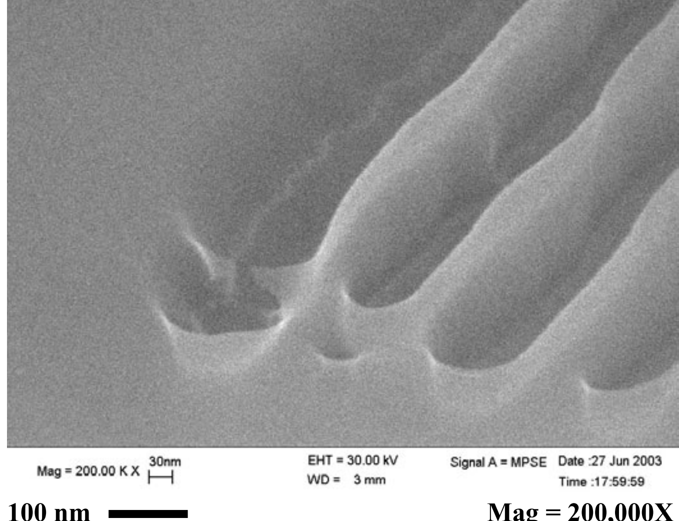
Figure 18: Photo-resist on silicon dioxide, uncoated, 30 kV beam energy, 70 degrees tilt, forward scattered electron imaging.



0.5 micron **Mag = 40,000X**

Figure 19: Photo-resist on silicon dioxide, uncoated, 30 kV beam energy, 70 degrees tilt, forward scattered electron imaging.

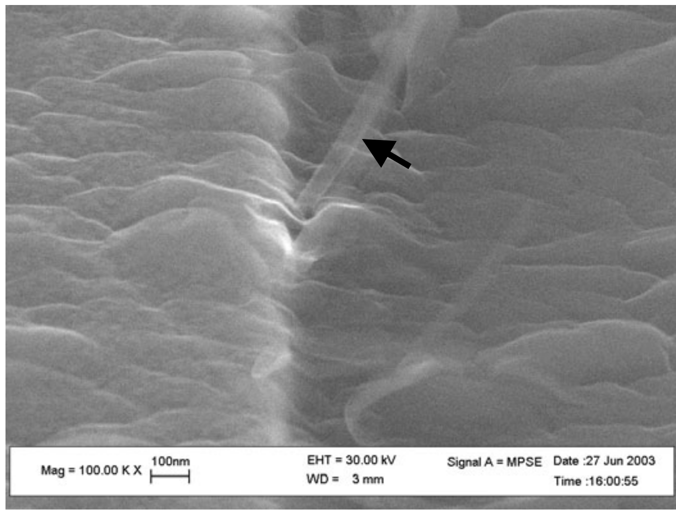
A PMMA (poly methyl methacrylate acid) e-beam resist layer 250 nm thick was patterned by electron beam lithography with 50 nm lines, see Figure 20. Although the contrast in this 200,000X image is marginal the 50 nm lines are clearly visible, which is a remarkable result since very fine resist structures such as this are almost always sputter coated before imaging. FSEI imaging of uncoated resist lines below 100 nm represents a new metrology capability.



100 nm **Mag = 200,000X**

Figure 20: PMMA e-beam resist on silicon dioxide, uncoated, 30 kV beam energy, 70 degrees tilt, forward scattered electron imaging.

Figure 21 is from an actual failure analysis problem, an investigation of the grain structure of the organic semiconductor pentacene. In Figure 21, a 50 nm thick pentacene film covers a gold electrode on the left side of the image, and thermal oxide on the right side of the image. The grain size is reduced on the gold electrode which affects device performance. It was also noted that 50 nm diameter extrusions of pentacene occur at the gold / oxide interface.



200 nm **Mag = 100,000X**

Figure 21: Pentacene thin film on gold (left) and oxide (right), uncoated, 30 kV beam energy, 70 degrees tilt, forward scattered electron imaging.

Image correction

The use of high tilt angle foreshortens the image resulting in image distortion. However, it is easy to correct this distortion by resizing the vertical scale of the image with any standard graphics program. Some scanning electron microscopes have this feature labeled as "tilt correction". Note that image correction works well for very flat samples, but any features that protrude out of the flat plane of the sample will be severely distorted by this procedure.

Beam damage

High energy electron beams will sometimes damage sensitive materials such as polymers. However, in taking several dozen very high magnification images on polymer films, damage from melting, vaporization, or electrostatic discharge was never observed, despite using slow scan at 200,000X and focusing and stigmating at 800,000X with 1/3 partial field. The lack of damage is somewhat surprising, but may be explained by the high tilt angle which distributes the deposited energy across a large surface area. Also, the energy absorption per path length falls as the beam energy increases so the actual energy deposition at the point of impact is less for higher beam energy.

Discussion

What physical effects produce the contrast in the FSEI images? In FSEI, electrons are scattered towards the target when they undergo collisions with atoms in the sample. In this aspect the technique is more similar to TEM (transmission electron microscopy) than to SEM. One might almost be tempted to refer to forward scattered imaging as "glancing angle TEM" especially since the magnification achieved by this technique is most often associated with TEM imaging.

Despite the far simpler hardware, FSEI images are in many cases superior to the low loss electron images published in the literature. This confirms that collection of a small solid angle of the forward scattered beam is more effective at selecting the high-resolution and surface-sensitive electrons than selecting out the low-loss electrons with an electron energy filter.

For FSEI, electrons will scatter towards the target when they undergo either elastic or inelastic collisions with atoms in the sample. In the case of low-loss imaging, only the inelastically scattered electrons are collected by the detector. Since the combined scattering cross-section for elastic and inelastic collisions is higher than for inelastic collisions alone, FSEI is highly effective compared to low-loss imaging.

Conclusion

Relatively simple techniques can enable very high resolution imaging in the SEM. Self-contained sample-holder/detectors for STEM-in-SEM and FSEI enable these technique to be performed in an ordinary unmodified SEM. Approximately 1 to 2 nm resolution images can be obtained on thin sections, uncoated polysilicon gates, and photoresist. These are promising techniques for sub-100 nm metrology.

References

1. L. Tsung, A. Anciso, R. Turner, T. Dixon, and N. Holloway, Proceedings of ISTFA 2001, p. 299-302 (2001).
2. W. E. Vanderlinde, Proceedings of ISTFA 2002 pp. 77-85 (2002).
3. E. Coyne, Proceedings of ISTFA 2002 pp. 93-99 (2002).
4. B. Tracy, Proceedings of ISTFA 2002, pp. 69-76 (2002).
5. W. E. Vanderlinde, Proceedings of ISTFA 2003 pp. 158-165 (2003).
6. Joseph I. Goldstein, Dale E. Newbury, Patrick Echlin, David C. Joy, A. D. Romig, Jr., Charles E. Lyman, Charles Fiori, and Eric Lifshin, *Scanning Electron Microscopy and X-ray Microanalysis, A Textbook for Biologists, Materials Scientists, and Geologists*, 2nd Edition, Plenum, New York, 1992, p. 219.
7. Private communication, Peter Gnauck, LEO Electron Microscopy GmbH.
8. Goldstein, et al., p. 89.
9. Available from Small World LLC, 2226 Chestertown Drive, Vienna, VA 22182.
10. Goldstein, et al., p. 269.
11. O.C. Wells, Appl. Phys. Lett. 19 (7) p. 232-235 (1971).
12. O.C. Wells, Appl. Phys. Lett. 49 (13) p. 764-766 (1986).
13. O.C. Wells, F.K. LeGoues, and R.T. Hodgson, Appl. Phys. Lett. 56 (23) p. 2351-2353 (1990).
14. A.N. Broers, B.J. Panessa, and J.F. Gennaro, *Proceedings of the Eighth Annual Scanning Electron Microscope Symposium (SEM/1975),* IIT Research Institute, Chicago, p. 233-242 (1975).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Transmission Electron Microscopy for Failure Analysis of Semiconductor Devices

Swaminathan Subramanian and Raghaw S. Rai ‡

Freescale Semiconductor Inc.
Austin, TX

1. Introduction

The ultimate goal of the failure analysis process is to find physical evidence that can identify the root cause of the failure. The aggressive scaling of semiconductor device features in evolving technologies has made identification and physical characterization of defects extremely challenging. Transmission electron microscopy (TEM), because of its superior spatial resolution and elemental analyses capabilities, has emerged as a powerful tool to characterize such subtle defects.

In a transmission electron microscope, a high energy (100 to 300 keV) electron beam is transmitted through the thinned area of interest of the sample. During the transmission process, a variety of beam-specimen interactions occur. These interactions yield transmitted electrons, elastically and inelastically scattered electrons, secondary electrons, back-scattered electrons, Auger electrons and X-ray photons. Most of the transmitted and elastically scattered electrons and some of the inelastically scattered electrons are used to form an image. TEM based elemental analysis techniques utilize X-ray photons in energy dispersive spectroscopy (EDS) [1-2] and inelastically scattered electrons or the 'energy-loss' electron in electron energy loss spectroscopy (EELS) and Energy-filtered transmission electron microscopy (EFTEM) [2-5]. An illustration of the physical location of the sample, imaging plane, EDS and EELS in a transmission electron microscope is shown in Figure 1.

A TEM sample containing the defect site has to be prepared from failing die for successful TEM analysis. The sample preparation involves thinning the area of interest containing the defect to achieve electron transparency, which is an irreversible destructive process. In most cases, final thickness at the area of interest exceeds the optimum thickness required for typical analysis. In such samples, a variety of TEM operation modes and techniques, including the conventional parallel beam illumination TEM, scanning transmission electron microscopy (STEM), energy-filtered TEM (EFTEM), and EELS and EDS based elemental mapping have to be employed for complete characterization of the defect and identification of root cause. Various applications of these techniques, as it pertains to semiconductor device failure analysis, are discussed in this review.

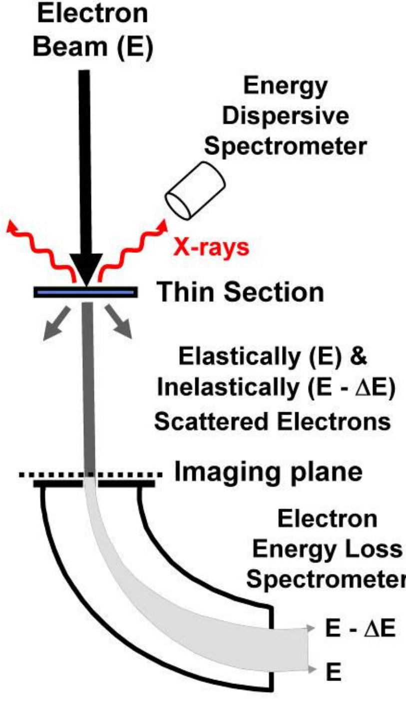


Figure 1: An illustration of a transmission electron microscope showing physical location of sample, imaging plane, energy dispersive spectrometer and electron energy loss spectrometer (E is in incident electron energy and ΔE is the energy loss because of inelastic scattering).

‡ Current Address: Nano Materials Characterization Group,
8402 Cornerwood Drive, Austin, TX 78717

2. TEM Sample

A thin section of the device, which is both transparent under electron beam illumination and encompasses the feature of interest, has to be prepared prior to TEM analysis. A variety of TEM sample preparation procedures [6-21] can be employed to construct the TEM sample. In FA, focused ion beam (FIB) instrument based procedures have become popular because of requirement to precisely thin the sample at specific failing site. In a FIB, a focused beam of gallium ion is used to ablate material around the failing site in a controlled manner. Advanced FIB instruments equipped with a scanning electron microscope (SEM) column permit simultaneous milling and imaging of defect location to successfully prepare a thin TEM section encompassing the feature of interest.

The thin TEM section is usually supported by a TEM grid. Typical TEM samples use 3 mm diameter circular or semi-circular TEM grid for support because most TEM sample holders are designed for such dimensions. In some cases, a rectangular bar with the length ≤3 mm, the width equaling to thickness of the die and the thickness less than 100 microns is used.

2.1 Thickness of TEM Section

Depending on analysis requirements, thickness of the TEM section can range between tens of nanometers to few hundreds of nanometers. Because the TEM sectioning is an irreversible destructive process, a section of optimum thickness should be prepared to successfully encompass the defect. In semiconductor applications, the final thickness can range between tens of nanometers to few hundreds of nanometers. The optimum TEM section thickness is decided based on the dimensions and composition of feature of interest, and the requirements of various TEM techniques. For example, if more than one device feature is included in TEM section, the resulting image may be a complex projection of overlapping features. It can become very difficult to extract useful information from such samples. To avoid this, the thickness of the section has to be less than dimension of the feature along the cross-sectioned direction. In advanced technologies (such as 22 nm or 45 nm nodes), the final section thickness may have to be less than tens of nanometers. This restriction is relaxed in the case of mature technologies, where the device features are larger.

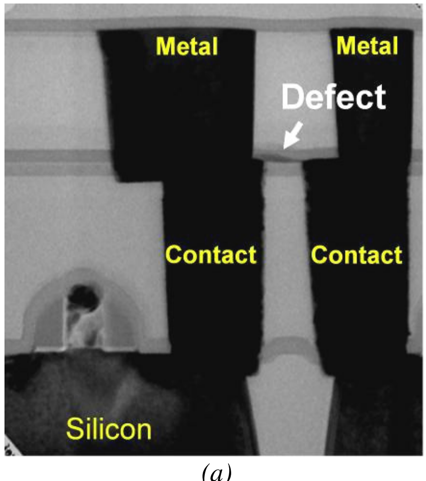
In silicon based devices, the image quality can seriously degrade in extremely thick (>250 nm) or extremely thin section. In extremely thick sections, the image can become blurry because of the noise from inelastically scattered electrons. The quality of images from such thick sections can be improved by employing EFTEM or scanning TEM (STEM) techniques. In extremely thin sections, the sample preparation induced damage results in poor quality images.

The preferred thickness of electron transparent section also depends on the requirements of various TEM techniques. To exploit the superior resolution (< 0.2 nm) capabilities of high resolution TEM (HRTEM) phase contrast imaging, the slice thickness should be maintained within several tens of nanometers. While EDS based elemental can work reasonably well in thicker sections, EELS and EFTEM based elemental analysis will yield best results only from thinner (tens of nanometers) sections. For dopant profiles analysis using electron holography thicker (<160 nm) [22] sections are preferred.

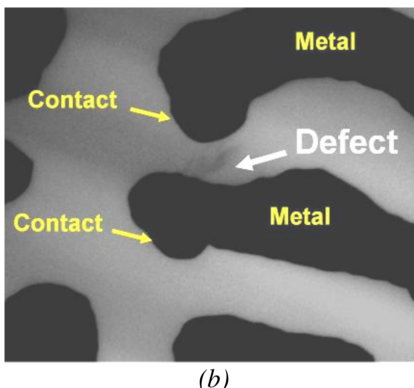
2.2 Cross-section *vs.* Planar TEM

In the cross-section TEM, the thin section is perpendicular to the surface of the die. This approach is useful when the suspected defect is at an interface or spans across multiple layers of the semiconductor device. Examples of such defects include interfacial layer between interconnects and metal lines, voiding in metal lines, particles and stringers connecting electrically isolated circuit elements, etc.

.



(a)



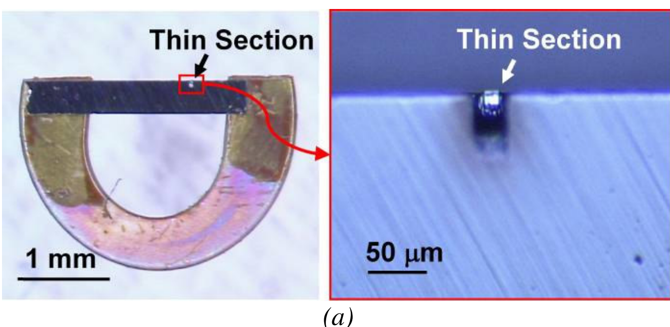
(b)

Figure 2: TEM images of similar defects causing leakage between adjacent contacts/metal lines: (a) cross-section and (b) planar.

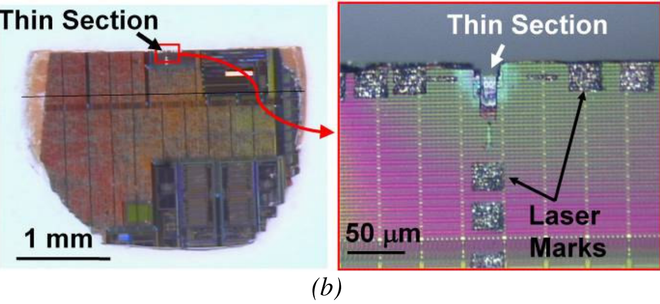
In planar TEM, the thin section is parallel to the surface of the die. This approach is preferred when defect is confined to a single layer or laterally dispersed. Examples of such defects include substrate level crystallographic defects (dislocation and stacking faults), stringers, particles etc. A cross-section and planar view of similar defect shorting adjacent contacts is shown in Figure 2. The choice between cross-section and planar TEM approach is made based on electrical symptoms of the failure.

2.3 Sample Preparation Procedures

The goal of the sample preparation procedures is to construct a sample that is both large enough to enable handling with a pair of fine tweezers and contains the thin section of the device encompassing the failing location. In semiconductor FA, the procedures for sample preparation can be classified in two generic approaches. The first approach [7-14] involves a combination of mechanical grinding, polishing or dicing of chip followed (in most cases) by FIB milling operations. Examples of images TEM samples prepared using this approach are shown in Figure 3.



(a)

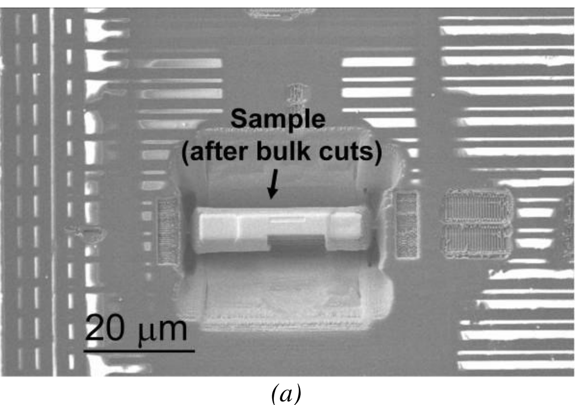


(b)

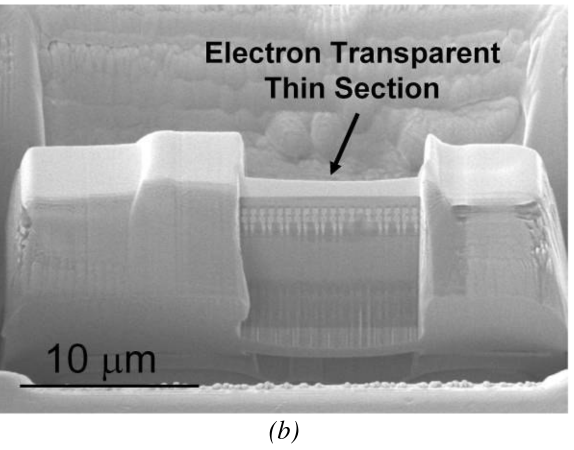
Figure 3: Optical images of samples prepared by mechanical grinding and subsequent FIB milling: (a) cross-section sample and (b) planar sample.

The second approach is commonly known as 'lift-out' [15-21], where the TEM sample is mostly constructed using a series of FIB milling (and welding) operations, with minimal damage to the substrate and transferred to a grid. The major advantage of the lift-out procedure is that it offers the potential for specific area TEM sample preparation of most components of integrated circuits with minimal geometric, dimensional or materials limitations. Also, because there is no mechanical grinding or dicing involved, TEM samples of exposed fragile defects can be prepared safely using lift-out method.

A cross-section sample prepared by the lift-out procedure is shown in Figure 4. In this version of the lift-out technique, a section of dimension 2 µm thickness x 30 µm width x 15 µm depth, encompassing the failure site is cut using the FIB. The sample is isolated from the die by cutting the bottom and milling the edges of the sample. The failure site is then thinned around the feature of interest to achieve electron transparency. The die along with sample is unloaded from the FIB chamber and the sample is lifted out of the die using a glass needle with 'static charge' at the tip acting as glue. The sample is then transferred from the glass needle to a membrane coated grid. The lift-out and transfer processes are performed under a high magnification optical microscope, equipped with a large working distance objective lens. Optical images of 3 mm diameter membrane coated grid, supporting the sample, are shown in Figures 5 and 6.



(a)



(b)

Figure 4: (a) A top down SEM image of a cross-section TEM sample after bulk and isolation cuts. (b) A cross-sectional scanning electron image of the lift-out TEM sample inside the FIB milled cavity. The surface of the die is tilted 52º with reference to electron beam.

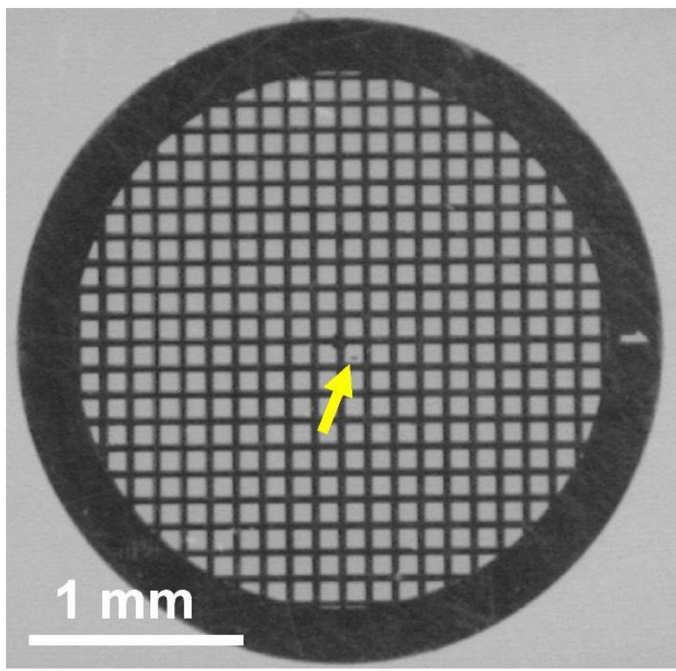


Figure 5: Low magnification optical image of the 3mm diameter membrane coated grid. The location of the sample is indicated by the arrow.

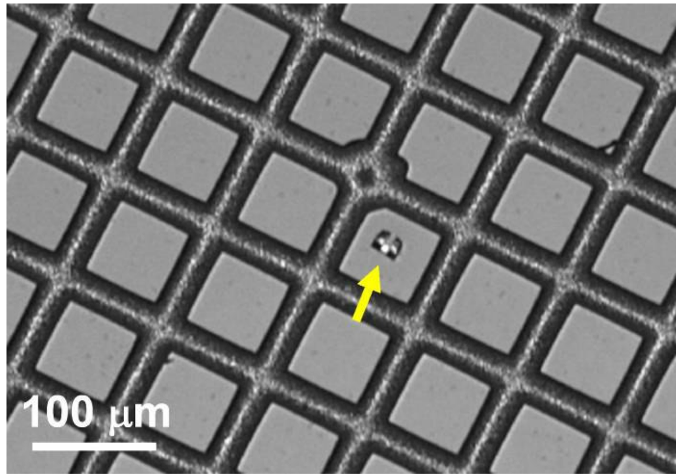


Figure 6: High magnification optical image of sample, placed on the membrane coated grid.

The success of the lift-out and transfer of sample using glass needle and static charge depends on the skills and experience of the operator. There is a reasonable chance of losing the sample during the lift-out process. In addition, this approach normally permits only one opportunity to thin the sample (even though the sample can be retrieved and rethinned or cleaned for further analysis in a TEM, such operations can be extremely labor intensive). These drawbacks of lift-out technique can be overcome by installing a probe or a manipulator in the FIB chamber [20-21]. The probe (or the manipulator) is used to transfer the TEM sample from the die to the grid by first welding the isolated sample, containing the area of interest, to a probe (or manipulator) with the aid of ion beam assisted metal deposition. The probe, along with the welded sample is retracted from the die. The sample is then transferred to a 3 mm diameter grid or similar supporting structure for ease of handling with fine tweezers. The transfer process involves positioning of the grid and sample attached to the probe in the viewing area of the FIB/SEM columns. The sample is then welded to the grid by ion beam assisted metal deposition and subsequently isolated from the probe by cutting the sample from the probe. The area of interest is then thinned for electron transparency using the FIB. This lift-out approach eliminates the uncertainties of the glass needle based lift-out step and is very useful for FA applications because the sample can be easily returned to the FIB chamber for further thinning after initial inspection in the transmission electron microscope.

The procedures for the planar TEM sample preparation lift-out technique [19, 21] will either require additional cross-sectional grinding operations prior to lift-out inside the FIB chamber or a combination of sample tilt, lift-out, weld and cut operations inside the FIB chamber to achieve the final geometry to allow milling of a thin section parallel to the surface of the die.

2.4 FIB Induced Damage

The FIB milling procedures used for TEM sample preparation often introduces undesirable damage to the thin section. The damage caused by the 30 keV focused gallium ions beam milling include implantation of gallium ions, amorphization of crystalline structures, mixing of components, and loss of fine structural detail [23-28]. When the surface of the die is imaged by an ion beam, it is milled and damaged. In order to prevent die-surface damage, a thin amorphous layer of platinum or tungsten is deposited over the region of interest. The amorphous metal layer also plays the critical role of reducing the thickness variations in thin section or the 'curtain effect', caused by uneven milling of device features that may arise due to differences in sputter rate of various materials present in the specimen. In addition, the metal layer forms a sacrificial layer and planarizes the surface topography that might cause uneven milling.

Exposure to 30 keV gallium ion beam can induce an amorphous layer (~20 nm [28]) on each surface of the thin section, even though the milled surface is parallel to the direction of the ion beam. When the section is extremely thin (only tens of nanometers for advanced technology nodes, HRTEM phase contrast imaging and EELS/EFTEM based elemental analysis), the total damaged layer thickness can easily exceed significant percent of material to be analyzed. Such damaged layer can be mostly removed by milling the final sliver with a 5 keV [28] and/or 2 keV ion beam, with the sliver surface oriented at an angle of five to ten degrees with respect to the direction of ion beam.

3. TEM Imaging

Various lenses, deflector coils and the imaging system settings can be selected (or stored and recalled) by the operator to achieve different TEM operation modes. The contrast variations observed in a TEM image will depend on beam-specimen interactions that occur under such conditions. This section describes principles behind commonly used imaging modes in semiconductor failure analysis and how these operation modes can be utilized to selectively maximize signal from specific beam-specimen interaction to yield useful information about the defect.

3.1 Parallel Beam illumination

In this mode, a parallel electron beam is transmitted through the sample and the image is formed, using several lenses, below the sample. The image can be recorded on a photographic film or a charge-coupled device (CCD) camera. The image formed under such condition is primarily dominated by three different types of contrast mechanisms: namely, diffraction, thickness-mass, and phase. The diffraction and thickness-mass contrast dominate during low magnification imaging of crystalline and amorphous semiconductor device features. The phase contrast supplemented by thickness-mass contrast is primarily used during high resolution imaging of sub-nanometer device features and defects.

Diffraction in Silicon Devices

Most semiconductor devices are fabricated on single crystal silicon wafer with the wafer normal parallel to [001] direction. In such a wafer (Figure 7), the [110] direction is parallel to the wafer flat or notch and lies on the (111) plane (cleavage plane) and perpendicular [001] direction (perpendicular to surface of wafer). It is important to note that the silicon crystal structure is diamond cubic and the x, y and z axes are interchangeble.

A cross-section image recorded with the electron beam oriented parallel to <110] crystallograpic directions of silicon substrate would ensure perpendicularity of cross-sectioned device features with respect to the electron beam and eliminate overlapping of various features of devices in the projected image. For this reason, most diffraction-contrast and HRTEM lattice images of cross-sectioned Si devices are recorded with electron beam direction parallel to the <110] crystallographic axis of the substrate.

It should be noted that, advanced semiconducter fabrication processes may utilize silicon wafers with different crystallographic orientations, including wafer normal along [001] and notch orientation at [100] or <111> and <110> wafer normals. In such cases, the cross-sectioned substrate should be oriented along proper crysallographic directions to achieve perpendicularity of device features with respect to the electron beam. In this article, the crystallographic directions shown in Figure 7 will be used as the reference.

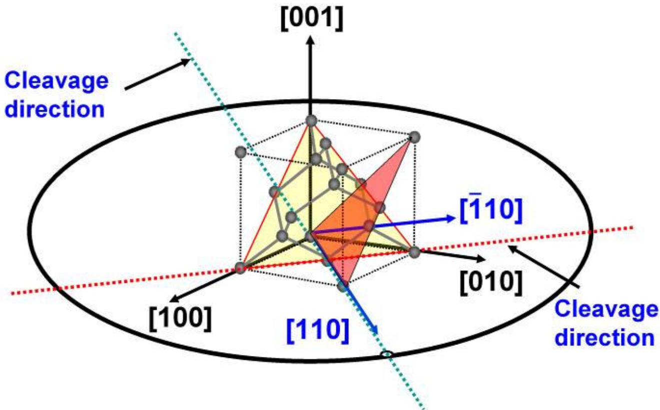


Figure 7: An illustration of crystallographic orientations in a silicon wafer with wafer normal parallel to [001] direction.

Diffraction Pattern: In a transmission electron microscope, electron diffraction by a crystallographic sample will result in several diffracted beams (or spots). An image of the diffracted beams or a diffraction pattern can be easily recorded in a TEM by switching to the lens settings to electron diffraction mode. The single-crystal spot diffraction pattern recorded with the electron beam parallel to [110] direction of the diamond cubic silicon lattice is shown in Figure 8.

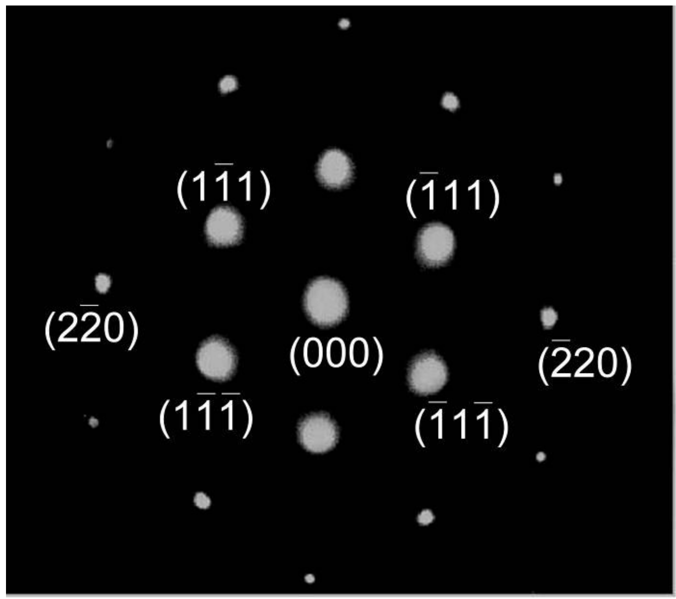


Figure 8: A single crystal diffraction pattern recorded with electron beam oriented parallel to the [110] direction in silicon. Selected diffraction spots from various crystallographic planes have been indexed.

Each spot corresponds to a set of crystallographic planes in the sample. The spacing of any spot or ring from the central (000)

or the 'transmitted' spot is inversely proportional to the crystallographic interplanar or the 'd' spacing corresponding to the diffraction spot.

Zone-axis imaging: In zone-axis imaging, the substrate is oriented along a direction parallel to a low-index i.e. high-symmetry crystal direction, with respect to the electron beam. Two common low-index crystal directions in silicon are [110] and [001]. The conventional bright-field TEM images are recorded selecting the transmitted beam or by intercepting the diffracted beams using a small objective aperture to achieve diffraction contrast, as is shown in Figure 9. In practice, the operator will switch to the diffraction mode to project the diffraction pattern on to the viewing screen or the CCD camera and insert the objective aperture around the trasmitted beam and switch back to imaging mode. The contrast observed in such image is a function (among other factors) of the size of the objective aperture. Examples of bright-field TEM images of polysilicon lines recorded with and without objective aperture are shown in Figure 10.

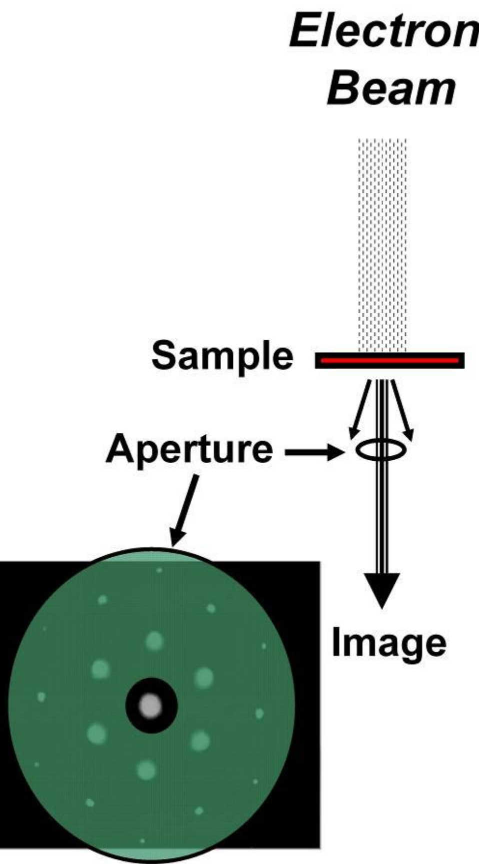
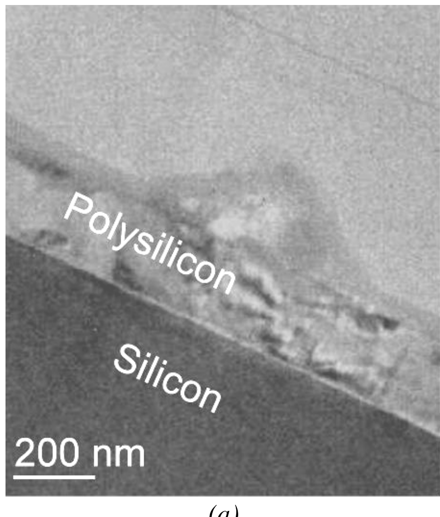
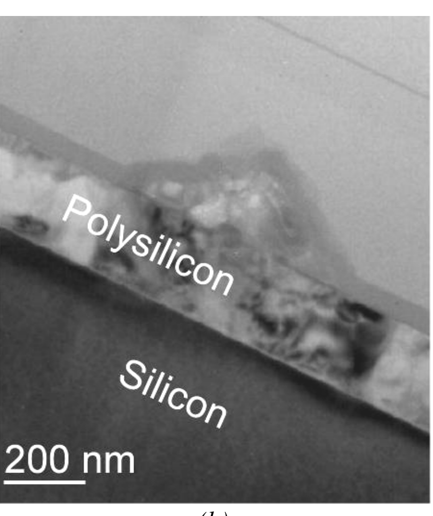


Figure 9: A schematic representation of the use of a small aperture to select transmitted beam for conventional bright-field imaging,



(a)



(b)

Figure 10: Diffraction contrast in [110] Si Zone-axis TEM images. (a) Image formed without objective aperture. Transmitted and diffracted beams contribute to the image. (b) Image formed by selecting only the transmitted beam with an objective aperture. The contrast in poly grains is significantly improved.

Applications of Diffraction Contrast

Diffraction contrast is unique to TEM and is sensitive to the crystallographic phase and thickness variations in the sample. The contrast introduced by diffraction in TEM images is dependent on the cryatallographic oriention of the sample with reference to the electron beam. This property is commonly used to delineate various components of semiconductor device and identify defects that introduce crystallographic phase variations in devices. An example of a diffraction contrast image of poly silicon lines is shown in Figure 11. The contrast variation in the polysilicon lines is a result of scattering incident electrons by grains of the polysilicon oriented

randomly with respect to the incident electron beam. Similar contrast variations are also noted in poly-crystalline silicide. However, the contrast is homogeneous in silicon-oxide and silicon-nitride because they are amorphous.

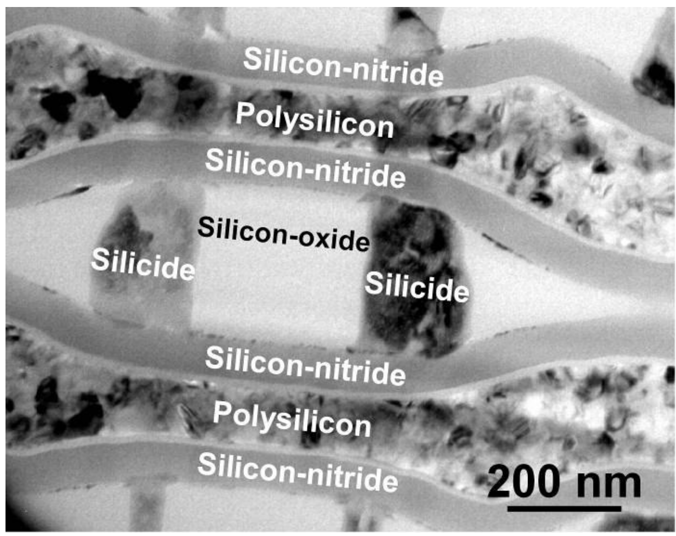


Figure 11: Planar TEM image of polysilicon lines. Contrast variation in polysilicon is a result of random crystallographic orientation of grains of silicon, with respect to the incident electron beam. No diffraction effects are noted within silicon-oxide and silicon-nitride because they are amorphous. Such amorphous layers are delineated based on the average atomic weight and density.

Defects in silicon substrate: The contrast due to diffraction in perfect single crystalline silicon substrate is usually homogeneous except for contrast because thickness, topographical variations, and changes in diffraction conditions because of slight warping of the thin section. Substrate defects such as dislocations and stacking faults disrupt the crystallographic planes. The strain around these defects causes strong contrast variations in the image. Examples of dislocations and stacking fault in substrate are shown in Figures 12 and 13.

Another example of a dislocation, shown in Figure 14, originating from the trench was observed in the silicon active region at the electrically leaky storage node of the failing SRAM bit. This planar section encompasses the silicon substrate and some parts of the gate-stack (the spacer, polysilicon and silicide).

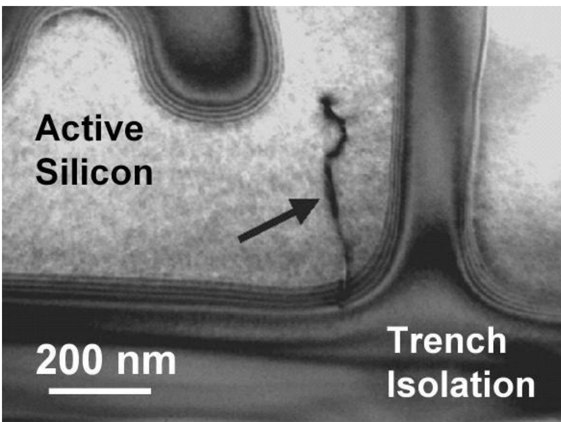


Figure 12: A diffraction contrast (planar) TEM image of a dislocation in the silicon substrate. All device layers have been milled away. The contrast from such defect is a result of strain from disruption of crystallographic planes of single crystal silicon, around the defect.

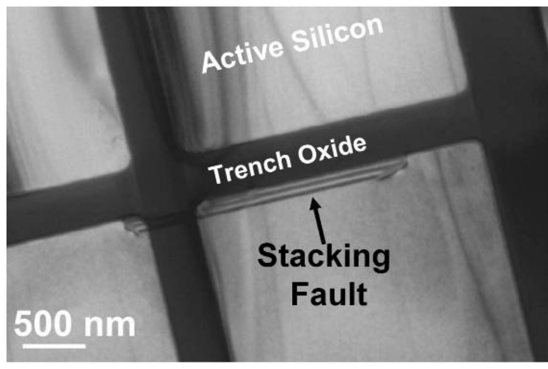


Figure 13: A diffraction contrast (planar) TEM image of a stacking fault in the silicon substrate, at the edge of active silicon.

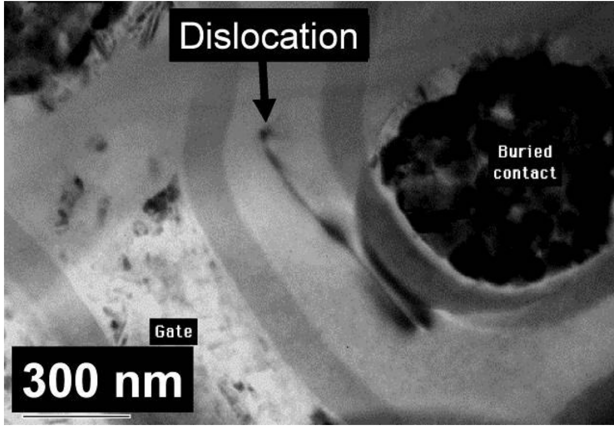


Figure 14: A diffraction contrast (planar) TEM image containing unique dislocation originating from the trench associated with the buried contact of a leaky node in an SRAM bit.

Gate oxide breakdown: Gate oxide breakdown sites are localized and difficult to detect. The large lateral area of analysis provided by TEM in conjunction with the strong diffraction contrast of the TEM can be used to observe the subtle effects (minor silicon damage) of gate oxide breakdown. The defect shown in Figure 15 was located in a SRAM cell. This procedure required careful milling during the final stages of the thinning to stop within 70 Å of the gate.

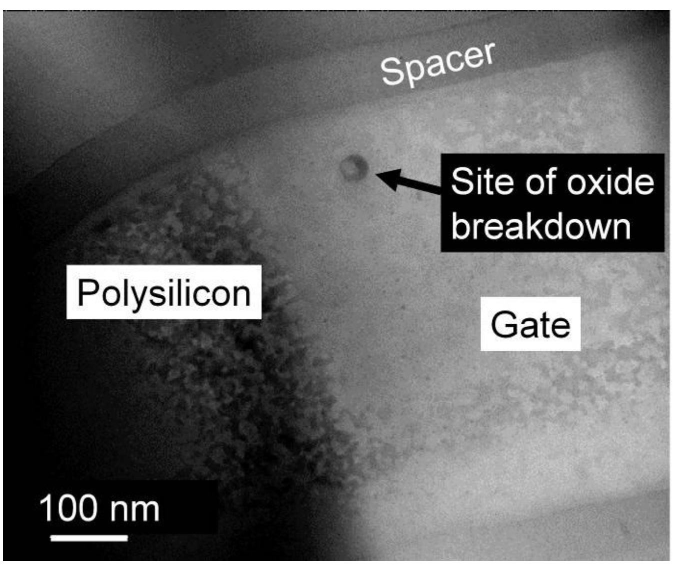


Figure 15: A planar TEM image of gate oxide breakdown site. The contrast variation at breakdown site is a result of damage in single crystal silicon substrate.

Stringers: A cross-section TEM image of stringers in the trench oxide, at the edge of active, is shown in Figure 16. Such stringers can cause leakage between adjacent circuit elements. The diffraction contrast variations in small stringer suggested that it is a polycrystalline material (silicide). The homogeneous contrast in the larger stringer is consistent with amorphous nitride, which is not expected to affect device performance.

Blocked contacts: A cross-section TEM image of polycrystalline material blocking the contact is shown in Figure 17. The contrast variations in the material blocking the contact suggested that material blocking the contact is poly crystalline.

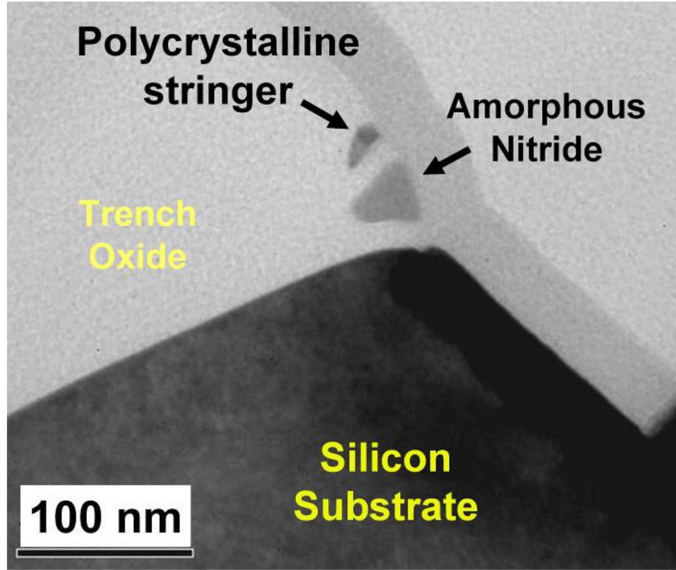


Figure 16: A cross-section TEM image of stringers in the trench oxide, at the edge of active. The contrast variations in the small stringer on the left suggest that it is a polycrystalline material (silicide). The homogeneous contrast in the large stringer on the right is consistent with amorphous nitride.

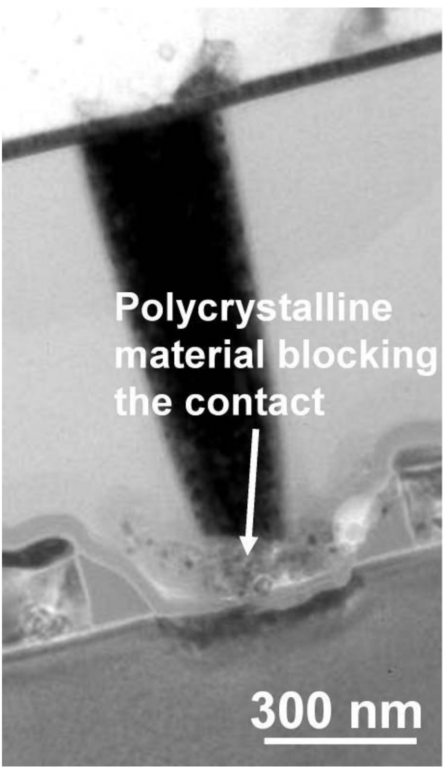


Figure 17: A cross-section TEM image of a contact blocked by polycrystalline material.

Thickness-Mass Contrast

Essentially all TEM thin sections have thickness variations, because of practical limitations of sample preparation techniques. In the case of semiconductor devices, various device features of differing densities are also composed of a number of elements with a wide range of atomic weights. When the incident electron beam interacts with such sample, the electrons can be inelastically scattered and/or absorbed to different extents, based on the average atomic weight, density and thickness variations. The resulting contrast is known as thickness-mass contrast. Such contrast is mostly independent crystallographic structure and is useful to delineate amorphous component in semiconductor devices such as silicon-oxide and silicon-nitride (noted in Figure 11). Thickness-mass effects contribute to contrast in images recorded from TEM sections that are inhomogeneous, which is often the case in semiconductor applications. For example, the contact in Figure 17 appears dark because of strong scattering of electrons by heavy atomic weight tungsten.

Phase Contrast

Phase contrast imaging is useful for measuring device features or identifying process defects that are only few nanometers in dimensions. Most of current generation transmission electron microscopes equipped with high resolution objective lens is capable of point to point resolution below 2.5 Å, whereas resolution below 1 Å can be achieved with the new aberration corrected instruments.

The phase contrast images are formed as a result of interference between the transmitted and diffracted beams included by the large objective aperture. A periodic two dimensional pattern resembling the atomic structure can be formed when three or more, non-collinear strong diffracted beams along with transmitted beam are symmetrically included with the objective aperture (Figure 18). The pattern or the fringe spacing corresponds to the interplanar spacing associated with the diffracted beams included in the aperture. In silicon based integrated circuits, cross-sectional high resolution images are usually recorded with the electron beam parallel to <110> direction using diffraction spots from {111} crystallographic planes. The periodic fringe spacing in such a image is equal to the planar spacing of {111} planes in silicon (3.14 Å). The fringe spacing is used as an internal magnification calibration to measure interfacial details at atomic level.

Applications of high resolution phase contrast imaging in semiconductor device include gate oxide thickness measurement (Figure 19), metrology of sub-nanometer features and identification of thin interfacial layers that can cause the devices to fail (Figure 20).

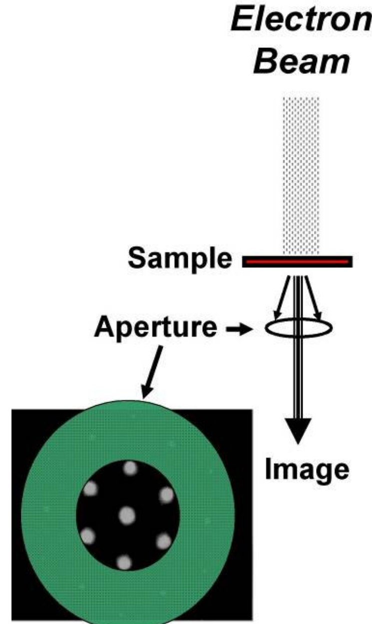


Figure 18: A schematic representation of the use of a large objective aperture for selecting multiple diffracted beams for phase contrast imaging.

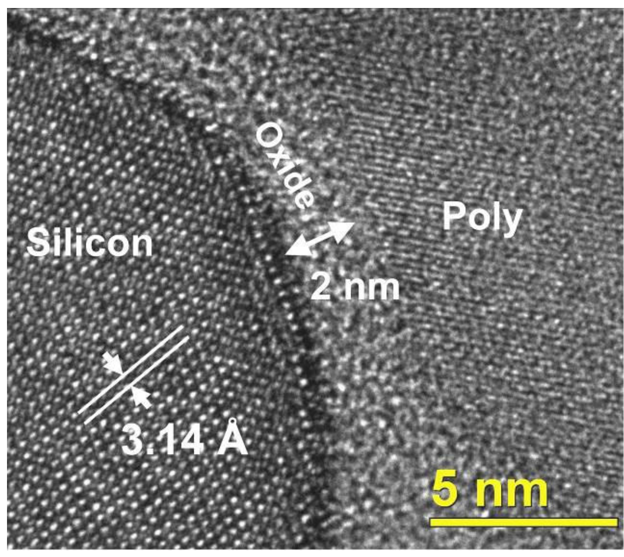
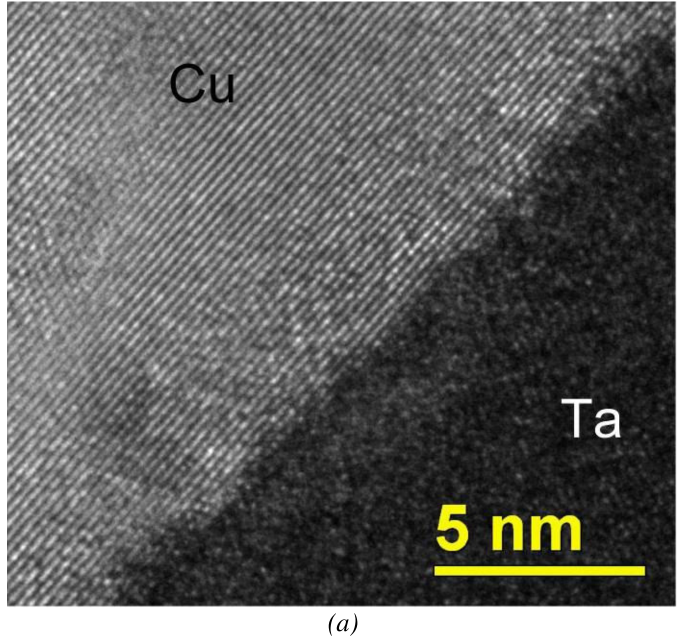
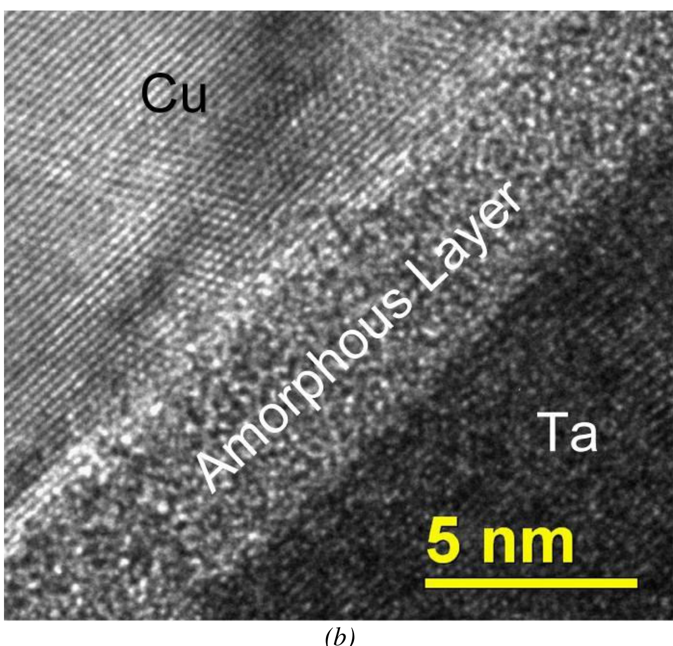


Figure 19: A phase contrast image of gate oxide wrap around the active silicon. The fringe spacing of 3.14Å corresponds to the (111) interplanar spacing of silicon, which serves as internal calibration.

The Fourier transform of the two dimensional periodic pattern in the image will produce an image with the spots of the diffraction pattern. The spots in Fourier transformed image can be used to identify the material based on the crystal structure information, by comparing crystallographic interplanar spacings extracted from diffraction patterns to the diffraction database.



(a)



(b)

Figure 20: High resolution phase contrast image of (a) normal copper-tantalum interface and (b) abnormal copper-tantalum interface separated by an amorphous layer that can lead to a resistive connection.

3.2 Scanning Transmission Electron Microscopy (STEM)

When the electron beam is transmitted through the sample, the electrons are elastically (no energy-loss) and inelastically (energy-loss) scattered. The fraction of inelastically scattered electrons increases with sample thickness. In parallel beam illumination mode, contribution from inelastically scattered electrons will negatively impact the quality of image because of the chromatic aberrations of lenses used to magnify the object and form the image. Scanning transmission electron microscopy (STEM) is less susceptive to such chromatic aberrations because lenses are not required for image formation or magnification. This advantage of STEM is very useful for imaging thick (>200 nm) samples, frequently encountered in failure analysis applications.

In scanning transmission electron microscopy, the electron beam is focused to a fine spot on the specimen and "scanned" over the area of interest [2, 29-31]. A STEM detector under the specimen is used to collect the transmitted electrons. The image is formed by converting the analog signal from the STEM detector to a pixel value in two-dimensional image. The magnification in a STEM image is the size of the area scanned to the size of the image on the computer monitor. The resolution is determined by the probe size.

A schematic representation of implementation STEM is shown in Figure 21. The STEM bright field (STEM-BF) image is recorded by collecting the transmitted electrons using the bright field detector. The bright field detector acts similar to an objective aperture used in the parallel beam illumination mode. As a result, the contrast variations observed in a STEM-BF image, shown in Figure 22, is similar to typical diffraction contrast observed in a bright field image recorded under parallel electron beam illumination.

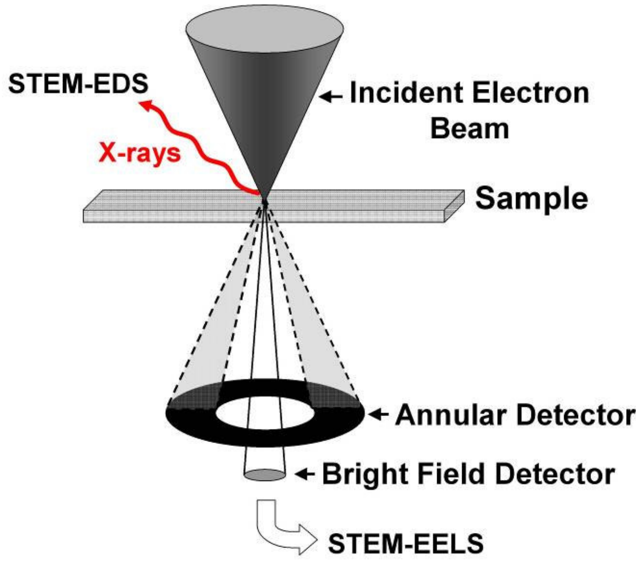


Figure 21: A schematic representation of STEM implementation in a transmission electron microscope. Electrons scattered at high angles are captured by the annular detector to form the Z-contrast image. X-ray photons are captured to form STEM-EDS elemental map. Bright field detector is retracted from the path of the transmitted beam for STEM-EELS elemental mapping.

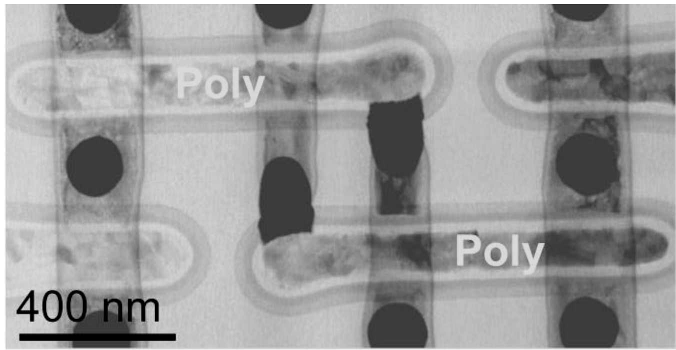


Figure 22: A planar STEM-BF image of a SRAM bit. The contrast is similar to conventional parallel beam bright-field image. Poly-silicon grains are delineated as a result of diffraction contrast.

The annular detector is usually installed at a specified port on the TEM column and is at a fixed height. The electromagnetic lenses of a TEM can be adjusted to selectively project electrons scattered at various angles on to the annular detector. The STEM-ADF image, shown in Figure 23, is formed by electrons scattered at lower angles, which carry crystallographic phase and grain structure information. The image shown in Figure 24 is recorded using electrons that are scattered at high angles (usually greater than 50 mrad). This operation mode is referred to as STEM high angle annular dark field (STEM-HAADF) imaging or Z-contrast imaging, where Z is atomic number. The contrast in STEM-HAADF image is primarily a result of mass variations in column(s) of atoms under the small electron probe. In a STEM-HAADF image, areas composed of heavy atomic weight elements (such as W) exhibit bright contrast because more electrons are scattered at high angles while areas composed of light atomic weight elements are relatively darker.

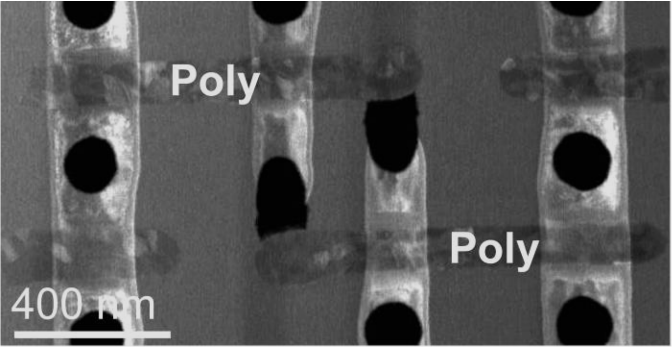


Figure 23: A planar STEM-ADF image of a SRAM bit (from Figure 22). The image is formed by collecting electrons scattered at low angles. Poly-silicon grain boundaries are delineated because of electron diffraction.

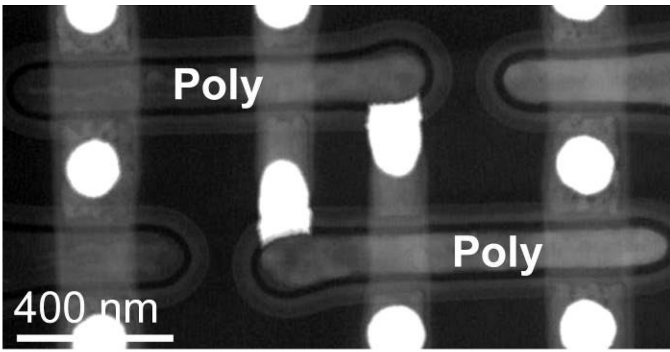


Figure 24: STEM-HAADF or Z-contrast image formed by collecting by electrons scattered at high angles (>50 mrad). In this image, the contacts are bright because it is composed of heavy atomic weight element (tungsten). The poly lines and oxide are darker because of relatively lower average mass.

Applications of STEM

Interfacial Defect: The enhanced mass contrast achieved STEM-HAADF technique can be used to identify interfacial defects by exploiting the mass change at the interface. A cross-section STEM-BF image of suspected via-metal interface, dominated by diffraction contrast, is shown in Figure 25. No conclusive evidence of interfacial issue is noted. The annular dark field Z-contrast image in Figure 26 clearly showed a dark interfacial layer (indicated by the arrow), which is consistent with an interfacial layer composed of light elements.

High Resolution STEM: Phase contrast imaging is commonly used for metrology of ultra-thin device features such as the gate oxide. In STEM mode, the electron lenses above the sample can be aligned to form a small probe (less than 2 nm diameter) to yield atomic resolution images [32]. Atomic resolution STEM-HAADF imaging can be used to clearly delineate interfacial boundaries, especially in thicker samples. This feature of STEM-HAADF is demonstrated by comparing Si/SiO_2 phase contrast TEM image and a STEM-HAADF image in Figure 27.

Phase contrast imaging in TEM is highly effective in thin TEM samples when there is change in crystal structure at the interface, for example the crystalline to amorphous transition at the Si/SiO_2 interface. However, it can be difficult to resolve the interface between two amorphous layers by phase contrast imaging. However, the Z-contrast STEM-HAADF imaging by virtue of its sensitivity to mass variations can be an excellent technique to resolve adjacent thin amorphous layers, such as silicon oxide and silicon nitride [29].

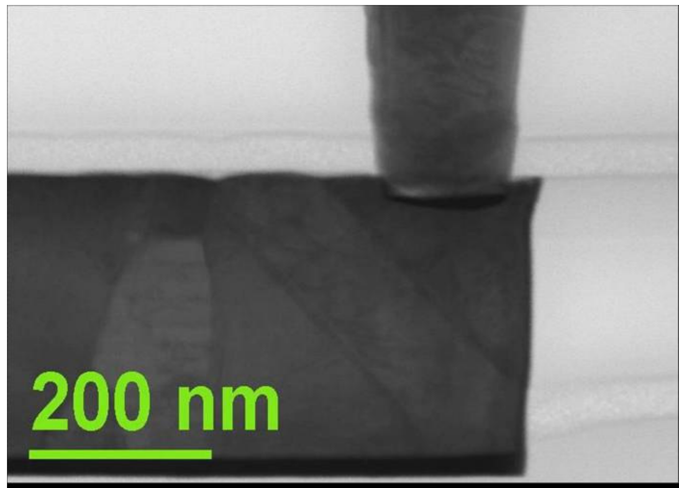


Figure 25: STEM-BF image of a via in a test structure. The contrast at the interface is masked by diffraction effects.

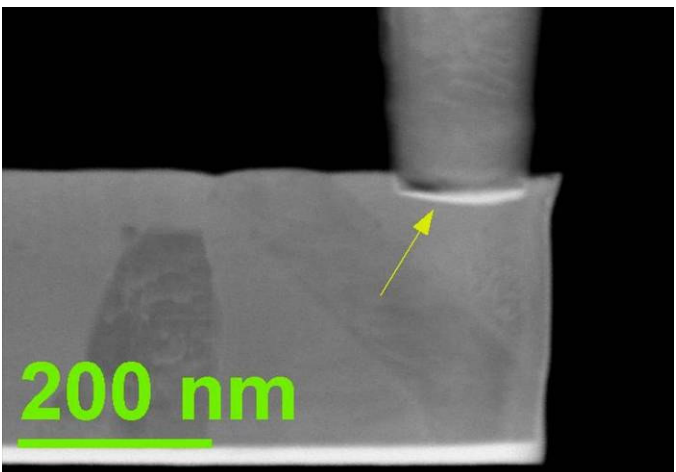
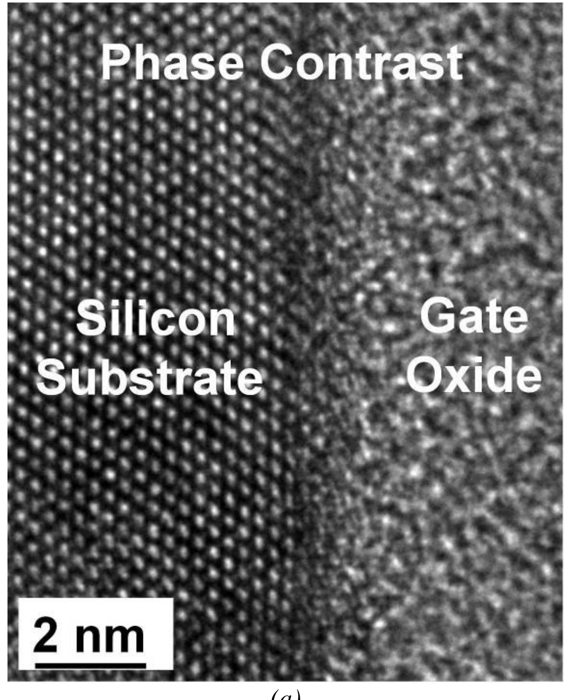


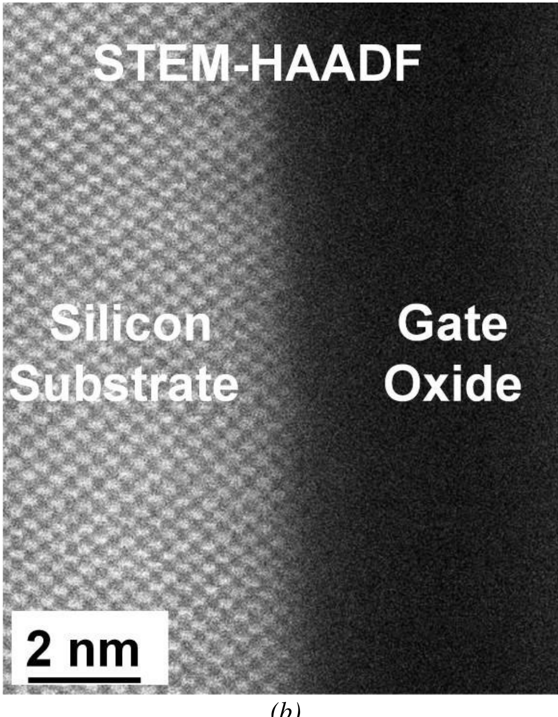
Figure 26: STEM-HAADF mass-contrast image clearly shows dark contrast at the interface, indicative of light elemental composition.

STEM Challenges

In the STEM mode, the image is acquired in a serial mode, one pixel at a time. As a result, STEM image acquisition can take several seconds or few minutes to achieve good signal to noise ratio. Any specimen drift or electro-magnetic (EM) interference can significantly degrade the resolution. To achieve consistent atomic resolution STEM imaging, it is critical that the sources of EM interference, acoustic noise, mechanical vibration and thermal instabilities in the room environment are minimized. Another challenge with the STEM technique is hydrocarbon contamination, because a converged high intensity electron beam is scanned over a small area. Contamination can be controlled by proper handling of the sample and the holder. Often an oxygen plasma cleaner is used to minimize or remove contamination after STEM imaging. However, it is important to note that plasma cleaning can also alter defect properties.



(a)



(b)

Figure 27: (a) High resolution phase contrast image of Si/SiO_2 interface in a thick sample. (b) STEM-HAADF image of the Si/SiO_2 interface of the same sample. The sharp contrast change at the interface in STEM-HAADF image permits easier identification of the interface.

STEM in an SEM

STEM imaging can be performed in scanning electron microscopes (SEM) after simple modifications [33-35]. In FA laboratories without access to a TEM, STEM in a SEM is an excellent economical option to achieve improved resolution using a SEM. It is important to note that this approach usually lacks other imaging and analytical capabilities of a TEM which can be a major handicap in FA samples. In addition, transmission electron microscopes operating at a high energy, 100 ~ 300 keV, are also better suited for thicker samples, often encountered in FA, than most scanning electron microscopes operating at 30 keV.

4. Elemental Analysis

Elemental composition of device features and defects can be characterized using energy dispersive spectroscopy (EDS) [1-2] or electron energy loss spectroscopy (EELS) [2-5, 36-39]. An illustration of the implementation of EDS and EELS in a transmission electron microscope is shown in Figure 1.

4.1 Energy Dispersive Spectroscopy (EDS)

In EDS, the characteristic X-rays emitted as a result of the interaction of incident electron beam with inner shell electrons of the atoms of various elements are collected by a spectrometer. EDS can be performed using scanning or transmission electron microscopes. In SEM based EDS, when the high energy electron beam is incident on the bulk sample, it can penetrate in to the sample and interact with sub-surface device features. Such interactions can easily degrade the spatial resolution of EDS at electron beam high accelerating voltages. In contrast, TEM based EDS offers inherently better spatial resolution for EDS because the interaction volume is limited by TEM section thickness, as is illustrated in Figure 28.

The characteristic X-rays generated by one element can be easily absorbed by a different element present in the same volume. Hence the sensitivity of the EDS technique is function of the element detected and the composition by volume in which the element is present. A reference sample with known composition is also required for meaningful quantitative analysis. Using state-of-the-art ultra-thin or windowless detector elements, atomic weight down to boron can be detected provided if a reasonable quantity is present within the volume analyzed.

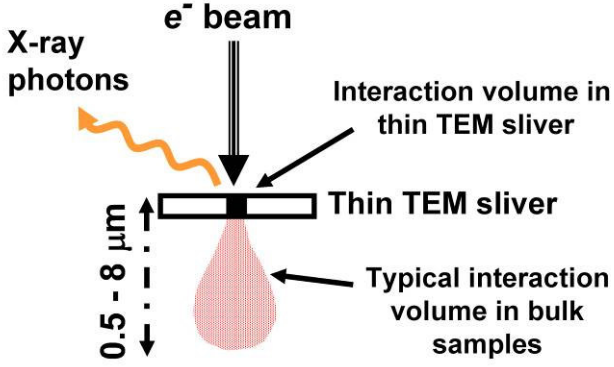
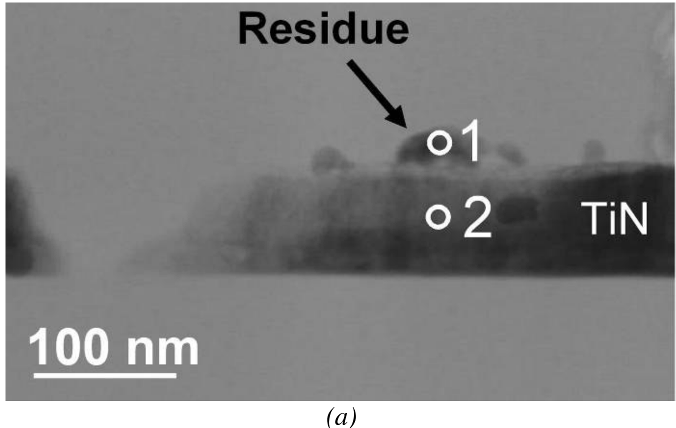
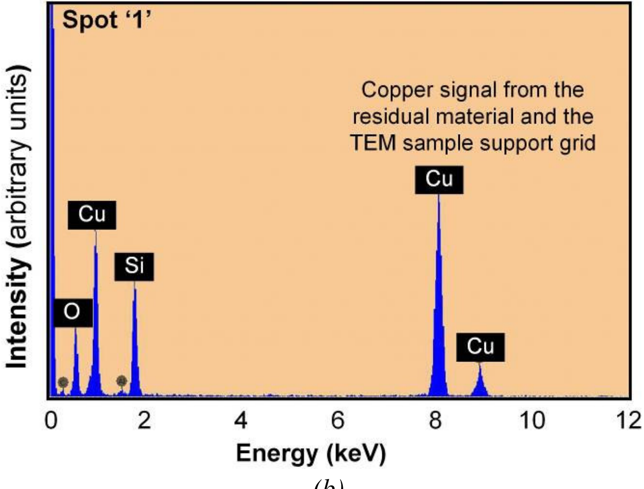


Figure 28: Comparison interaction volume in thin section of TEM sample and bulk material in SEM. The electrons can penetrate as deep as 8 µm in to the bulk sample at 30 keV [33] and degrade spatial resolution.. In a thin TEM section, the interaction volume is limited by the sample thickness (few tens to hundreds of nm).

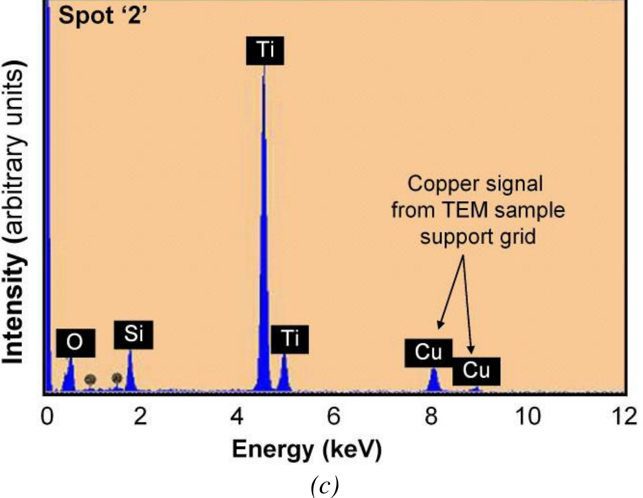
In failure analysis, EDS is commonly used to identify elemental composition of a defects and anomalies. Such analysis can be easily convoluted by spurious x-ray signals from material surrounding the defect. For example, copper grids (Figures 3, 5 and 6) are commonly used for support and handling of TEM samples. Frequently, spectrum from the thin TEM section supported by such grids will exhibit background copper peaks. The copper X-ray signal in those spectra is usually generated as a result of electrons scattered by the sample the sample interacting with the support grid. If the feature of interest is also composed of copper, the background copper signal can lead to ambiguous results. While this problem can be mitigated by using a support grid made of different element (such as molybdenum or nickel), it is hard to predict composition of the defect before the sample is prepared. When such spurious signal is present, the composition of the feature of interest can be determined by comparing the intensity of the peaks in the spectra from the defect and a site adjacent to the defect. An example of such analysis is shown in Figure 29. The relatively higher copper signal detected from spot ‘1’ in comparison to spot ‘2’ indicated that the residue is primarily composed of copper.



(a)



(b)



(c)

Figure 29: (a) STEM-BF image of residue on TiN layer. (b) The energy dispersion spectrum from residue at spot '1' show strong Cu peaks. (c) The energy dispersion spectrum from spot '2' exhibit strong Ti peaks (originating from the TiN layer) and the weak copper peaks. The copper signal at spot '2' is primarily a result of electrons scattered by the sample interacting with the copper grid supporting the sample.

4.2 Electron Energy Loss Spectroscopy (EELS)

As noted previously, as the electron beam is transmitted through the sample, a variety of beam-specimen interactions occur. In EELS, the elastically (zero-loss) and inelastically (energy-loss) scattered electrons from the feature of interest in the sample are directed into a spectrometer attached to the bottom of the TEM column [5, 37]. The spectrometer consists of a sector magnet and a detection system. The sector magnet deflects the transmitted electron beam by 90 degree (Figure 1). During this process, electrons with different energies are deflected to different extent by the magnetic field. The process results in an electron energy loss (EEL) spectrum.

The amount of energy lost by inelastically scattered electrons or the 'energy-loss' electrons depends on various inelastic scattering processes that occur within the sample. An example of EEL spectrum, in the low energy-loss range, from silicon-oxide based dielectric overlaid on a spectrum from silicon substrate (with shaded area) is shown in Figure 30. The x-axis of the plot represents the energy-loss (ΔE) from inelastic scattering events, when the incident electron beam with energy (E) of 200 keV is transmitted through the sample. The peak at $\Delta E = 0$ eV, known as zero-loss peak (ZLP), usually dominates the EEL spectrum when the sample is reasonably thin and free of sample preparation induced damage. The low energy-loss peaks for silicon at $\Delta E = 17$ eV and silicon-oxide dielectric at $\Delta E = 24$ eV are known as plasmon peaks. These low energy-loss plasmon peaks (typically at $\Delta E < 50$ eV) are a result of weakly bound valence electrons of the material in the sample collectively interacting with electron beam [2, 5]. These interactions will also yield multiple plasmon peaks in thicker samples [5], which is observed in the EEL spectrum from silicon in Figure 30.

During the inelastic scattering process, electrons from the incident beam also lose energy because of interaction with the inner shell (K, L …) electrons of atoms. These energy-loss electrons appear as a step or an edge in the higher energy-loss regime (above 40 eV to few thousands of eV) of EEL spectrum and are referred to as the ionization-edge. The ionization-edge reflects the atomic structure of element and is useful for elemental analysis. The EEL spectrum, from a sample containing N, Ti, and O with ionization-edges at 401eV, 456 eV and 532 eV, respectively, is shown in Figure 31. It is also important to note the background under the ionization-edges of various elements. This background signal is primarily a result of tails from plasmon peaks and ionization-edges from lower energy-loss and increases with sample thickness. In thick samples, the background electron counts can reach extremely high numbers and render EELS based elemental analysis impractical [38].

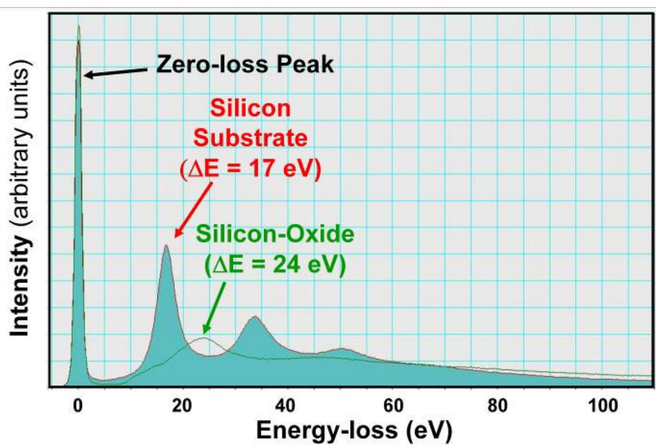


Figure 30: Examples of EEL spectrum in the low energy-loss regime. EEL spectrum from silicon-oxide based dielectric is superimposed on EEL spectrum from silicon. EEL spectrum from silicon exhibits several plasmon peaks (at multiples of 17 eV) because it is recorded from a thicker area of the sample.

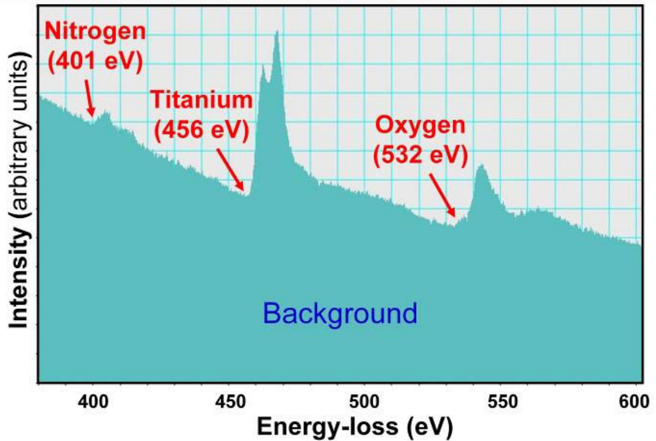


Figure 31: EEL spectrum from an area of sample composed of N, Ti and O. Background under ionization-edges of various elements is primarily a result of tails of plasmon peaks and ionization-edges of other elements with lower energy-loss.

4.3 EELS *vs.* EDS

TEM based EDS offers superior spatial resolution because of reduced interaction volume in a thinned electron transparent TEM section (~ 20 nm to 250 nm). However, when the electron beam is incident on the sample, the X-ray photons carrying the elemental information propagate along all directions is space. In practice, only a fraction of these X-ray photons are collected by the X-ray detector because of the difficulties in inserting a large detector in the microscope column without affecting other capabilities of the TEM. The EEL spectrometer does not have this problem because it is introduced in the path of the electron beam that has gone through the sample or attached to the bottom of the column. And, significantly high percentage of electrons carrying elemental information from the area of interest can be directed into the spectrometer. As a result, EELS offers better signal collection efficiency to detect elements from very limited volume of material at the area of interest, which is often the case in advanced technologies with shrinking device features.

The X-ray energy-resolution achieved by standard X-ray detectors is approximately two orders of magnitude worse than energy-resolution of an EEL spectrometer (< 1 eV). In semiconductor device analysis, the poor energy resolution of EDS can often yield in inconclusive results because the characteristic X-ray peaks of various elements contained in the sample could not be resolved. In addition, EDS analysis of light elements can also be impacted by overlap of low energy X-ray peaks with electronic noise of the detector. In spite these issues, EDS has remained more popular than EELS because it can work well in thick and imperfect samples commonly encountered in semiconductor failure analysis.

In recent years, EELS based elemental analysis has become more popular in semiconductor FA because of the necessity to analyze shrinking device features (smaller volume that require higher sensitivity) in advanced technologies. In addition, advancements in FIB based sample preparation techniques permit ultra-thin (< 50 nm) site specific sample preparation with minimal damage or contamination.

It is also important to note that the success of any elemental analysis depends on various practical limitations such as composition of material analyzed, matrix the surrounding element that is being detected, volume of material analyzed and sample preparation challenges. The relative strengths of EELS and EDS in TEM should be exploited by optimizing these factors and constraints, and collecting complementary data using all of these techniques to perform accurate elemental analysis.

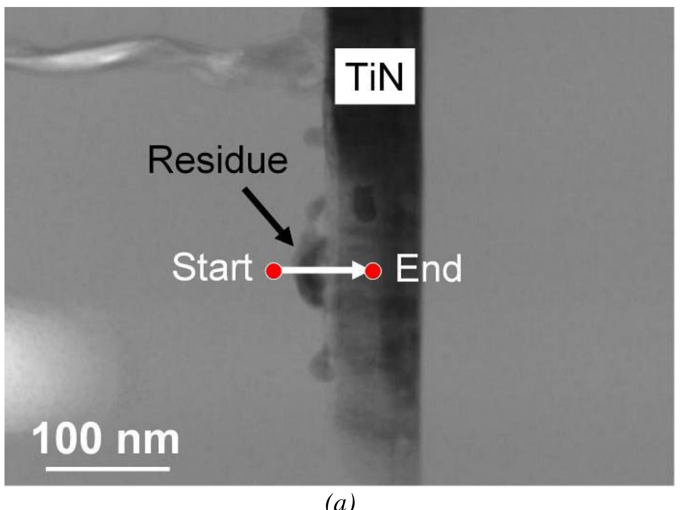
4.4 Elemental Mapping

STEM-EDS and STEM-EELS

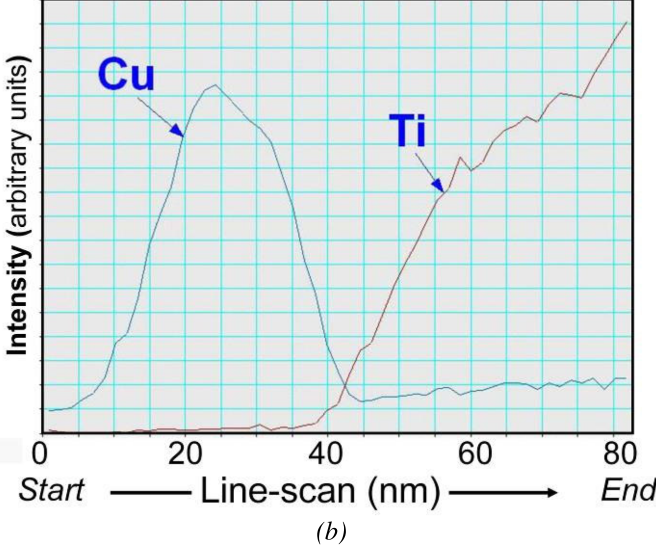
In STEM mode, the transmission electron microscope can be used to record the “energy” based image from the specimen. The energy from the sample information manifests itself as the energy of X-rays from various areas of the specimen or the energy-loss in transmitted electrons from various areas of the specimen. These signals carry a wealth of elemental and chemical information about the specimen.

STEM-EDS is a technique in which X-ray signals are collected from the area scanned by STEM. The EDS system processes and stores the data. The scanned area could be in the shape of a line (line-scan) or a rectangle (area-scan). The output is either in the form of intensity plots for x-ray signals

in case of line-scan or in the form of an elemental map image. Figure 32 demonstrates the advantage of line-scan over the spot mode EDS. The line-scan in Figure 32(b) clearly shows elevated Cu signal on the residue.



(a)



(b)

Figure 32: Example of the EDS line-scan analysis. (a) STEM-BF image of residual material (b) A plot of Cu-Kα and Ti-Kα x-ray signal intensities extracted from energy dispersive spectral data recorded in the STEM-EDS line-scan mode. Residual material clearly shows higher copper signal indicating that is primarily composed of copper.

STEM-EELS is a technique in which transmitted electrons are collected by the EELS spectrometer. The EELS system processes and stores the data. Again, line- and area-scans are the possible modes. The results can be displayed as intensity plots or as elemental maps.

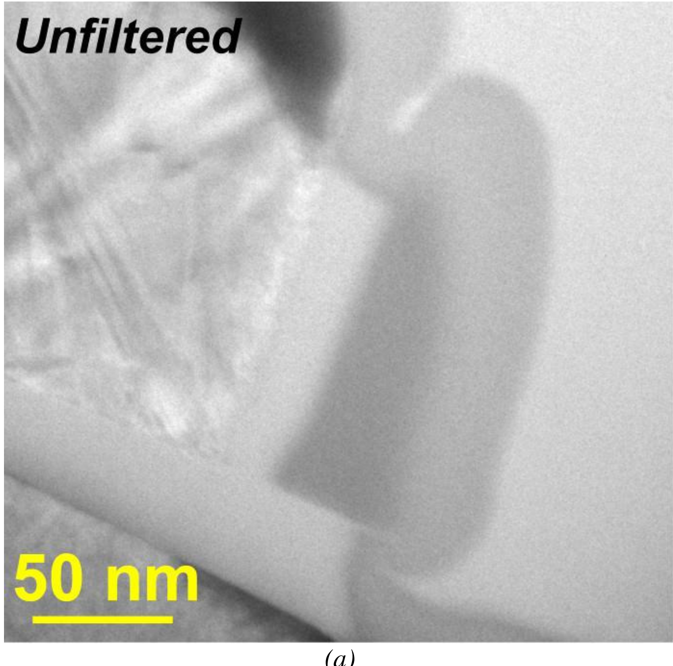
The data acquisition process for STEM-EDS or STEM-EELS map can be time-consuming because the data is acquired one pixel at a time in a serial mode. The total acquisition time depends on various factors including the size of the map (pixels) and the dwell time of the beam. The minimum required dwell time is dependent on the electron beam intensity, element of interest mapped and composition of feature to achieve meaningful signal to noise ratio. In typical semiconductor FA applications, the total acquisition can range from several tens of minutes or several hours. In addition, hydrocarbon contamination caused by exposure to electron beam and specimen drift can also affect the analysis. Following section describes the EFTEM based mapping technique to avoid these issues.

Energy-Filtered TEM (EFTEM)

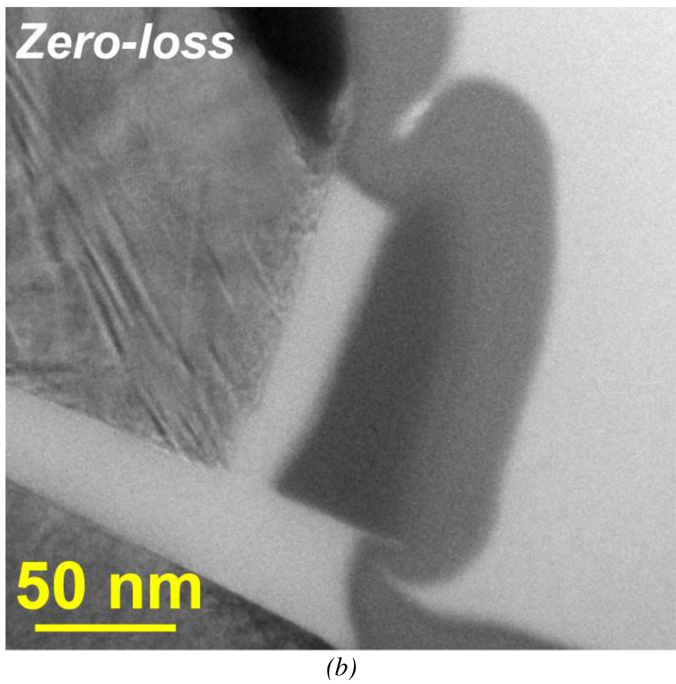
In energy-filtered TEM, the parallel beam imaging mode in a TEM is coupled with principles of EELS to yield a filtered image or an elemental map. Unlike the serial pixel by pixel acquisition in STEM based mapping, the EFTEM image is recorded in parallel so that an energy-filtered image or an elemental map can be obtained in several seconds or few minutes. The acquisition time will depend on various factors including the resolution of the image, signal strength of ionization-edge, sample quality etc.

Energy filters are available from various vendors in either in-column (introduced after the sample in TEM column), or post-column (attached to the bottom of the TEM column). These filters utilize a series of magnetic prisms or magnetic sectors coupled with a series of quadrapole and sectapole magnets [5, 37, 39]. Energy filters change the direction of electron beam by 90 degrees at least once to disperse the electron beam that has gone through the sample on the basis of energy of electrons. The following sections will focus on various EFTEM based techniques.

Zero-Loss Imaging: In zero-loss imaging, the image is formed by using only those electrons contributing to the zero-loss peak in the EEL spectrum shown in Figure 30. The inelastically scattered energy-loss electrons are excluded from the image with the aid of a slit (width calibrated in eV), which allows only the zero-loss electrons to pass through. In failure analysis applications, thicker (>150nm) samples have to be prepared when there is uncertainty about location of the defect within a circuit. Features in unfiltered images from such thick sample can be blurry and washed-out because of the noisy background from inelastically scattered electrons. The contrast in thick samples can be significantly enhanced in a zero-loss image (as is demonstrated in Figures 33 and 34).

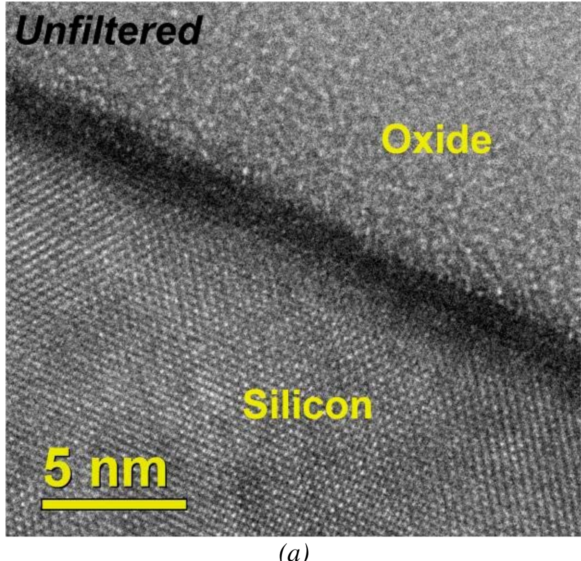


(a)

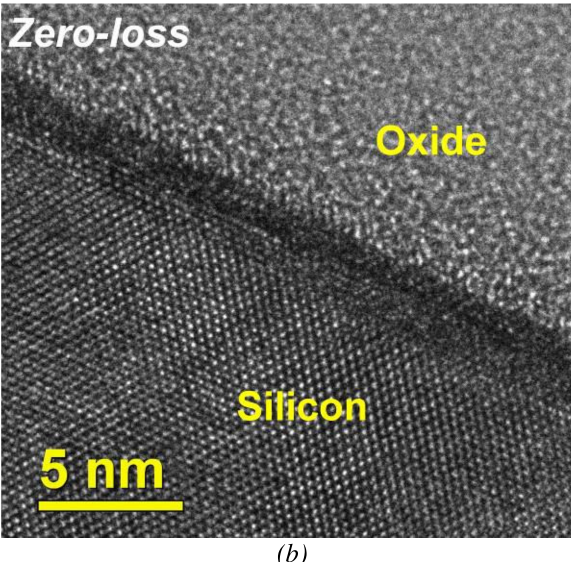


(b)

Figure 33: (a) Unfiltered TEM image from a thick sample is affected by inelastically scattered electrons. (b) Zero-loss image recorded using 10 eV energy window clearly delineates the boundaries of various device components.



(a)

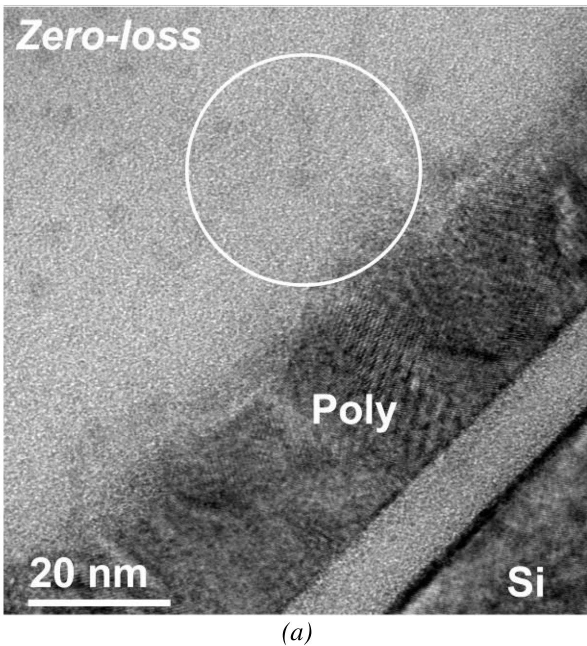


(b)

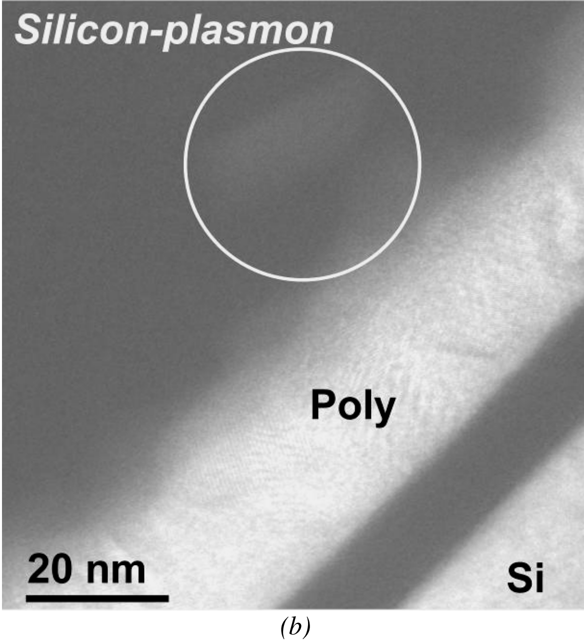
Figure 34: (a) Unfiltered high resolution phase contrast TEM image from a thick sample. (b) Zero-loss high resolution phase contrast TEM image recorded using 10 eV energy window. Significant contrast enhancement noted in zero-loss image is a result of excluding background noise because of inelastically scattered electrons.

Low-loss Imaging: In silicon based integrated circuits, the silicon plasmon peak at ΔE = 17 eV and silicon-oxide plasmon peak at ΔE = 24 eV (Figure 30) can be used to delineate device features and defects. By selectively forming an image using the energy-loss electrons contributing to the silicon-plasmon peak, one can delineate silicon and silicon-oxide feature [40] by introducing bright contrast in silicon containing areas. Areas containing the silicon-oxide will be dark in the silicon-plasmon image. Examples of the zero-loss and the silicon-plasmon images of a gate stack are shown in Figure 35. The silicon-plasmon image recorded using 5 eV slit

centered on plasmon peak (at ΔE ~ 17 eV), clearly shows the silicon grain embedded in oxide (in circled area) above poly-silicon (Figure 35(b)). This grain is invisible in zero-loss image.



(a)



(b)

Figure 35: (a) Zero-loss electrons image of a gate stack. No unique contrast indicative of silicon grain is visible in the circled area above poly. (b) Silicon-plasmon image recorded by selecting the plasmon peak (using 5 eV energy window at ΔE = 17 eV). Regions containing silicon and poly-silicon are bright in this image. The poly- silicon grain in circled region in the oxide is clearly visible.

EFTEM Elemental Mapping: Another benefit of EFTEM for failure analysis comes from elemental (B, C, O, N, Ti, Co, Ni, F, Cu, etc.) mapping of anomalies and defects in semiconductor devices, using the energy-loss electrons contributing to their ionization-edges. The electron counts at ionization-edges of elements can be several orders of magnitude weaker than zero-loss or plasmon peaks. The background signal under the ionization-edge of elements of interest in the EEL spectrum (Figure 31) has to be subtracted to extract the elemental information. In thick samples, background will dominate the EEL spectrum and limit any meaningful extraction of elemental maps. In thin samples, the background can be subtracted either by 'jump-ratio' method or 'three-window' method [5]. In the 'jump-ratio' method, elemental map is obtained from the ratio of two images recorded by selecting an energy window before and after the ionization-edge of the element of interest. This method is not computationally intensive and is more suited for qualitative mapping low elemental concentration from noisy images. In the 'three-window' method, the background subtraction is achieved by using two images recorded by selecting two energy windows immediately before (pre-edge 1 and 2) ionization-edge and one image by selecting an energy window immediately after (post) ionization-edge (Figure 36). This approach offers better background subtraction but may fail when the images are noisy. Since the pre-edge and post-edge images are acquired sequentially, the background subtraction algorithm should also account for specimen drift that may occur during the collection process.

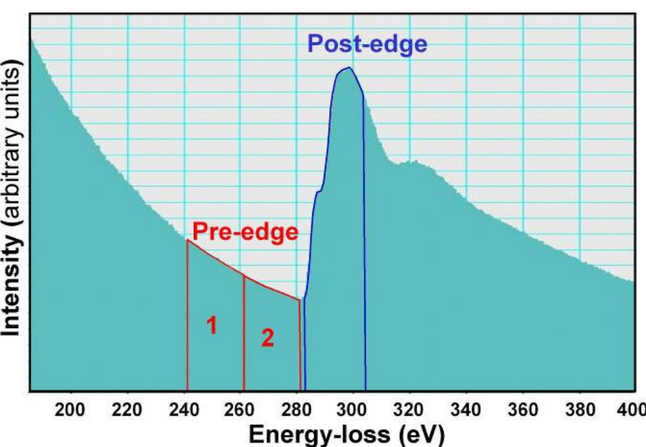
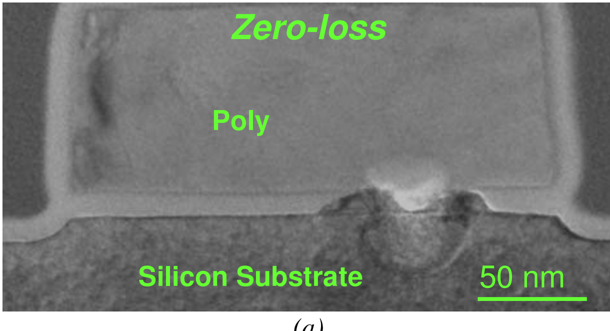
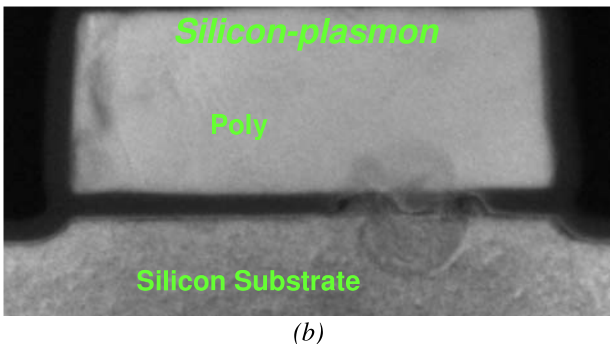


Figure 36: Representation of the three-window method at the ionization-edge of carbon. The background is estimated from pre-edge 1 and 2 images, and subtracted from the post-edge image..

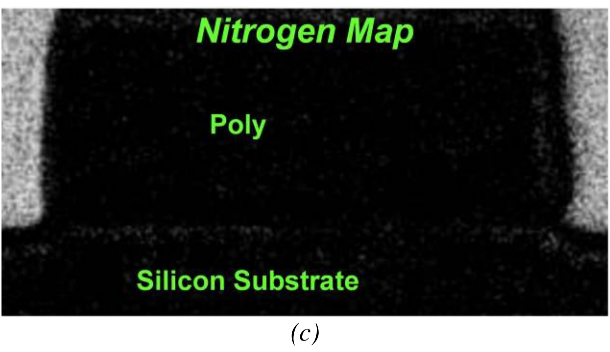
An illustration of application of EFTEM for elemental analysis of gate oxide breakdown is shown in Figure 37. The nitrogen and oxygen element maps were derived by employing the 'three-window' method, using 30 eV energy-windows for pre- and post-edge images



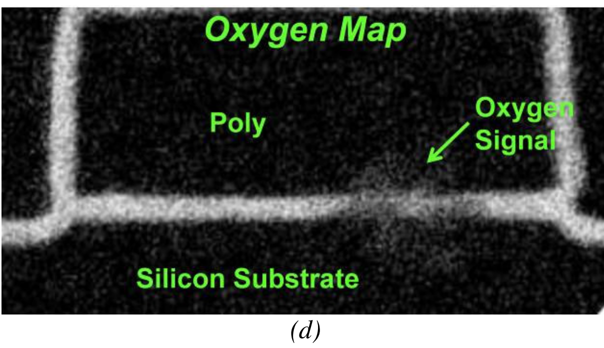
(a)



(b)



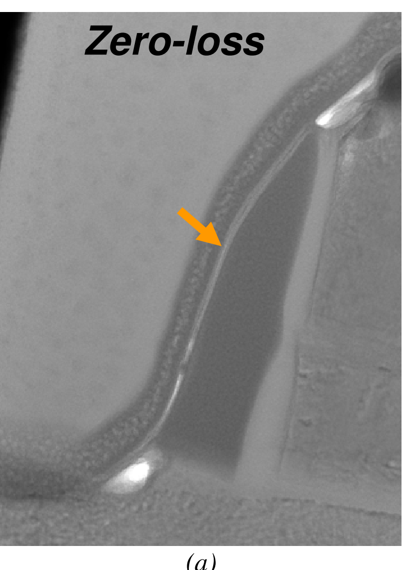
(c)



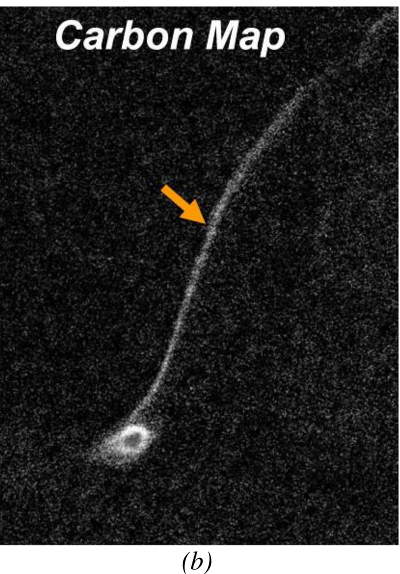
(d)

Figure 37: (a) Zero-loss TEM image of a gate oxide breakdown, (b) silicon-plasmon (5 eV slit at 17 eV energy-loss) image, (c) nitrogen map (white areas are rich in N), (d) oxygen map (white areas are rich in O). The analysis showed that the site of the gate oxide breakdown is primarily composed of Si and O.

Elemental analysis of defects and anomalies containing carbon can often lead to inconclusive results because of spurious carbon signal from hydrocarbon contamination caused by exposure to electron beam. There can be elevated contamination rate during STEM-EDS or STEM-EELS because of the focused electron beam dwelling on every pixel for a significant period of time. The contamination rate and the spurious carbon signal can be minimized in EFTEM because of the use of a parallel beam of electrons and relatively shorter e-beam exposure times. An example of application of EFTEM to identify a carbon containing residue is shown is Figure 38.



(a)



(b)

Figure 38: (a) Zero-loss electron image of the anomalous layer (indicated by the arrow) on the spacer side-wall. (b) Carbon map showed elevated levels of C in the anomalous layer.

5. Off-Axis Electron Holography

Off-axis electron holography [22, 41-47] can be used for 2D dopant profiling of shallow p-n junctions in source/drain regions of the transistors. When an electron transparent single crystalline substrate containing the p-n junction is illuminated with a coherent electron beam or a 'plane electron wave', the electrostatic potential gradient across the junction causes a local phase shift in the electron wave. This phase shift can be extracted from holograms recorded using off-axis electron holography technique.

A transmission electron microscope equipped with a coherent field emission source, a Lorentz or an objective mini lens with a large field of view and a rotatable electron biprism wire (> 1 μm diameter) is required for dopant profiling using off-axis electron holography technique. An illustration of implementation of off-axis electron holography is shown in Figure 39. In the holography operation mode, the standard objective lens of the microscope is turned off and a Lorentz or an objective mini lens is turned on. A flat electron transparent sample is partially inserted in the path of the coherent electron beam or the 'plane electron wave' such that a fraction of electrons are allowed pass through sample (modulated wave) and rest of electrons transmit through vacuum (reference wave) of the electron column. The biprism wire is inserted below the Lorentz or an objective mini lens and oriented parallel to the edge of the sample. A positive bias to the biprism wire will cause the modulated wave and reference wave to interfere with each other to produce a hologram. A reference hologram after removing the sample from the path of the beam is also recorded. The phase image containing the dopant profile information is then extracted from the holograms by Fourier analysis. An example of off-axis electron holography application is shown in Figure 40.

The quality of the sample is critical for the success of electron holography in failure analysis. It is important that the sample is flat and homogeneous at the location of junction of interest. For best results, a sample thickness between 160 nm [22] and 300 nm is preferred, which is easily achievable using latest FIB based sample preparation techniques. The sensitivity of off-axis holography can be seriously degraded by of additional material such as the trench oxide, dead amorphous layer from ion beam exposure, sample thickness variations resulting from uneven milling ('curtain' or 'water-fall' effect) or redeposition of various sputtered materials. The thickness of amorphous dead layer can be reduced low energy (< 5keV) ion beam cleaning of the sample. The undesirable curtain' or 'water-fall' effect can be eliminated by FIB cross-sectioning the sample from back-side [48], which involves additional sample manipulation to orient/flip the sample to position silicon substrate above the device features. The homogeneous single crystal substrate can be milled evenly to achieve a uniform flat sliver at the site of p-n junction.

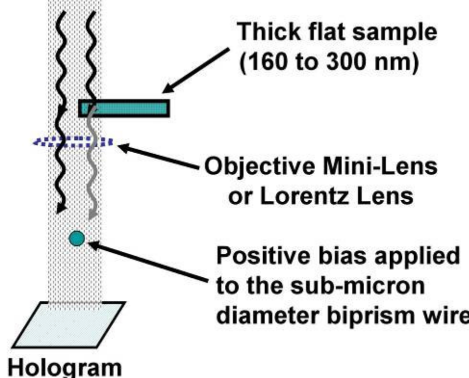


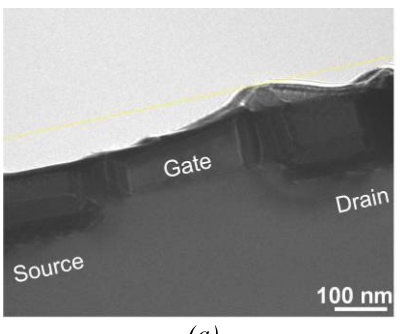
Figure 39: An illustration of implementation of off-axis electron holography in a TEM for two-dimensional dopant profiling. A Lorentz or objective mini lens is used for achieving a large field of view. The coherent electron beam passing through the sample and vacuum interfere with each other as a result of positive bias applied to the biprism wire, aligned parallel to the edge of the sample to produce the hologram.

Another sample requirement for off-axis electron holography is that the p-n junction of interest should be close to edge of the sample (typically less than 0.5 μm), to allow the reference wave through the vacuum. The edge of the sample is approximately 300 nm from the junction in Figure 40(d). In this case, all metal layers and part of polysilicon gate have been removed.

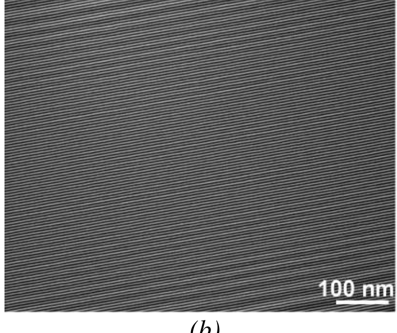
6. Summary and Conclusions

The role of TEM in the failure analysis of semiconductor devices is discussed. The principles of various TEM imaging and elemental analysis techniques/modes are described using examples encountered in failure analysis. The origin of different image contrast mechanisms, their interpretation, and analytical techniques for composition analysis are reviewed.

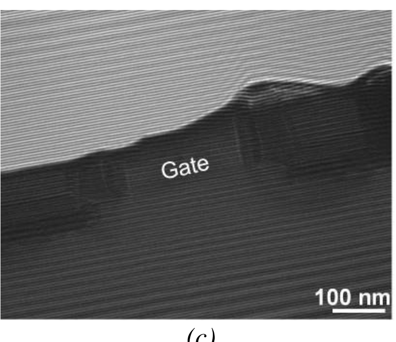
With parallel beam illumination, the advantages of electron diffraction in crystalline samples can be exploited to image substrate crystallographic defects such as dislocations and stacking faults. Other defects that alter or interrupt crystallographic structure at the failure site can also be visualized using diffraction contrast. Examples of such defects include stringers and interfacial defects. At high magnifications, phase contrast imaging is commonly used for metrology and identification of sub-nanometer sized defect. Thickness-mass contrast is always super-imposed on diffraction and phase contrast images and offers the ability to delineate various amorphous layers based on the average atomic weight and density differences.



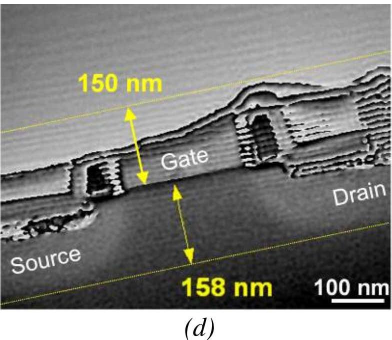
(a)



(b)



(c)



(d)

Figure 40: (a) A cross-section TEM image NMOS transistor, tilted few degrees away from [110] Si zone-axis to minimize diffraction effects. (b) An image of reference hologram without the sample in the path of the beam, recorded after positively biasing the biprism wire. (c) Image of the hologram recorded with sample partially inserted in the path of the beam. (d) The phase image showing the dopant profile signal, extracted using by Fourier analysis from images shown in (b) and (c).

In STEM mode, a converged probe is scanned over the sample and the image is formed by acquiring transmitted electrons, one pixel at a time. Advantages of STEM include the capability to image thicker sample encountered in failure analysis and dedicated Z-contrast or HAADF imaging by collecting electrons scattered at high angles. High resolution STEM-HAADF imaging can be used to delineate interfacial defects.

EDS and EELS based elemental analysis techniques are useful for identifying elemental composition of defects and device features. EDS technique can work well with thick imperfect TEM sections. However, the energy resolution of EDS detector may limit its applications in certain cases. The superior energy resolution of EELS offers better light element analysis capability, but also requires a thin damage-free sample. The relative strengths of EELS and EDS in TEM can be exploited by optimizing various factors and collecting complementary data to perform accurate elemental analysis. EDS and EELS can be coupled with STEM to produce line-scan or two dimensional element maps of defects.

The zero-loss EFTEM imaging is useful for enhancing contrast in imperfect thick TEM samples encountered in semiconductor failure analysis. In silicon IC's, the contrast in a silicon-plasmon image can be used to delineate silicon and poly-silicon device features and anomalies that are invisible in the standard TEM image. Since EFTEM is based on EELS, it also offers superior energy-resolution for elemental mapping. The elemental maps can be derived by recording filtered images at user selected energy windows in the vicinity ionization-edges of elements of interest (in thin samples). Also, because such elemental mapping is performed in parallel electron beam TEM imaging mode, localized hydrocarbon contamination at defect site is minimized to permit mapping of anomalies containing carbon. EFTEM based elemental mapping can lead to inconclusive results as the samples gets thicker, because of the increasing background under the ionization edges.

Electron holography offers the capability to visualize dopant profiles in source and drain regions of the transistors. The sample requirements are stringent for dopant profiling. Defect free samples can be prepared using state-of-art FIB instruments or ion millers.

Finally, it should be noted that the successful application of TEM in failure analysis depends on how precisely the fault has been isolated and whether an appropriate sample can be prepared at the isolated site. Considering that the sample preparation process involves destruction of area around the target site, it is often advisable to prepare thicker sample to minimize the chance of destroying the defect. The strengths of various TEM techniques outlined in this article should then be optimized to extract useful information from the sample.

Acknowledgements

The authors would like to thank Tony Chrastecky, Khiem Ly, Charles Buckley, Gary Clark, Wan Foong Kho (Freescale, Malaysia) and the members of the Austin Product Analysis Laboratory for their technical contributions and support.

References

1. D.C. Joy, A.D. Romig, J.I. Goldstein, Principles of Analytical Electron Microscopy. Plenum Press, 1986.
2. D. B. Williams and C. B. Carter, Transmission Electron Microscopy, Plenum Press, NY, 1996.
3. R.F. Egerton, Electron Energy Loss Spectroscopy in the Electron Microscope, Plenum Press, 1996.
4. M. M. Disko, C. C. Ahn, and B. Fultz, Transmission Electron Energy Loss Spectrometry, TMS, 1992.
5. R. Brydson, Electron Energy Loss Spectroscopy, Taylor & Francis, 2001.
6. R. S. Rai and S. Subramanian, Progress in Crystal Growth and Characterization of Materials, 55, 2009, p.63.
7. R. J. Young, E. C. G. Kirk, D. A. Williams, and H. Ahmed, Mater. Res. Soc. Symp. Proc. 199**,** 1990, 271.
8. S. Morris, S. Tatti, E.Black, N. Dickson, H. Mendez, B. Schwiesow, and R. Pyle, Proceedings from the 17th International Symposium for Testing and Failure Analysis, 1991, p417.
9. E. C. G. Kirk, D. A. Williams, and H. Ahmed, Inst Phys. Conf. Ser. No. 100 Section 7, 1989, p501.
10. D. M. Schraub and R. S. Rai, Progress in Crystal Growth and Characterization of Materials, Vol 36, 1998, p99.
11. L. Tsung, A. Anciso, R. Turner, T. Dixon and N. Holloway, Proceedings from the 27th International Symposium for Testing and Failure Analysis, 2001, p299.
12. R. M. Anderson and J. Benedict, Mater. Res. Soc. Symp. Proc. Vol 254**,** 1992, p141.
13. R. Anderson and S. J. Klepeis, Specimen Preparation for Transmission Electron Microscopy of Materials IV, Materials Research Society, Pittsburgh, Vol 480, 1997, p187.
14. S. Subramanian, P. Schani, E.Widener, J. Moss and V. Soorholtz, Proceedings from the 24th International symposium for Testing and Failure Analysis, 1998, p131.
15. M. H. F. Overwijk, F. C. Van den Henvel, and C. W. T. Bulle-Lieuwma, J. Vac. Sci Technol. , Vol B11, 1993, p 2021.
16. L. R. Herlinger, S. Chevacharoenkul, D. C. Erwin, Proceedings from the 22nd International Symposium for Testing and Failure Analysis, 1996, p199.
17. L. A. Giannuzi, L. L. Drown, S. R. Brown, R. B. Irwin, and F. A. Stevie, Mater. Res. Soc. Symp. Proc., Vol 480, 1997, p19.
18. F. A. Stevie, R. B. Irvin, T. L. Shofner, S. R. Brown, J. L. Drown and L. A. Giannuzzi, Charact. And Metrology for ULSI Technology: 1998 Conference, eds Seiler *et al.*, AIP, 1998, p868.
19. R. Rai, S. Subramanian, S. Rose, J. Conner, P. Schani, and J. Moss, Proceedings from the 26th International Symposium for Testing and Failure Analysis, 2000, p415.
20. T. Moore, Private Communication, Texas Instruments, Presented at FEI FIB Users Group Meeting, SEMICON South-West, 1997.
21. T. Ohnishi, H. Koike, T. Ishitani, S. Tomimatsu, K. Umemura and T. Kamino, Proceedings from the 25th International symposium for Testing and Failure Analysis, 1999, p449.
22. M. G. Han, P. Fejes, Q. Xie, S. Bagchi, B. Taylor, J. Conner, M. R. McCartney, IEEE Trans. Electron Dev., 54, 2007, p3336.
23. A. J. Leslie, K. L. Pey, K. S. Sim, M.T. F. Beh, and G. P. Goh, Proc. 21st International symposium for Testing and failure Analysis, 1995, 353.
24. R. Jamison, J. Mardinly, D. Susnitzky, and R. Gronsky. Microscopy and Microanalysis, Vol.6 (Suppl.2), 2000, p526.
25. C. Liu, C. Chen, J. Yan-Chiou, and D. Su, Proceedings from the 28th International Symposium for Testing and Failure Analysis, 2002, p313.
26. T. Ishitani, H. Koike, T. Yaguchi, and T. Kamino, J. Vac. Sci. Technol., Vol (B)16, 1998, p1907.
27. R. M. Langford and A. K. Petford-Long, Proceedings of the 12th EUREM, Brno Czech Republic, July 9-14, 2000.
28. N. I. Kato, N. Miura, and N. Tsutsui, J. Vac. Sci. Technol. Vol. (A)16, 1998, p1127.
29. R Keyse, A Garrett-Reed, P J Goodhew & G W Lorimer, "An Introduction to Scanning Transmission Electron Microscopy", Springer/BIOS Scientific Publishers, 1998
30. K. Li, E. Er and S. Redkar, Proceedings of the 10th International Symposium on the Physical and Failure Analysis of Integrated Circuits, 2003, P. 206.
31. S. Subramanian, R.S. Rai, K. Ly, T. Chrastecky, R. Mulder, K. Harber, Electron Dev. Failure Anal. 10 (2) 2008, p.20.
32. P.M. Voyles, D. A. Mueller, J. L. Grazul, P. H. Citrin, H. –J. L. Gossman, Nature, 416, 2002, p. 826
33. W. Vanderlinde, Proceedings of the International Symposium for Testing and Failure Analysis, 2002, p. 77.
34. E. Coyne, Proceedings of the International Symposium for Testing and Failure Analysis, 2002, p. 93.
35. P. Gnauck, U Zelle, P. Hoffrogge, G. Benner, A Orchowski and W. D. Rau. Proceedings of International Symposium for Testing and Failure analysis, p. 132, 2003.
36. C. C. Ahn and O. L. Krivanek, EELS Atlas, ASU Center for Solid State Science, Tempe, AZ and Gatan Inc., Warrendale, PA, 1983.
37. R. Egerton, Experimental Techniques and Instrumentation, p.29, Transmission Electron Energy Loss Spectrometry in Materials Science, Eds. M. M. Disko, C. C. Ahn, and B. Fultz, TMS, Warrendale, PA,1992.
38. N. J. Zaluzec, Electron Energy Loss Spectroscopy of Advanced Materials, p.241, Transmission Electron Energy Loss Spectrometry in Materials Science, Eds. M.

M. Disko, C. C. Ahn, and B. Fultz, TMS, Warrendale, PA,1992.

39. R. F. Egerton, Electron Energy Loss Spectroscopy in the Electron Microscope, Plenum Press, NY, 1996.
40. H. Heidemeyer, C. Single, F. Zhou, F. E. Prins, D. P. Kern, E. Flies, Self-limiting and pattern dependent oxidation of silicon dots fabricated on silicon-on-insulator material, Journal of Applied Physics, 87(9), p. 4580, 2000.
41. W. -D. Rau, F. H. Baumann, H. H. Vuong, B. Heinemann, W. Hoppner, C. S Rafferty, H. Rucker, P. Schwander and A. Ourmazd, Phys. Review Letters Vol. 82, 1999, p2614
42. W. -D. Rau, P. Schwander, F. H. Baumann, W. Hoppner, and A. Ourmazd, IEDM-98, 1998, p713.
43. W. -D. Rau, P. Schwadner, and A. Ourmazd, Phys. Stat. Sol. Vol. (B) 222, 2000, p213.
44. M. A. Gribdyuk, M. R. McCartney, J. Li, C. S. Murthy, P. Ronsheim, B. Doris, J. S. McMurray, S. Hegde, and D. J. Smith, Phys. Review Letters, Vol. 89, 2002, p25502.
45. U. Muehle, A. Lenk, M. Lehmann and H. Lichte, Proceedings from the 28th International symposium for Testing and Failure Analysis, 2002, p39.
46. K. Vogel, A. Lenk, H. Lichte, H. J. Engelmann, U. Muhle, and B. Breitag, Microscopy and Microanalysis, Vol. 9 (suppl.3), 2003, p240.
47. Y. Y. Wang, A. Domenicucci, M. Kawasaki, J. Bruley, M. Gribelyuk and J. Gaudiello, Microscopy and Microanalysis, P. 564, 2005.
48. R. E. Dunin-Borowski, S. B. Newcomb, T. Kasama, M. R. McCartney, M. Weyland, P. A. Midgley, Ultramicroscopy, 103, 2006, p.67.

.

SELECTED REFERENCES

M. M. Disko, C. C. Ahn and B. Fultz, Transmission Electron Energy Loss Spectrometry in Materials Science, TMS, 1992.

R.F. Egerton, Electron Energy Loss Spectroscopy in the Electron Microscope, Plenum Press, 1989.

J.W. Edington and K.C. Thompson-Russell, Practical Electron Microscopy, Vol. 1-5: Monographs in Material Science, Macmillan Press, 1975.

P.B. Hirsch, A. Howie, R.B. Nicholson, D.W. Pashley, M.J. Whelan, Electron Microscopy of Thin Crystals, Krieger, New York ,1977.

D.C. Joy, A.D. Romig, and J.I. Goldstein, Principles of Analytical Electron Microscopy, Plenum Press, 1986.

M.H. Loretto and R.E. Smallman, Defect Analysis in Electron Microscopy, Halsted Press, 1976.

M.H. Loretto, Electron Beam Analysis of Materials, Chapman and Hall., 1984.

G. Thomas, M.J. Goringe, Transmission Electron Microscopy of Materials, Wiley and Sons, New York, 1979.

D. B Williams and C. Barry Carter, Transmission Electron Microscopy, Vol I-IV, Plenum press, 1996.

R.F. Egerton, Electron Energy Loss Spectroscopy in the Electron Microscope, Plenum Press, 1996.

M. M. Disko, C. C. Ahn, and B. Fultz, Transmission Electron Energy Loss Spectrometry, TMS, 1992.

R. Brydson, Electron Energy Loss Spectroscopy, Taylor & Francis, 2001.

E. Völkl, D. C. Joy, and L. F. Allard (Eds.), Introduction to Electron Holography, Plenum Press, New York, 1999.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

X-ray Imaging Tools for Electronic Device Failure Analysis

Steve Wang
Advanced Photon Source, Argonne National Laboratory, Argonne, IL 60439

1. Introduction

X-ray imaging provides direct visualization of devices' superficial and internal structures, typically with little need of sample preparation or modification. Compared with imaging techniques based on electron and visible light, x-ray technique offers several favorable traits that makes x-rays uniquely well suited for non-destructive evaluation and testing:

1. Large penetration depth of up to many mm with silicon substrate for hard x-ray radiation[1] (Figure 1) to image complete packaged integrated circuits devices and circuit board without destructive de-processing.
2. No charging effect and much lower radiation damage than electrons. Devices remains fully functional when imaging at micrometer resolution[2-4].
3. When combined with computed tomography (CT) technique, produces distortion-free 3D images that represent the devices' structures, including both surface and internal features[5].
4. Non-contact and non-destructive measurements with minimal sample preparation and modification [6].
5. Different material compositions can be distinguished by absorption differences. As shown in Figure 1, heavy metal, light metal, organic and silicon substrate exhibit significantly different absorption density in x-ray images. Furthermore, the samples' exact three-dimensional elemental composition can be mapped using their absorption edges.

These attributes have made x-ray inspection systems a critical non-destructive imaging and analysis tool in the failure analysis laboratories. The majority of commercial x-ray equipment are projection type systems that place the sample in the x-ray beam emitted from micro-focus or nano-focus sources and record the images after the sample. The resolution of these system depends on the x-ray source spot size, magnification, detector pixel size[7-9]. About 1-um resolution is achievable in practical applications. These systems provide high imaging speed so that operator can manipulate samples' position and view angle and make real-time observations. Some manufacturers have also provided integration to CAD design data to allow live overlay on 3D x-ray images for direct comparison and defect analysis[10].

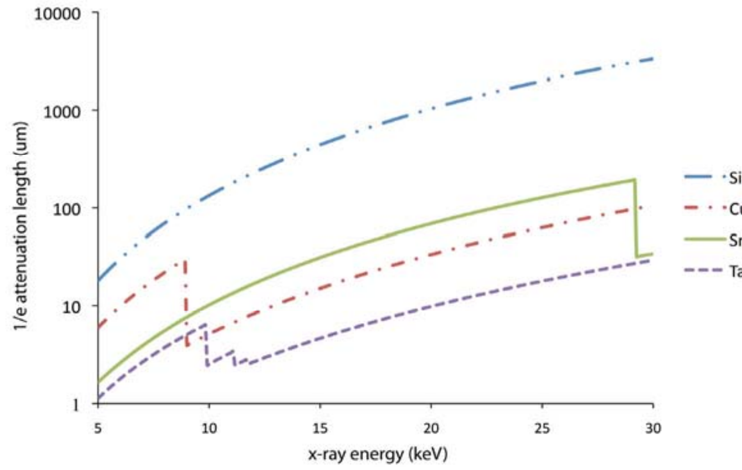


Figure 1. Attenuation length (1/e) of several materials commonly used in electronic devices as a function of x-ray energy.

During the past decade, the increasing complexity and 3D nature of semiconductor packaging structures have prompted x-ray equipment manufacturer to significantly improve the 3D micro-CT imaging capability[11,12]. Most x-ray equipment manufacturer now provide 3D models with fully automated data acquisition, reconstruction, and data analysis packages. In a typical 3D imaging sequence, the operator will locate the region of interest (ROI) and initiate the 3D imaging sequence. The data acquisition will be completed many minutes at low resolution and hours at highest resolution. The user is then presented with the 3D volume data that allow user to make distance and density measurements, observe structures with volume rendering, and extract slices in arbitrary direction to observe "virtual" cross-sectional details.

In order to achieve better than micrometer resolution in practical application, an x-ray optics is needed to magnify the x-ray images. Transmission X-ray Microscope (TXM) using Fresnel zone plate lenses have achieved better than 50 nm resolution using laboratory x-ray sources and better than 30 nm at synchrotron radiation facilities[13,14]. Furthermore, 3D nano-CT techniques using TXM systems have been refined to routine practices. This class of instruments provides over an order of magnitude higher resolution than projections system and is capable of resolving interconnects in modern IC devices. However, their complexity and the difficulty in optics fabrication makes them more costly than projection systems and more time-consuming to maintain. The exposure time is typically in the range of tens of seconds to minutes for each 2D image and hours for each 3D data set.

Figure 2 provides a size scale comparison of features in electronic devices and suitable microscopy techniques. X-ray techniques are particularly well suited for non-destructive failure analysis. The resolution and field of view (FOV) size scale of projection systems makes it a versatile tool for studying a wide range of devices such IC packaging, MEMS, and printed circuit boards. TXM system provides higher resolution, but limited field of view and restricted sample size. They are well suited for studying interconnects, through-silicon vias, and specific components of MEMS devices.

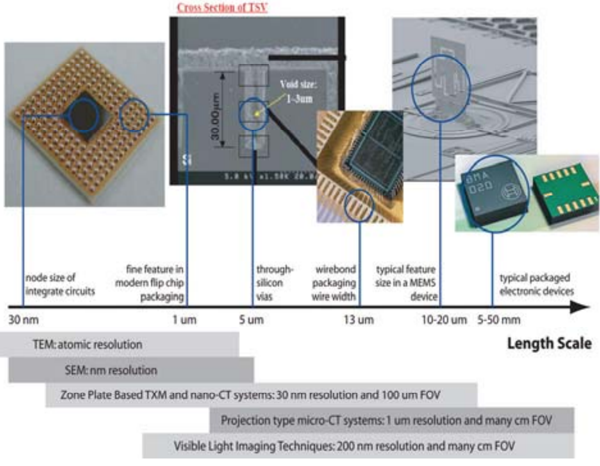


Figure 2. Comparison of size scales of features in electronic devices and imaging techniques.

2. Packaging Failure Analysis with X-ray Instruments

Projection x-ray system are widely used in quality control and failure analysis with IC packaging and integrated circuit boards. In 2D imaging mode, samples can be manipulated to position the region of interest (ROI) in the x-ray imaging field and adjust view angle to best visualize and analyze the defect. Even at high magnification settings, these systems can provide sufficient throughput for real-time imaging. With samples of low complexity, defects can often be identified quickly in 2D imaging

mode if their approximate location is known. Several example 2D images are shown in Figure 3. At a normal view angle, Figure 3(a) shows a die attachment voids on the backside of a semiconductor inside an IC-case. Void size, coverage, and distribution can be determined automatically by manufacturer supplied software. Broken wires can also be observed at higher magnification as shown in Figure 3(b). Figure 3(c) shows an image of through-hole solder joints at an oblique angle. The through-hole plating in the left hole is not properly soldered.

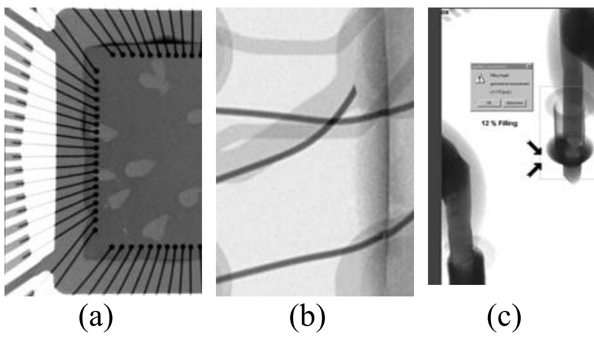


Figure 3. Examples of packaging failure analysis with 2D projection type x-ray imaging system. Images provided by Phoenix |x-ray, a division of GE Sensing and Inspection Technology.

Some x-ray systems also allow operators to import CAD design data and provide a live overlay onto the x-ray image (Figure 4). The overlay is automatically adjusted for sample location and view angle. In addition to design dimensions, component identification such as pad ID is available to the operator at any time. The pad specific inspection results can be accessible by mouse click on the image. The overlay technique provides a convenient way for the operator to pinpoint specific fault location within a device and correlate x-ray analysis other test results.

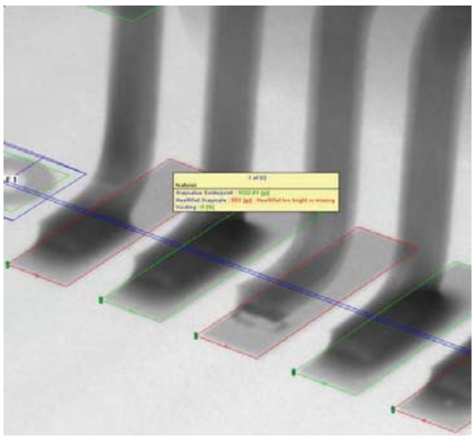
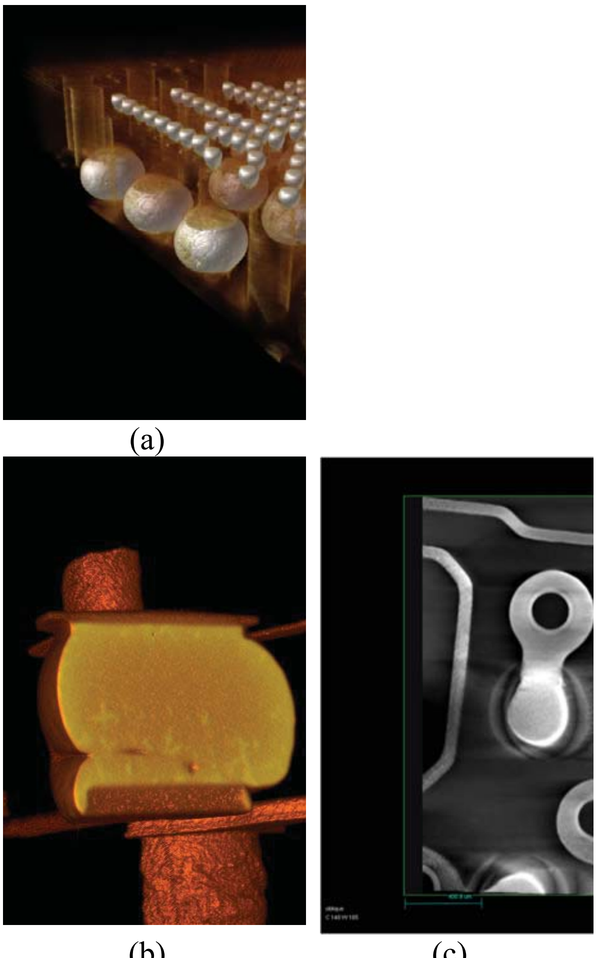


Figure 4. Live overlay of CAD design in the x-ray image. Missing heel fillet and complete missing solder can be identified at two gullwing solder joints. Image provided by Phoenix |x-ray.

With more complex samples, such as multi-level packaging with dense interconnects, 2D images are often contain too many overlapping structures to reliably perform visual inspection and 3D imaging techniques are needed to separate the features in depth. Figure 5(a) shows a volume rendered image of a flip chip packaging. From such 3D data volume, specific regions and slices can be extracted to examine regions of interest in detail. For example, Figure 5(b) shows two neighboring BGA solder joints with cross-sectional images showing the contact with pads. The solder joint at the left is open: The soldering paste melted and wetted the pads, but did not form a junction with the solder balls. The intensity differences in the cross-section areas indicate different metallic phases. Figure 5(c) shows a virtual delayering slice showing a crack occur between solder ball and via. These two defects are nearly impossible to identify from 2D images alone as the defective features are small and are typically obscured by other feature in the packaging when viewed from almost any angle. As the complexity of packaging increases, 3D x-ray micro-CT imaging technique is expected to

play increasingly importatnt roles in the failure analysis laboratory.



(a)

(b) (c)

Figure 5. Figure 5. (a) Volume rendering of a flip chip packaging with voxel size of 7 um. From Xradia, Inc. (b) Volume rendering of two neighboring BGA solder joints with cross-sectional images to show solder contact with pad. From Phoenix| x-ray. (c) A broken trace from crack. From Xradia, Inc.

3. TXM Application in Electronic Devices Failure Analysis

Since its introduction to the semiconductor equipment market in the early 2000s, TXM systems has demonstrated its unique capability to non-destructively study um-scale and sub-um structures in electronic devices. In the die-level failure analysis workflow, the TXM is used directly after electronic functional testing and fault isolation to nondestructively examine structures without physical de-processing. This can in some cases replace the SEM approach with physical cross-sectioning, but can also function as an intermediate step before SEM imaging. The use of TXM saves time, avoids the risk of polishing away the defective layer, and keeps the rest of the surrounding circuitry intact in the event the defect is elsewhere.

Figure 6 shows a bank of fuses are used in a memory redundancy scheme to disconnect defective circuits allowing them to be replaced with functioning replacements. A fuse that is open prematurely disables this critical capability and prevents a fully processed and packaged memory device from exercising its redundancy when and if needed later. When the memory is tested at wafer level, prematurely open fuses show up on failing bit maps when entire blocks of circuitry are nonfunctional. Fault isolation (e.g. laser scanning, probing) is then required to isolate the problem to a smaller area. In this example, Figure 6(a) shows a transmission image acquired at normal angle to the chip. The Cu interconnect line width is 120 nm. A small low density circular feature in the top left fuse indicates material loss resulting as open circuit. Figures 6(b) and (c) shows the same region imaged at different oblique angles. The layered structures of the fuse can already be observed from these two oblique views, and the layer with open can also be identified. The 3D structure of the sample was reconstructed from a series of projects at different views. Slices containing the open layer was extracted from the reconstructed 3D volume data and shown in Figure 6(d) and 6(e) to allow further quantitative analysis.

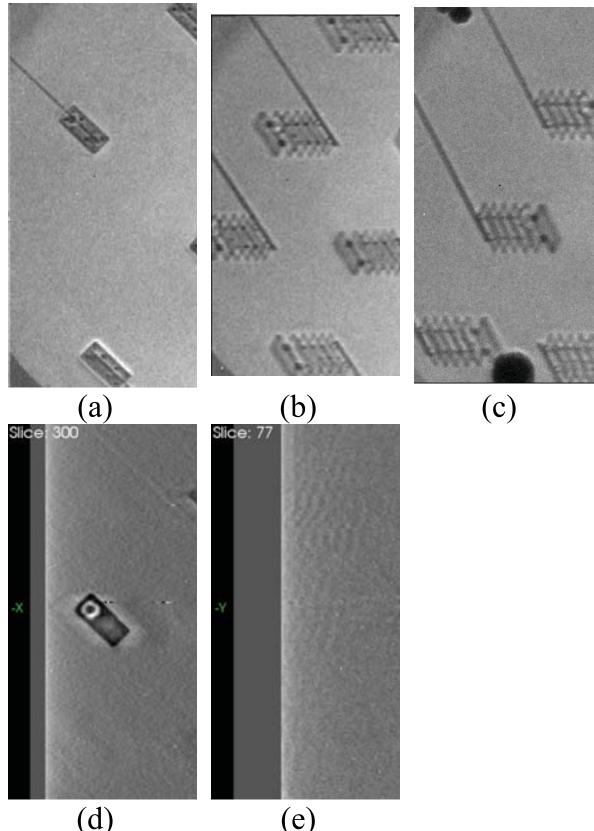


Figure 6. Images of a fuse sample imaged with a TXM and nano-CT system: (a) Image acquired at normal angle, and (b-c) at oblique angles. (d-e) Slices from the reconstructed 3D volume containing the defect slice. Image from Xradia, Inc. Sample provided by IBM.

This TXM's high-resolution non-destructive 3D imaging and analysis capability has become particularly important with the development of 3D IC. For example, Figure 7(a) shows series of through-silicon vias (TSV) acquired with the nanoXCT system at Xradia, Inc. The 3D volumetric data was imaged with a die sample that has been thinned to 100 um. The vias were designed to have a 5-um width and 20 um depth. These features are very difficult to image directly and is typically studied with SEM by cross-sectioning. With x-ray nano-CT technique, however, the complete 3D structures can be studied non-destructively and the geometry of the vias can be measured directly from the volume data as shown in Figure 7(b) and 7(c). These images were acquired with 50 nm diffraction-limited resolution and 20 nm pixel size. The measurement accuracy is better than 50 nm. In this sample, a slight taper was measured with the width of 4.9 um at top and 4.6 um at bottom. Furthermore, a small pyramid shaped fabrication artifact was visible at the bottom of the via. This feature was not previous identified with other imaging techniques.

In addition to failure analysis, x-ray nano-CT system has also being used in competitive analysis and integrated circuit verification. These applications are becoming increasingly important as more integrated circuits devices are being fabricated by foundries. As an example, DARPA is currently engaged in the Trust in Integrated Circuits program to investigate the use of TXM to verify devices designed in the US and fabricated overseas[15]. As the part the program, the TXM is used to map the entire IC sample's 3D structure and compare it with the design to identify possible differences.

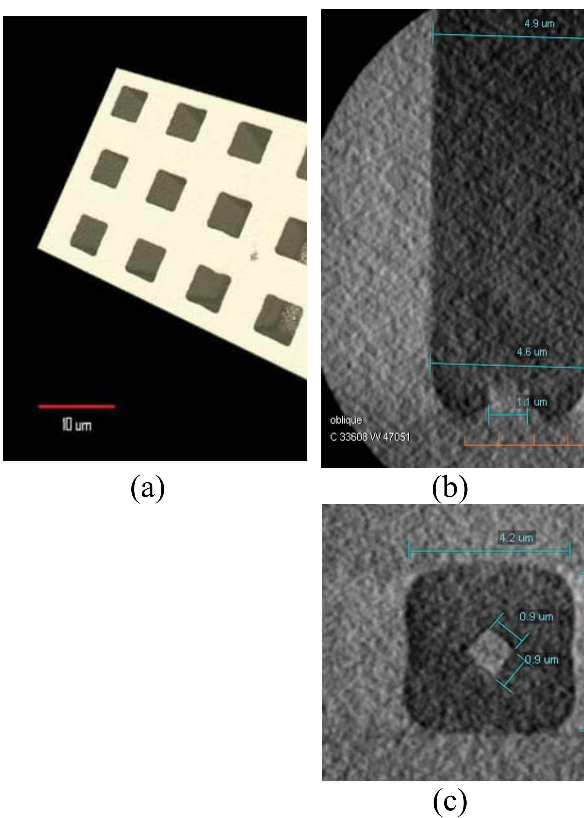


Figure 7. (a) Volume rendering of a TSV sample obtained with a TXM system. (b-c) Via geometry measurements made from the 3D data. Image from Xradia, Inc.

4. Summary

X-ray imaging systems play a critical role in failure analysis laboratories. The non-destructive nature of x-ray technique avoids sample destruction and helps users save time, reduce cost, and diminish the risk processing errors. To further improve tool productivity, equipment manufacturers have developed innovative features such as live CAD-overlay to provide multiple sources of information that help users understand samples structure and defect source. The trend of increasing complexity, higher density, and finer feature size of packaging and MEMS devices is placing stronger demand on high-resolution 3D imaging capability. Commercial micro-CT systems provide highly automated 3D imaging and analysis capabilities at micron level resolution that can help users identify and investigate nearly all common defect types. Furthermore, nano-CT tools based on TXM systems are beginning to play a role in failure analysis with sub-micron size structures. In both types of x-ray imaging systems, the virtual delayer technique use 3D image data has in many cases eliminated the need of destructive physical cross-sectioning. With improvement resolution, throughput, and 3D analytical features, as well as potential reduction in cost, the non-destructive imaging capabilities of x-ray tools have the potential to significantly improve the practice of failure analysis in the future.

5. References

[1] B.L. Henke, E.M. Gullikson, and J.C. Davis. X-ray interactions: photoabsorption, scattering, transmission, and reflection at E=50-30000 eV, Z=1-92, Atomic Data and Nuclear Data Tables Vol. 54 (no.2), 181-342 (July 1993).

[2] Colangelo, J., Microelectronic Failure Analysis Desk Reference, Fourth Edition, ASM International, 2002.

[3] Effects of x-ray irradiation on the channel hot-carrier reliability of thin-oxide re-channel MOSFETs," J. Y. C. Sunet al., Conf. Solid State Devices and Mater. (Japan), 1986, pp. 479.

[4] Comparison of effects of ionizing radiation and high-current stress on characteristics of self-aligned bi-polar transistors, T. C. Chen, J. P. Norum and T. H. Ning, Proc. 20th Conf. on Solid-state Devices and Mater., Tokyo, 1988, pp. 523.

[5] Haddad, W.S. and J.E. Trebes. Developments in Limited Data Image Reconstruction Techniques for Ultrahigh-Resolution X-ray Tomographic Imaging of Microchips. in Developments in X-ray Tomography. 1997. San Diego, CA: SPIE.

[6] Wang, Y., et al., A transmission x-ray microscope (TXM) for non-destructive 3D imaging of integrated circuits at sub-100 nm resolution, Conference Proceeding from International 29th Symposium for Testing and Failure Analysis, 29 227-233, 2002.
[7] Bonse, U. (Editor.), "Developments in X-Ray Tomography IV" SPIE, Wellingham, (2004).
[8] Brockdorf, K. et al., "Sub-micron CT: visualization of internal structures" in Developments in X-ray Tomography VI, edited by Stuart Stock, Proceedings of SPIE, Vol. 7078, (2008).
[9] Wang, Y., X-ray Microtomography Tools for Advanced IC Packaging Failure Analysis, Microelectronic Failure Analysis Desk Reference, pp 261, Fifth Edition, ASM International, 2004.
[10] Neubrand, T., Can AXI Meet Zero Defect Quality Standards?, Surface Mount Technology, Oct 26, 2009.
[11] Brunke, O., High-resolution CT-based defect analysis and dimensional measurement, Insight Vol 52 No 2 February 2010.
[12] M Feser et al, Meas. Sci. Technol. 19 (2008) 094001.
[13] Mau-Tsu Tang, et. al., Hard X-ray Microscopy with Sub-30 nm Spatial Resolution at NSRRC, Proc. 8th Int. Conf. X-ray Microscopy, IPAP Conf. Series 7 pp.15-17.
[14] Lau, S. H., Non Destructive Failure Analysis Technique With a Laboratory Based 3D X-ray Nanotomography System, LSI Testing Symposium 2006, Osaka, Japan.
[15] Collins, D., TRUST, A Proposed Plan for Trusted Integrated Circuits, DARPA report.

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Atomic Force Microscopy: Modes and Analytical Techniques with Scanning Probe Microscopy

J. Colvin
Jim Colvin Consulting Services, Newark, CA, USA

K. Jarausch
Intel Corporation, Santa Clara, CA, USA

Abstract

The Scanning Probe Microscope (SPM) has matured rapidly and is already a valuable tool for Failure Analysis and indispensable for FAB process control. The SPM or more commonly known as the Atomic Force Microscope (AFM) can do much more than just image a surface as will be shown in each section. The theory of operation and various modes will be discussed and routine failure analysis presented.

Introduction
There are many other types of scanning and imaging tools but what sets the SPM apart from other scan methods is that a tip is physically rastered across the sample while force or some other sample interaction is recorded. Scanning probe microscopes have been commercially available for just over 16 years now. In this time they have matured quite rapidly compared to other analytical instruments, such as scanning electron microscopes, surface analysis tools (Auger electron spectrometers, secondary ion mass spectrometers, etc.), and other frequently used instruments which have developed over the last 40 years. Scanning probe microscopes have not only improved the original scanning technologies (scanning tunneling microscopy and contact atomic force microscopy) but they have broadened their utility through the development of many new scanning technologies. In this article we will briefly review all of the scanning technologies which are currently helpful in IC failure analysis.

During this first approximately 16 years, the applications of SPM's have also grown rapidly. Initially the applications were in fairly basic research areas while scientists began to become familiar with these new tools. Because, at least in part, of the ease of sample preparation and use of the SPM's the tool quickly found uses in more applied research and even production areas. They found uses in production process control in the data storage industry within the first five years of their availability. They were in use in analytical laboratories of the major semiconductor manufacturers within the first six to seven years of their availability. By now SPM's are in routine use in production control applications in data storage, semiconductor, polymer, contact lens, and other manufacturing facilities. These applications all focus on the measurement of some property, associated with a surface, at high spatial resolution.

Initially, the use of SPM's in IC failure analysis (ICFA) had not progressed very rapidly. The familiarity of ICFA personnel to SPM's had not been as high as it is, for example, to SEM's. This has changed with the development of new technologies such as scanning capacitance microscopy, Conductive AFM (C-AFM) also known as Tunneling AFM or (TUNA), Scanning Spreading Resistance (SSRM) and Atomic Force Probe (AFP) methods. Sample preparation has also evolved allowing FIB assisted SCM measurements, for example.

The SPM has improved dramatically over the original contact mode with the development of TappingMode AFM (AC mode scanning) in 1993. An offshoot from the availability of AC mode is phase contrast measurements in AFM. Lift Mode, introduced in 1994, enabled the separation of topography from magnetic or electric field measurements to make those measurements practical. Detailed surface electric potential measurements are also possible making high spatial resolution Kelvin probe maps of surface work functions and other potentials. High spatial resolution surface temperature measurements are available through scanning thermal microscopy (SThM). Optical probes or optical signal pickup can now be accomplished at dimensions below the traditional "diffraction limit" through the use of near field scanning optical microscopy (NSOM). Carriers in semiconductors can be mapped out showing carrier type, concentration, junction locations, and even very local C-V measurements using scanning capacitance microscopy (SCM).

The SPM and its modes

Contact mode (DC)
A sharp tip mounted on the end of a long flexible cantilever is pressed against the sample surface with a known force (known from the curvature of the cantilever) and dragged across the surface while keeping the force constant. The vertical movements required to maintain the force constant are the measurements of the surface topography. DC AFM produces a shear force from the dragging of the tip against the surface. Figure 1 shows the basic layout of an SPM for both DC and AC operation. The tip or sample can be scanned depending on the

configuration of the SPM with motion provided by piezoelectric transducers for X, Y and Z. The cantilever/tip combination is held at a constant height during the scan by monitoring a laser bounced off the top of the cantilever and changing Z to compensate. A topographical image is generated based on the change in Z to maintain a constant force on the tip for each X-Y data point. [1]

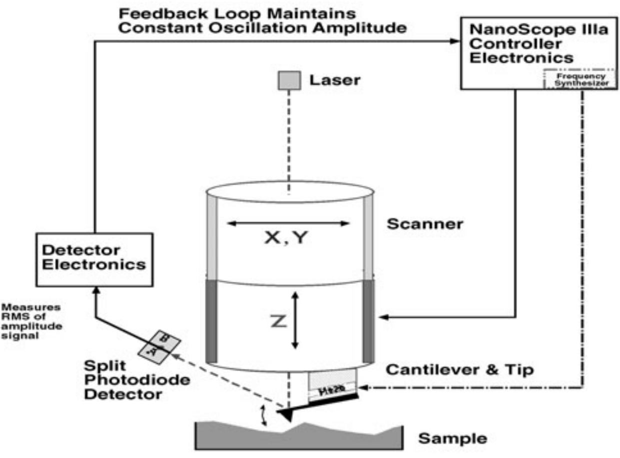


Figure 1: Schematic diagram of TappingMode AFM. Courtesy Digital Instruments.

Lateral Force
Lateral force operates the same as DC with an additional split on the photodetector to monitor the cantilever torsion or twist as a result of dragging the tip across the sample. The frictional force of the tip/sample can be mapped allowing some material characterization. [2]

Tapping mode (AC)
Tapping mode nearly eliminates the tip/sample damage issues present in DC mode. For this version of the technique the tip/cantilever assembly is attached to the piezoelectric scanner with the sample surface stationary. [3] In this case the cantilever is oscillated so that the tip oscillates at an amplitude in the range from 5 to 100 nm before it is brought down to where it “taps” against the sample surface. When it is brought down and linked to the surface, the oscillation amplitude reduces due to the tapping against the surface and the percentage reduction in the amplitude determines the force of the tip on the surface set by the operator and controlled by the feedback loop. The tip is then scanned across the surface while tapping against it and while maintaining a constant oscillation amplitude. The vertical motion required of the z piezoelectric crystal to maintain the constant oscillation amplitude is a precise measurement of the surface topography to .05 nm in all three axes. Compare this to the SEM, which is only capable of X-Y measurements since Z is relative topography only. Figures 2-7. [4,5]

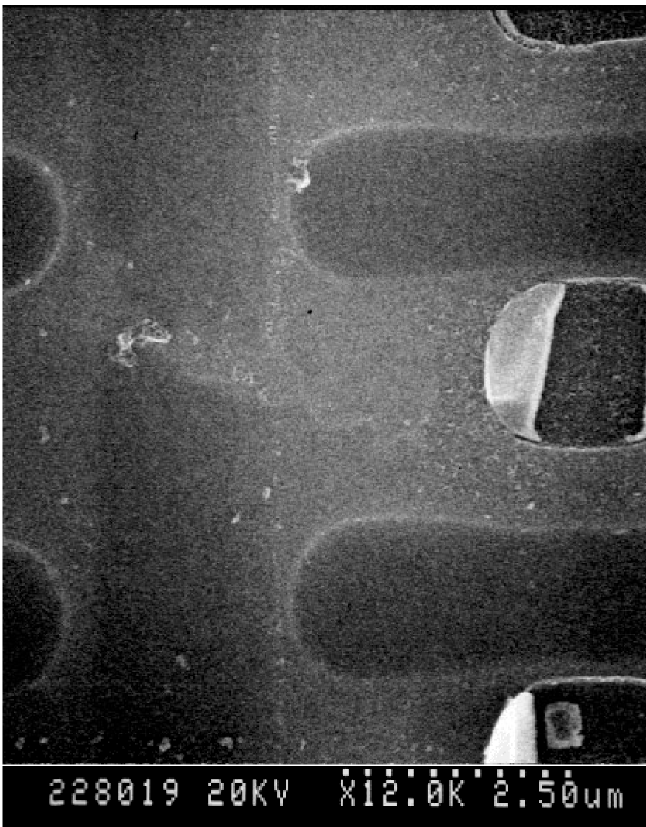


Figure 2: SEM image of resist residue at substrate affecting an implant. Note the poor Z resolution.

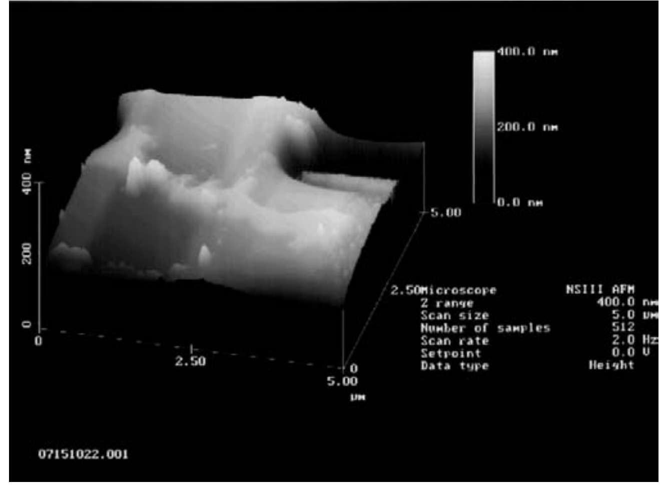


Figure 3: AFM image of the same residue on the substrate.

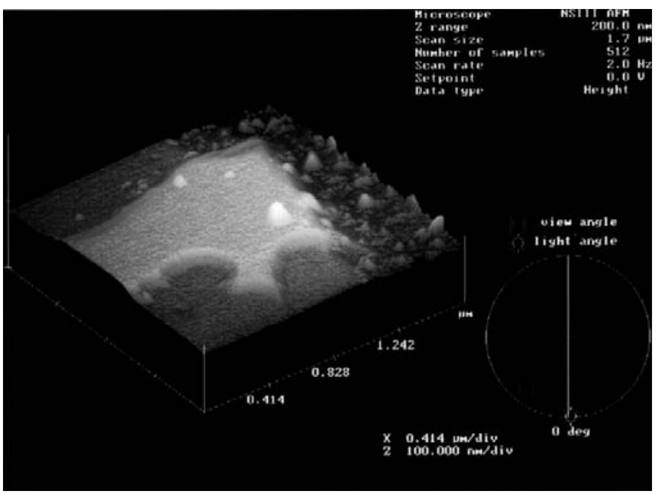


Figure 4: Increased magnification view of the residue clearly showing topographical details unattainable in the SEM.

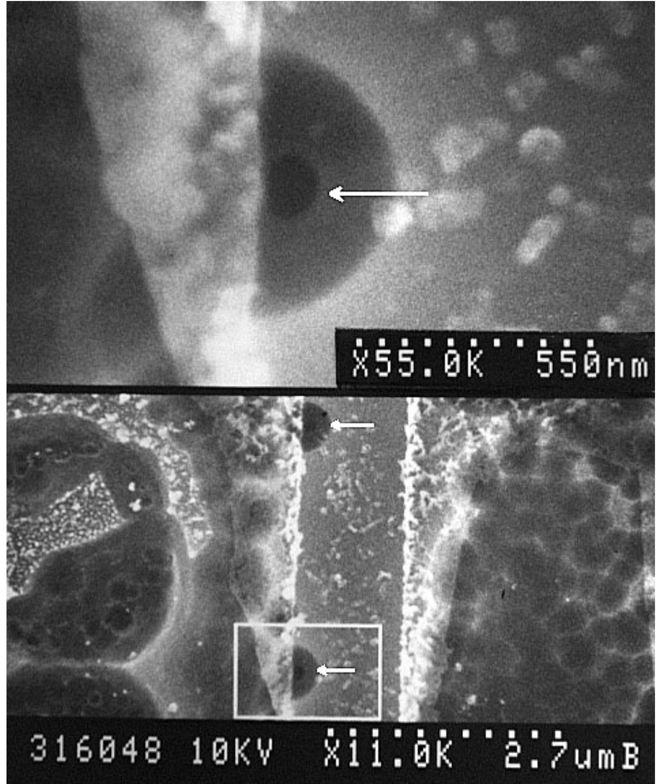


Figure 5: SEM split view of the LESD damage sites. Note the hole within a hole. The outside dark ring is due to plasma undercut that removed the underlying channel radially from the rupture. The size of the rupture ranges from .15 µm to .26 µm in the Y direction and from .12 µm to .19 µm in the X direction for various samples.

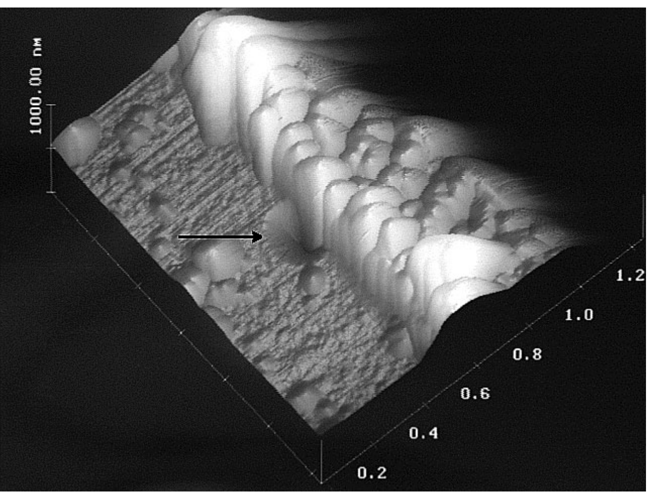


Figure 6: Atomic Force Microscope image of an LESD damage site. Compare the appearance of the oxide sidewall to the SEM image in Figure 5.

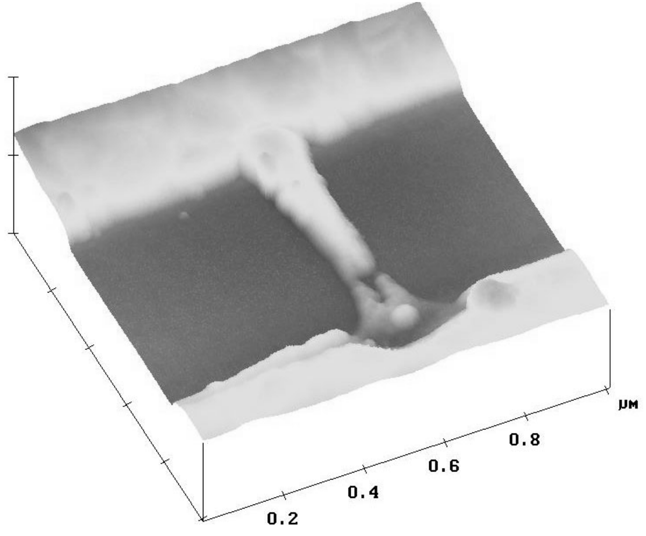


Figure 7: Backside AC SPM image of lateral ESD damage (melt filament) between poly. Deprocessed with Choline Hydroxide.

Phase contrast

Phase contrast is an offshoot of the oscillating cantilever. [6] The cantilever is driven in oscillation by an AC signal generator and a piezoelectric crystal. At the same time, the motion of the cantilever is monitored by the photodetector, which monitors the oscillation amplitude. We can compare the drive signal and the motion response signal and measure the phase shift between the two. When the cantilever bounces elastically off the surface in its tapping, there is a set phase shift. When the tip is pulled toward the surface by any of a variety of forces this extra pull required to get away from the surface causes an additional phase shift. Using the phase shift to produce contrast in an image shows a map of those locations on the sample where the tip/sample interaction is stronger than at others. An example of the use of phase imaging is shown in Figure 8, which is a topography image (left) and phase image

(right) taken simultaneously on the same area of a bond pad. This pad had been coated with a polyimide layer, which was to have been etched off. This bond pad was on a wafer, which showed poor bond pad adhesion. The phase image shows small spots of high contrast, which were later shown to be residual polyimide, which caused the poor bonding. This quick and easy measurement showed high sensitivity for the detection of the polyimide.

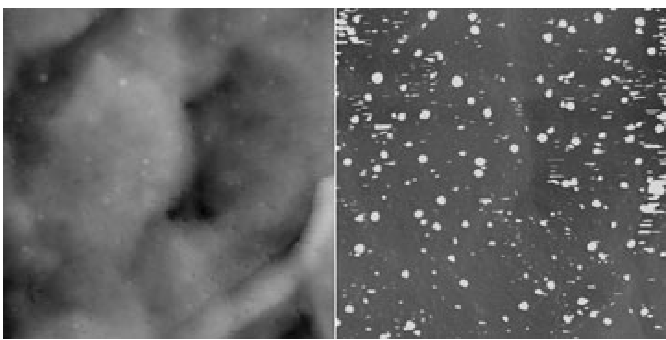


Figure 8: Topography (left) and phase contrast (right) images of an Al bond pad, which was on a wafer producing poor bond adhesion. The topography shows typical Al grains while the phase contrast shows polyimide residue.

Nanoindentation

Nanoindentation provides nanohardness data by forcing a diamond tip down on the sample in select locations with stepped force. Scanning over the area after indentation yields data about the compressibility of the material. Note the difference in the extruded (displaced) material in the right image for these two different materials. [7,8]

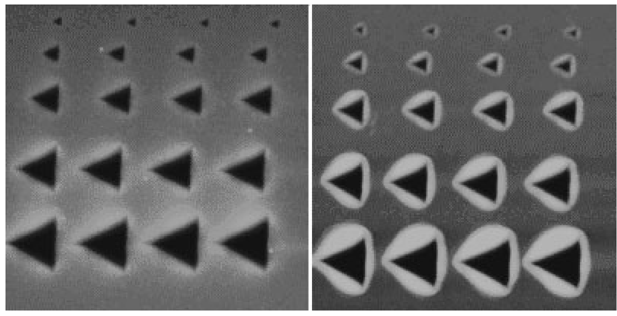


Figure 9: Nanoindentation of 2 differing materials.

Lift Mode

Lift mode enables separate but concurrent measurements of surface topography and a field associated with the surface, through the use of interleaved scans. The principle is illustrated in Figure 10. It works by making one scan line of the measurement area (left to right and then back, right to left) in AC Mode to measure the topography of the surface. Next the probe is lifted some height above the surface (typically 5 to 100 nm, chosen by the operator) and follows the same height pattern as it makes the next scan line. On this scan line it is lifted above the surface, oscillating, and sensing the field at this constant distance above the surface. This provides a field measurement at a constant height above the surface, without any interference from the tip striking the surface and thus gives a clean measurement of the electric or magnetic field described below. In this way both a topographical image and a field image are built up concurrently.

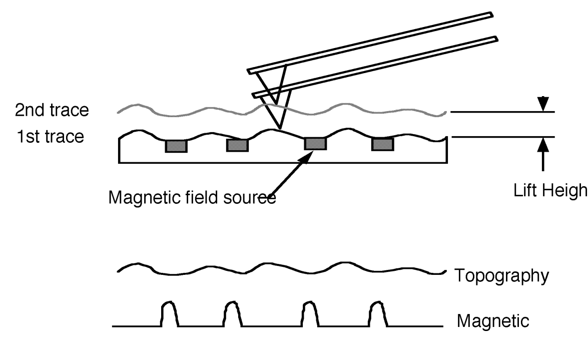


Figure 10: Schematic diagram of Lift Mode operation.

Magnetic field

Magnetic field measurements require a tip which is either a small piece of magnetic material or a regular Si tip coated with a magnetic film. With the tip oscillating while it scans across and above the surface the vertical component of the magnetic field can be sensed by the force it exerts on the magnetic tip because this alters the resonant frequency and phase of the oscillating cantilever. The shift in either the phase or the frequency can be used to produce contrast in a map of the magnetic field associated with the sample surface.

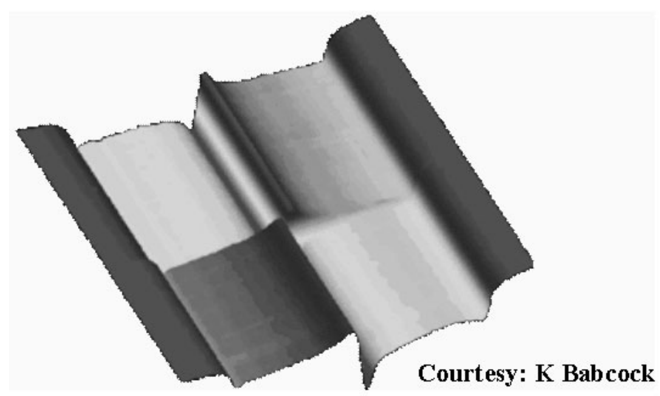


Figure 11: Fringing field due to current in a magnetic recording head write element. The current was reversed mid-scan to show the contrast.

Magnetic force microscopy is used to map currents in IC's with a scanning probe. Sensitivities of around 1 mA DC and around 1 μA AC are attainable while comparable E-beam current mapping techniques are typically limited to greater than 100 mA of ac resolution. A sufficient magnetic dipole moment is required at the tip in order to image the weak H field that is on the order of a few microteslas, whereas a coated tip for magnetic imaging has a weak magnetic domain and is suited to image intense magnetic domains such as recording media. [9]

Electric field

Electric field measurements can be made in a similar way except that the tip only needs to be coated with a conductor, thus producing a capacitor with the tip as one plate and the region of the sample on which some charge is applied as the other plate. The tip scanning and oscillating above the surface again feels a force, which is dependent upon the vertical component of the electric field gradient at each pixel of the scan line. This causes a shift in the phase and frequency of the cantilever's oscillation, either of which can be used to produce contrast in a map of the field above the surface. As an example of an EFM image Figure 12 shows an area of an operating device, with passivation intact. The right image is an AC SPM image of the surface showing nothing out of place. The left image is an EFM image showing the potential distribution across this area and showing a transistor which is running in saturation because of a gate oxide short at another location on the device. The variations in intensity of the EFM signal are due in combination to variations in spacing between the tip and potential on the buried conductors. [10]

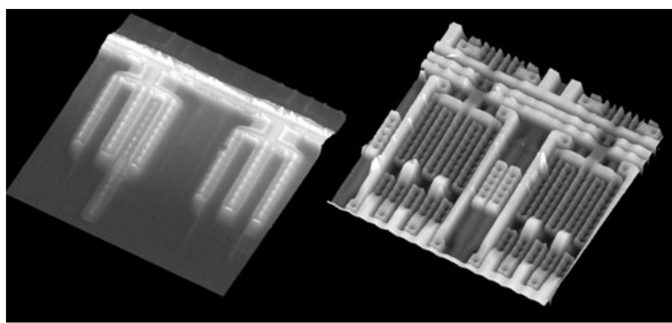


Figure 12: Electric force microscopy (left) and topography (right) images of an operating device. The potential distribution shown in the EFM image shows a transistor operating in saturation. This voltage distribution measurement was made through the passivation coating.

Scanning thermal microscopy (SThM)

Scanning thermal microscopy requires a tip that is constructed so as to measure the temperature or thermal conductivity of the surface, at high spatial resolution. As thermal energy diffuses to a surface on which SThM can measure it, it also diffuses laterally. A point source of heat buried deep under the surface being measured shows considerable lateral extent in its signature on the surface. An example of this is seen in Figure 13, which was taken on an operating device that had a gate oxide short. In the topography image, on the left, taken on top of the passivation, there is no indication of any problem. In the SThM image, on the right, the whiter area is the hot area. This shows where the gate oxide short is located but the heat has diffused outward to make a large area hot spot. In this instance the heat source is 2 to 3 microns below the surface which is being mapped. Temperature resolution to 5 mK and spatial resolution to 30 nm have been reported. [11,12]

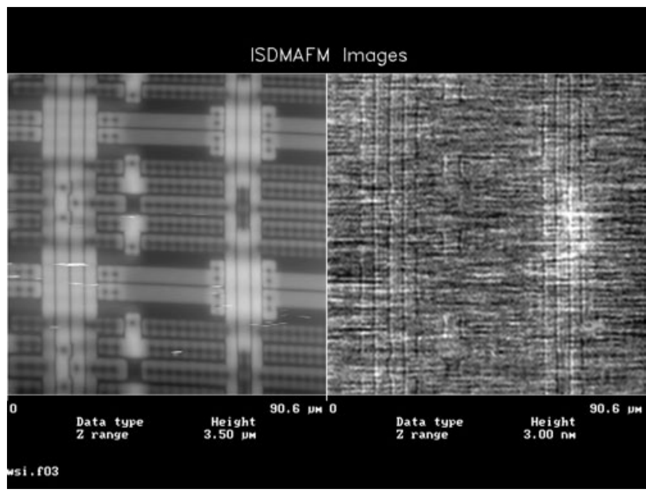


Figure 13: Topography (left) and temperature distribution (right) on an operating device. The temperature image points out the location of a gate oxide short, below 2 to 3 microns of material.

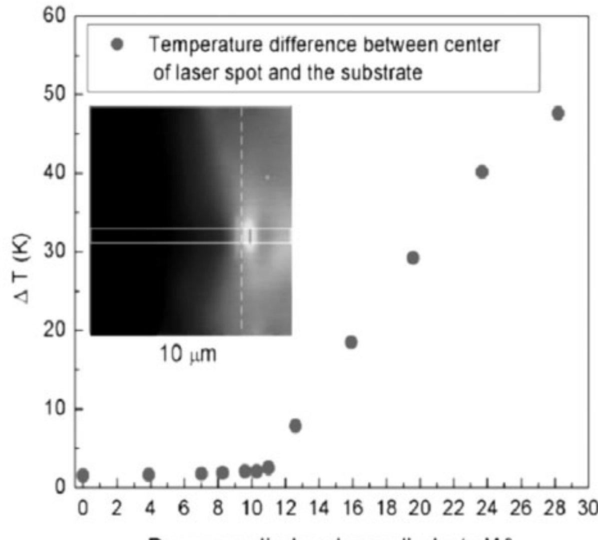


Figure 14: Temp profile of a laser diode measured by SThM. Courtesy Digital Instruments.

Implant imaging

Implant imaging techniques based on scanning probe microscopy (SPM) can be applied to diagnose failures. [13, 14] Qualitative images are often sufficient for a failure analysis diagnosis since images of a failing location can be compared to a passing location to identify differences. Dopant profiles can obtained indirectly, by first decorating with a selective etch and then imaging topography, or directly by imaging local variations in capacitance or spreading resistance. The wet chemical decorations used for SEM and TEM (Dash, Wright's etch…) will highlight/pit areas with of high implant concentration (>10^{17} carriers/cm^3) and this pitting can then be imaged with an AFM. Scanning capacitance microscopy (SCM) and scanning spreading resistance microscopy (SSRM)

are able to resolve implant distributions over a greater dynamic range. [15,16]

Scanning Capacitance Microscopy

Scanning capacitance microscopy measures capacitance variations between a conductive tip and a semiconductor sample while scanning in contact mode. [15] This technique produces a 2-D image of near surface variations in carrier concentration with a sensitivity range of 10^{15} to 10^{20} carriers/cm^3 and a lateral resolution of 10 to 20 nm. An AC bias is applied to the sample and the capacitive coupling of the metallic or conductive probe is recorded point by point as shown in Figure 15. SCM data provides polarity of dopant information and relative concentration data but is not easily quantifiable due to variance in tip to sample interaction and sample preparation. The relatively low imaging forces of contact mode (DC) allow for repeated SCM imaging of a samples surface.

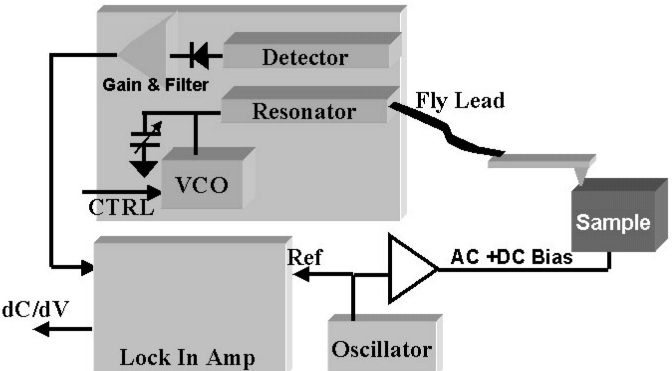


Figure 15: Schematic of SCM system.

SCM requires a sample's surface to be free of damage, trapped charge and ionic contamination. [15] Damage and trapped charge at the surface effectively passivates the capacitive coupling between the SPM probe and the dopants of interest. To prepare low damage plan view samples, selective wet chemical etches have proven useful. For example, during deprocessing, HF can be used to remove the final layer of oxide without damaging the underlying Si. One such plan-view SCM example is shown in Figure 16. [17] Cross-sections with low-damage surfaces can be produced reliably using traditional polish techniques. To enable SCM on both plan view and cross-section samples, a thin oxide free of additional trapped charge must be grown on the surface to provide a uniform dielectric. Placing samples on a hotplate (at 300 C) for twenty minutes under UV illumination has proven effective. A damage free surface, covered by a uniform dielectric, and good ohmic contact to the sample's substrate, are all requirements for successful SCM imaging.

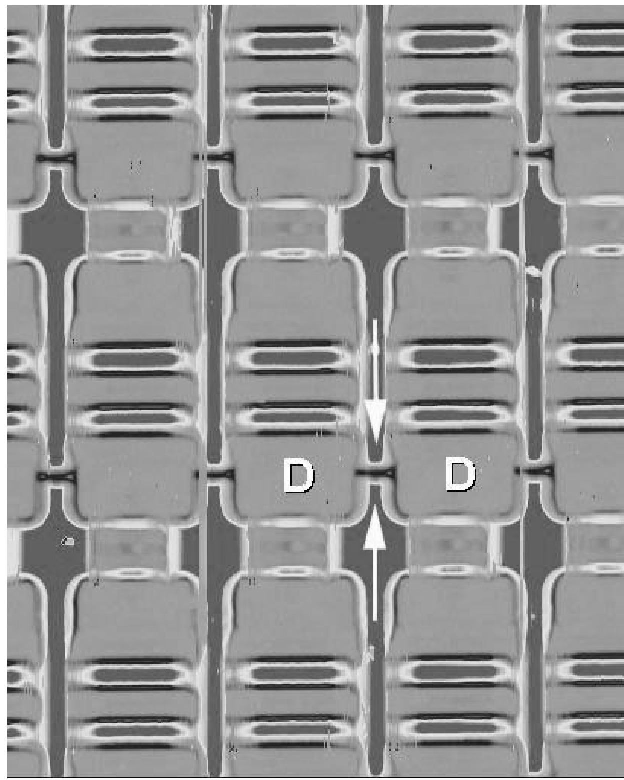


Figure 16: Plan-view SCM image of non-volatile embedded memory array used to diagnose drain leakage failure. The arrows illustrate leakage paths across every other row related to a "bullseye" yield pattern on the wafer. Note that the light blue region indicates n+ type and dark blue indicates n– type dopants. The red regions are the p– areas.

To obtain site-specific implant profiles the final steps of conventional polish sample preparation can applied to FIB generated cross-sections. [18] Figure 17 depicts the integrated FIB and polish sample preparation flow, which can be used to target single bit failures. The dual-beam is used to generate a site-specific cross-section and the final polish is needed to remove the ion damage layer. From start to finish this sample conversion procedure (from a finished dual-beam cross-section to an SCM-ready sample) requires roughly 4 hours per sample.

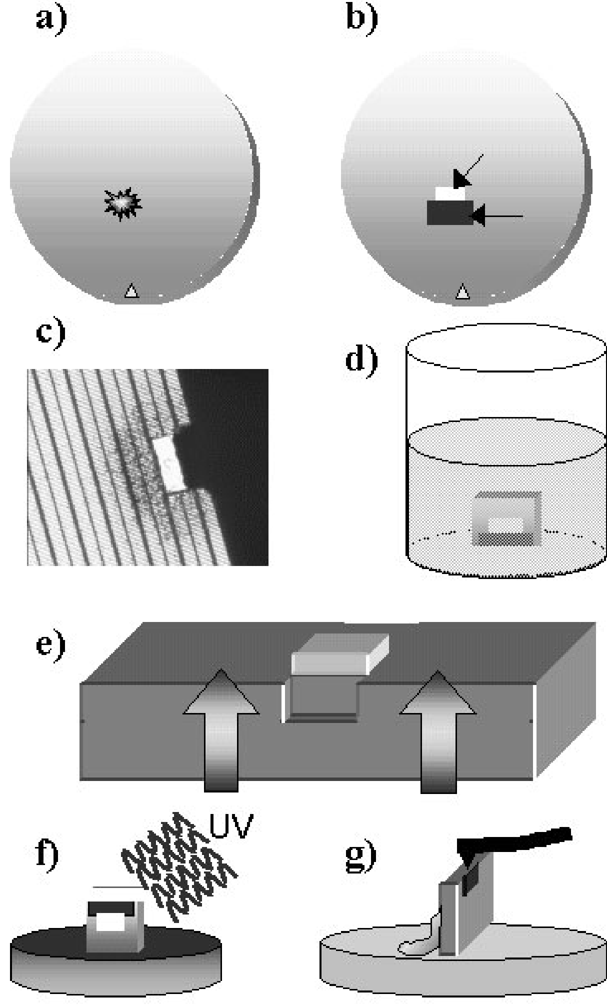


Figure 17. Schematic illustration of FIB assisted SCM.

(a) Failing location is identified using electrical testing and/or fault isolation techniques.
(b) Failing location is cross-sectioned using standard dual-beam techniques.
(c) Optical micrograph showing how the FIB trench is cleaved or polished so that the cross-section face is within 5 um of the cleave surface.
(d) W plugs can be etched away on the cross-section face using 1:3 $NH_4OH:H_20_2$ to improve the imaging (this step is optional).
(e) Cross-section face is polished for 25 seconds with 0.02 um colloidal silica slurry on a felt pad, with FIB cross-section face on the trailing edge. Arrows indicate the direction of the polish wheel rotation.
(f) Sample is heated at 300 C under UV illumination for 20 minutes to produce a 2-3 nm thick oxide.
(g) Sample is mounted on SPM sample stub using hot glue and silver paint which provides a conductive path to the backside of the wafer.

The distribution of implants at a failing location can provide both root cause and information about when the processing defect occurred. An SEM image of a silicon particle causing a single-bit contact to gate short in a 0.l3 um flash memory cell is shown in Figure 18(a). From this image it is apparent that the defect was present after the gates were patterned and before the spacer was deposited. However there are still several dozen processing steps to consider between these two operations.

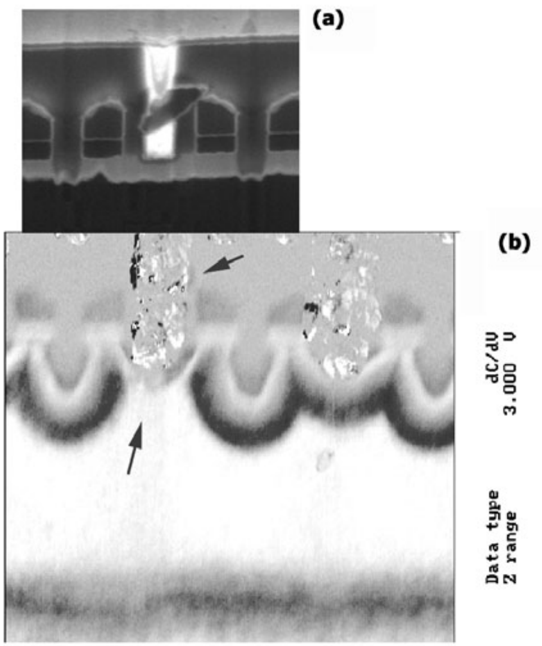


Figure 18: a) SEM image of a failing flash cell prepared by FIB milling. b) SCM image of the same failing site where n-type is dark and p-type is white. The SCM image clearly shows that the particle blocked the drain implant. This indicates the defect was present prior to the drain implant step.

The SCM image of the same failing location is shown in Figure 18(b). The failing bit can be compared to the neighboring good bit. The obvious difference in the drain implant clearly indicates that this particle was present prior to this implant. A similar image was obtained for another particle positioned above a source. Together these SCM images of single-bit failures were used to narrow the search for these defects from 100 down to less than 8 process steps. This example illustrates how site-specific SCM can be used to isolate the source of single bit failures.

Scanning spreading resistance microscopy (SSRM)

SSRM and SCM are complimentary techniques. SSRM gives quantitative data based on local resistance [16] with dynamic range and spatial resolution similar to SCM. In SSRM a bias is applied to the sample while diamond or metallic probe is scanned across the surface and the resistance is recorded point by point (Figure 19). For successful SSRM imaging, the probe must remain in ohmic contact with the surface during the scanning. For SSRM of Si surfaces, this requires high force

scans (>10^{10}N/m^2 or 1,000kg/mm^2) to ensure that the Si at the point of contact undergoes the phase change to metallic Si. [16] Sample preparation requirements are less stringent than for SCM since the technique is less sensitive to surface contamination, damage or dielectrics. Unfortunately the high scan forces required by SSRM do generally cause damage and therefore samples cannot be scanned repeatedly. Although dopant polarity information is not obtained with SSRM, good quantitative data can be obtained of dopant concentrations. Figure 20 illustrates the ability of SSRM to resolve dopants over a wide concentration range.

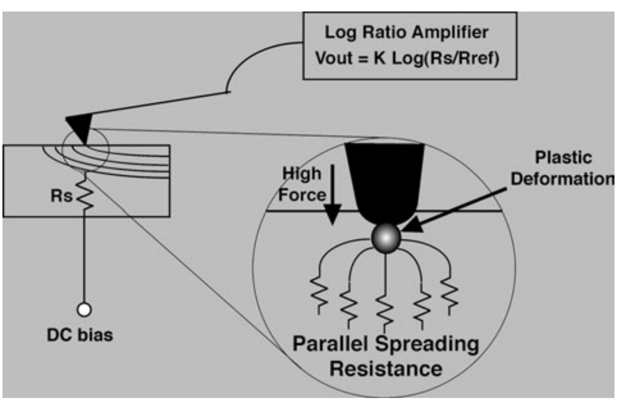
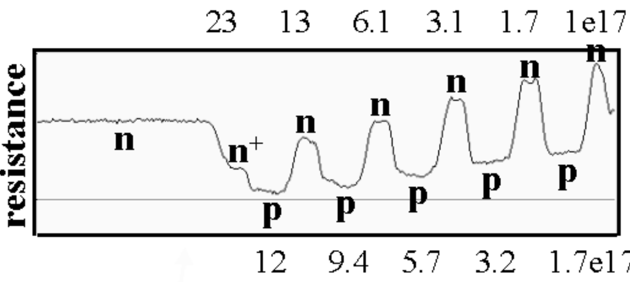


Figure 19: Schematic representation of SSRM



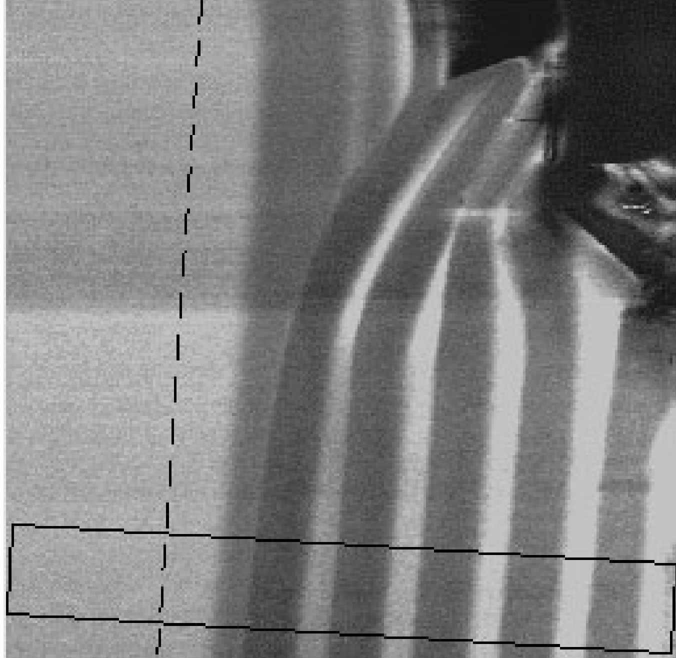


Figure 20: SSRM on InP Test Structure shows dopant concentrations.

Tunneling mode

Tunneling mode AFM also known as Conductive AFM (C-AFM) produces a topography map and a map of the voltage required to tunnel at a set current level. [19] The SPM feedback loop maintains a constant force on the tip while the tunneling current feedback loop adjusts bias voltage to maintain a constant tunneling current between tip and substrate. The current is typically set at around 1 pA for tunneling measurements. Do not confuse this with the early Scanning tunneling microscope. Key differences are that this technique does not normally use the tunneling current for feedback control due to instabilities of oxides and operates in either contact or intermittent contact modes to generate the topographical image. Typical tunneling fields are reached at 10 V/nm of oxide thickness. Although this method was originally recognized as an inline measurement for the FAB, this technique has merit in FA for tunnel oxide/oxide failures, dielectric integrity measurements, passive voltage contrast capabilities, and nodal I-V curves with single or multiple AFM probes. Figure 21 is a schematic diagram of a typical conductive AFM configuration. The ability of C-AFM to image very low level currents ranging from 30fA to 100 pA with 5-10 nm spatial resolution, makes this mode very useful for semiconductor analysis. Dielectric integrity measurements and

the ability to map current paths on IC's are possible since this mode provides 2D imaging and local I-V spectra. The dynamic range of the amplifier is large, allowing the 60fA to 120pA imaging bandwidth. Figure 22 illustrates gate oxide thinning at the "birds beak." This data correlated well with TEM physical measurements of the product.

For analysis of post fabricated devices, delayering techniques are used. Figure 23 shows the desired target. Surface lapping techniques are used (Colloidal silica) to polish through the layers until the poly cap of the flash cell is exposed. This endpointing can be done optically based on the color of the surrounding remaining oxides. Since the device will lap nonuniformly, the SPM can be used to image in several areas of the array and correlate the color viewed in brightfield microscope to the SPM (or SEM) topography. Choline Hydroxide or TMAH is then used to etch the remaining poly gate, exposing the underlying gate oxide as shown in Figure 24. Figure 25 is the resulting 2D tunneling data taken from the prepared sample of a normal gate structure. It is important to pay attention to bias voltage and polarity (protect the sample from anodic oxidation) to avoid growing more oxide under the field or zapping the area of interest during imaging. [20,21]

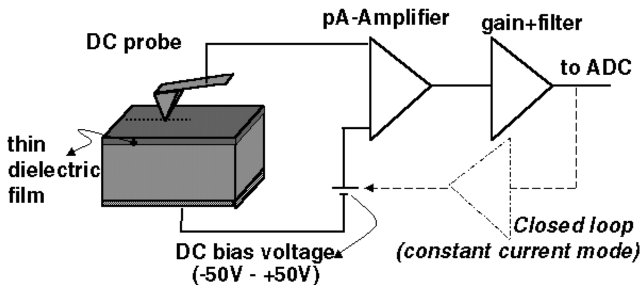


Figure 21: Schematic of typical conductive SPM configuration.

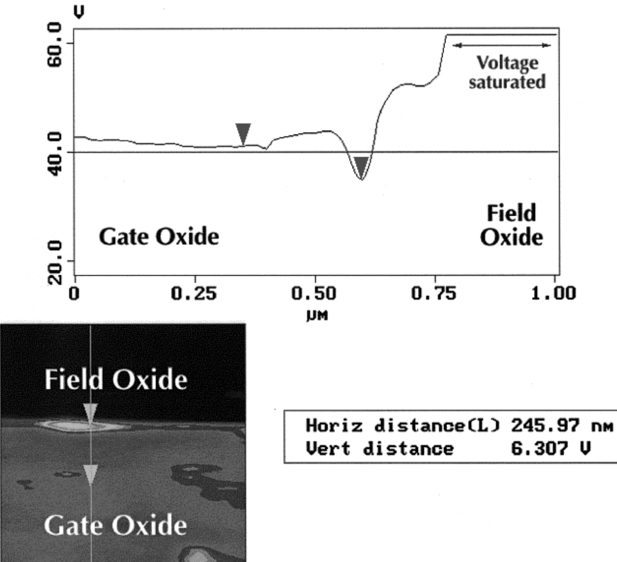


Figure 22: Gate oxide thinning at bird's beak measured by C-AFM at 1 pA.

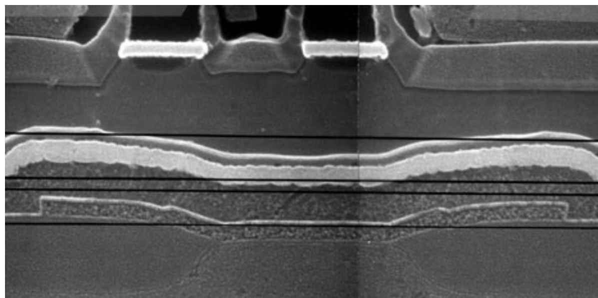


Figure 23: Cross sectional view of a flash cell showing the preparation target for surface lap deprocessing. Bottom black line is the endpoint target.

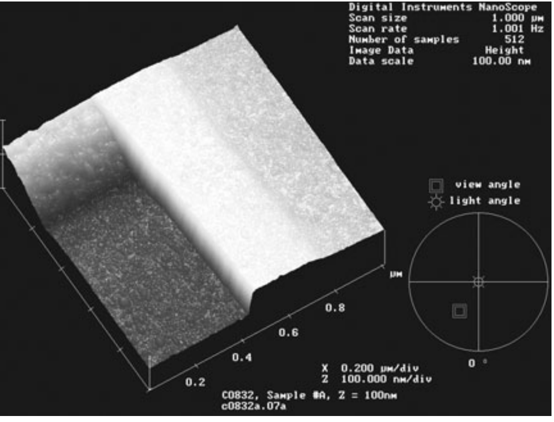


Figure 24: Topographical view of the exposed tunnel oxide after poly removal with Choline Hydroxide. Step height is 62nm.

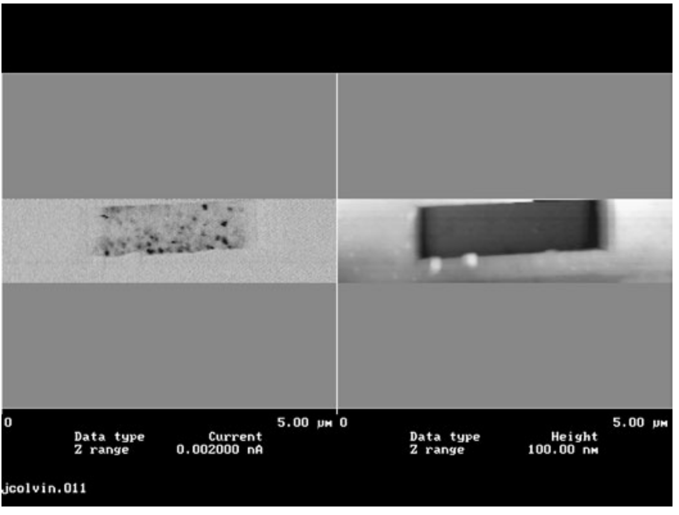


Figure 25. C-AFM image of the exposed gate oxide from Figure 24.

C-AFM and Passive Voltage Contrast

Passive Voltage Contrast (PVC) was originally developed by Jim Colvin in 1990. [22,23] The technique involves imaging the charging of gates contacts or metal lines to identify floating nodes vs. junctions as well as thin film contamination issues in

the SEM at low accelerating voltages. Using the AFM in C-AFM mode, PVC similar imaging data can be obtained with the advantage of obtaining I-V data. [24,25]

Figure 26 is typical data obtained from PVC. The contacts which are bright are P+, Gray=N+, Black=poly gate or floating. The C-AFM scan is shown in Figure 27. The abnormal contact has the same contrast as a P+ contact inferring a short to P+. Deprocessing confirms the short to P+ as shown in Figure 28.

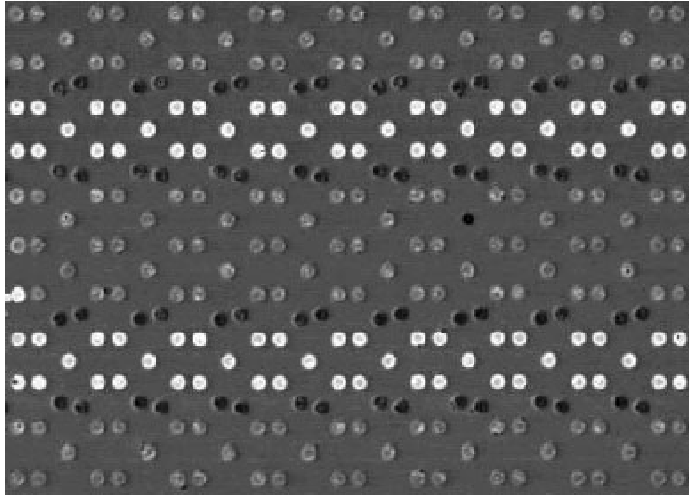


Figure 26: PVC image at 1KV accelerating voltage in the SEM. [26]

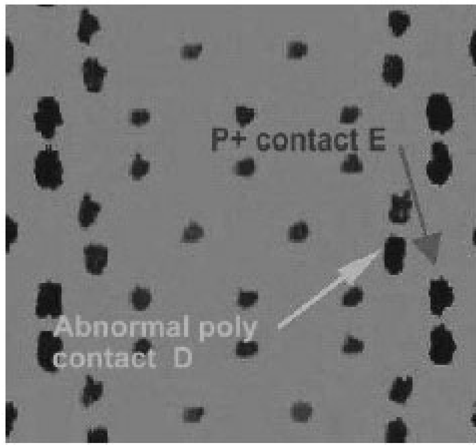


Figure 27: C-AFM current map with abnormal poly contact identified. [27]

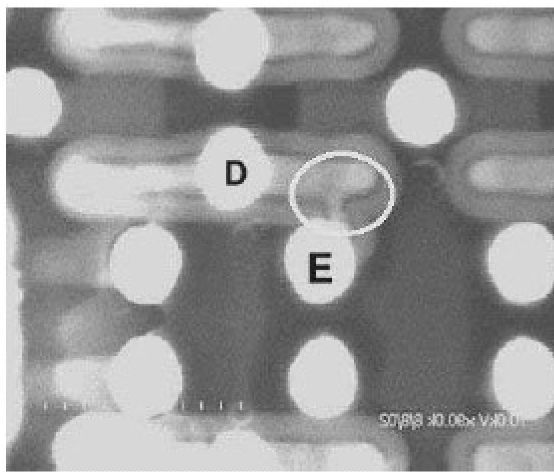


Figure 28: Deprocessing reveals a poly residue to P+ contact short. [27]

Atomic Force Probing (AFP)
The AFP is an SPM designed to allow probing of deep submicron nodes with multiple probes. The reasoning behind AFP is to extend AFM capabilities to a conventional probe station platform. Since we can no longer rely on optics to image and probe, the FIB has traditionally been used to lay down pads to allow probing. Success rate, quality of data due to beam charging/Vt drift and time to lay pads is a problem for the FIB (Figure 29). The AFM traditionally is difficult to navigate with, due to piezo drift and low magnification optics. Drift in the AFP is addressed with closed loop control, allowing accurate, repeatable probe placement. Transistor probing to obtain a family of curves is attainable with the AFP, even on 90nm nodes.

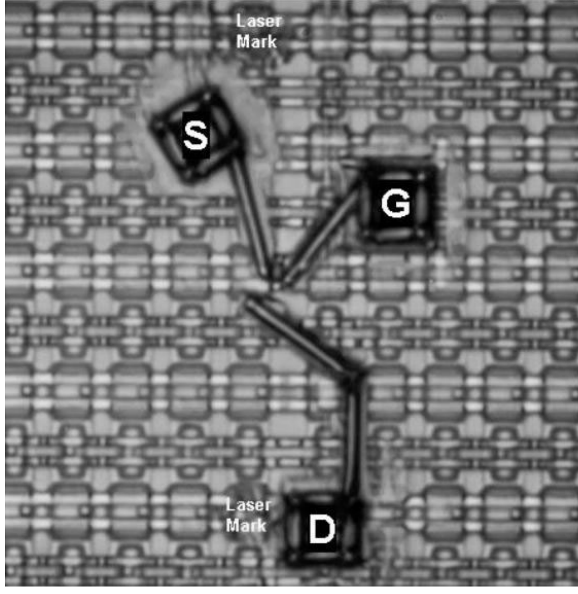


Figure 29: Manual contacts deposited on a select transistor at Contact1 with a FIB.

Since the probe tips for the AFP can probe features with multiple probes, spaced 250nm or greater, the need for FIB

pads is eliminated. Tip collision is avoided by scanning and retracting from a known feature with each probe. Afterwards the software keeps track of each tip allowing simultaneous scanning/probing of nodes. Tip diameters are typically 100nm with the capability to scan/image up to 5 probes. [28] Figure 30 illustrates probe placement and entrant angle. Figure 31 show typical views of the equipment. AFP's are most commonly operated in topographical mode and C-AFM mode to acquire waveforms or I-V data.

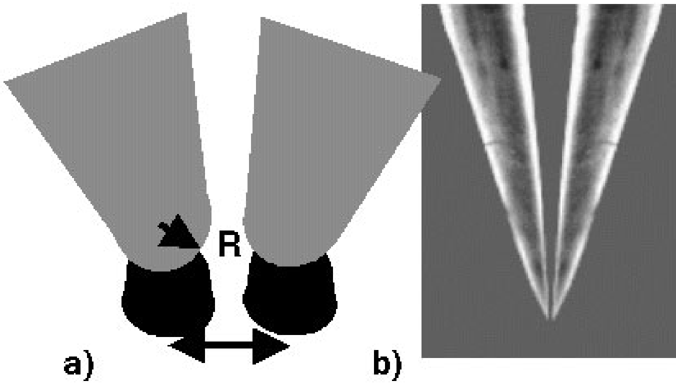


Figure 30: AFP probe tips showing the probe entrant angle. Courtesy Multiprobe Inc.

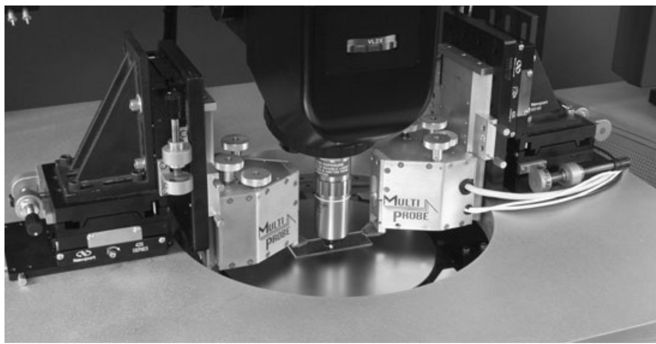


Figure 31: View of multiple AFM probes on a probe station for acquisition of I-V data. Courtesy MFI/Suss (upper image) and Multiprobe Inc. (Lower image)

About Sample Preparation Techniques

Sample preparation is straightforward and a few guidelines will result in repeatable clean results.

Use high selectivity etches such as TMAH or Choline Hydroxide for poly removal. Other etches include 49% HF, or metal etchants. Etches which are not highly selective such as a poly etch containing nitric and HF will not leave crisp definition associated with silicon to oxide boundary and will, in many cases, remove too much gate oxide. Mechanical surface lapping techniques allow layers to be partially exposed and then removed using a selective etch. In order to expose the gate oxide of a failing transistor to examine the gate oxide integrity topographically, the following procedure is employed:

For plan view with removal of poly for C-AFM:
1. Surface polish with colloidal silica.
2. Endpoint inspect for silicide removal.
3. Etch @ 90^0C with Choline Hydroxide 50% solution in water for 1 minute. (Aldrich Catalog 29,225-7)
4. Image exposed gate oxide.

For plan view at contact for C-AFM:
1. Surface lap with a chem-pol or polyimide pad using standard mechanical delayering procedures.
2. Endpoint at contact and image.

For SCM:
Preparation for capacitive probe requires that the surface be non-contaminated, as is true with any surface science technique. Avoid finger oils or other contaminants as capacitive probe is capable of detecting into the attofarad level with a spatial resolution of 10-15 nm. Polished or lapped surfaces must be smooth in order to avoid local trapping of ionic contaminants. Several methods may be used as follows:

A. Older method: Use a glass wheel with water. This method will result in a differential surface height due to differing material hardness of the metal and silicide layers but will yield reasonable SCM results.

B. Use diamond mylar films down to .1 um and finish with colloidal silica and a polishing cloth such as Final B to buff out the damage areas at the surface due to the diamond mylar film. Colloidal silica and a chem pol pad also works well. Do not attempt SCM without the final polish or all that will be seen is the damage from the diamond. This method is preferred due to the improved planarity of the surface coupled with good SCM results. Bake and UV are necessary to condition the surface for SCM.

Conclusions

The theory and practical application of the AFM has been presented with various modes of operation. Each year new modes of operation are added to the SPM which enhances its capability and allows us to routinely obtain data such as junction profiling, Fowler Nordheim tunneling maps of the surface with nanometer precision, probing and imaging of individual 90nm plugs that just a few years ago would have been deemed nothing more than science fiction.

Additional Sources of Information

J. Ebel et al., “Cross-sectional Atomic Force Microscopy of Focused Ion Beam Milled Devices”, International Reliability Physics Symposium Proceedings 1998. IEEE. pp. 157-162.

G.M. Fiege et al., “Temperature Profiling with Highest Spatial and Temperature Resolution by Means of Scanning Thermal Microscopy (SThM)”, ISTFA 1997 pp. 51-56.

R.M. Cramer et al., “The use of Near-Field Scanning Optical Microscopy for Failure Analysis of ULSI Circuits”, ISTFA 1996 pp. 19-26.

H. Yamashita, Y. Hata, “Grains Observation Using FIB Anisotropic Etch Followed by AFM Imaging”, ISTFA 1996 pp. 89-94.

B. Ebersberger et al., “Thickness Mapping of Thin Dielectrics with Emission Microscopy and Conductive Atomic Force Microscopy for Assessment of Dielectric Reliability”, International Reliability Physics Symposium Proceedings 1996. IEEE. pp. 126-130.

T. Hochwitz, et al., “DRAM Failure Analysis with the Force-Based Scanning Kelvin Probe”, International Reliability Physics Symposium Proceedings 1995. IEEE. pp. 217-222.

M. Masters et al., “Qualitative Kelvin Probing for Diffusion Profiling”, ISTFA 1995 pp. 9-14.

L. Ghislain and V. Elings, Near-field scanning solid immersion microscope, Appl. Phys. Lett. 72, 2779-2781 (1998).

A. Erickson, D. Adderton, Y. Strausser, and R. Tench, Scanning capacitance microscopy for carrier profiling in semiconductors, Solid State Technology 40, 125-133 (June 1997).

P. Hansen, Y. Strausser, A. Erickson, E. Tarsa, P. Kozodoy, E. Brazel, J. Ibbotson, U. Mishra, V. Narayanamurti, S. DenBaars, and J. Speck, Scanning capacitance microscopy imaging of threading dislocations in GaN films grown on (0001) sapphire by metalorganic chemical vapor deposition, Appl. Phys. Lett. 72, 2247-2249 (1998).

S. Magonov, V. Elings, and M.-H. Whangbo, Phase imaging and stiffness in tapping-mode atomic force microscopy, Surf. Sci. Lett. 375, L385-L391 (1997).

K Babcock, M. Dugas, S. Manalis, and V. Elings, Magnetic force microscopy: recent advances and applications, Materials Research Society Symposium Proceedings 355, 311-322 (1995).

K. Luo, Z. Shi, J. Lai, and A. Majumder, Nanofabrication of sensors on cantilever probe tips for scanning multiprobe microscopy, Appl. Phys. Lett. 68, 325-327 (1996).

Jon C. Lee, C. H. Chen, David Su, J. H. Chuang, “Investigation of Sensitivity improvement on Passive Voltage Contrast for Defect Isolation“, ESREF Proceedings (2002).

Kenneth Krieg, Richard Qi, Douglas Thomson, Greg Bridges, “Electrical Probing and Surface Imaging of Deep Sub-micron Integrated Circuits”, ISTFA Proceedings (1999).

Kenneth Krieg, Douglas Thomson, Greg Bridges, “Multiple probe Deep Sub-micron Electrical Measurements Using Leading Edge Micro-machined Scanning Probes”, ISTFA Proceedings (2001).

Yongxia Zhang, Yanwei Zhang, J. Blaser, T. S. Sriram, A. Enver, and R. B. Marcus, “A thermal microprobe fabricated with wafer-stage processing”, Rev. Sci. Instr. 69, 2081-2084 (1998).

A. Olbrich, “Nanoscale Electrical Characterization of Thin Oxides with Conducting Atomic Force Microscopy”, International Reliability Physics Symposium Proceedings 1998. IEEE. pp. 163-168.

Ann N. Campbell, Edward I. Cole Jr., Bruce A. Dodd, and Richard E. Anderson, “Internal Current Probing of Integrated Circuits Using Magnetic Force Microscopy”, International Reliability Physics Symposium Proceedings 1993. IEEE. pp. 168-177.

References

1. Sergei N. Magonov, "Surface Characterization of Materials at Ambient Conditions by Scanning Tunneling Microscopy (STM) and Atomic Force Microscopy (AFM)", Applied Spectroscopy Reviews, 28(1&2), pp.30-33.
2. IBID p. 38.
3. Q. Zhong, D. Inniss, K. Kjoller, and V. Elings, Fractured polymer/silica fiber surface studied by tapping mode atomic force microscopy, Surf. Sci. Lett. 290, L688-L692 (1993).
4. Jim Colvin, "Advanced Methods for Imaging Gate Oxide Defects with the Atomic Force Microscope", Proceedings of the International Symposium for Testing and Failure Analysis, pp 271-276, 1992.
5. Jim Colvin, "The Identification and Analysis of Latent ESD Damage on CMOS Input Gates", 1993 EOS/ESD Symposium Proceedings, EOS-15, pp 109-116.
6. S. Magonov, V. Elings, and M.-H. Whangbo, Phase imaging and stiffness in tapping-mode atomic force microscopy, Surf. Sci. Lett. 375, L385-L391 (1997).
7. S.A. Joyce, J.E. Houston, T.A. Michalske, "Differentiation of topographical and chemical structures using an interfacial force microscope," Appl. Phys. Lett. 60 (10) 1992, 1175-1177.
8. N.A. Burnham, R.J. Colton, H.M. Pollock, "Interpretation of force curves in force microscopy" Nanotechnology 4 (1993) 64-80.
9. W. Mertin, R. Weber, F. Seifert, E. Kubalek, G. Zimmermann, C. Boit, "Contactless Failure Analysis of Integrated Circuits Via Current Contrast Imaging with Magnetic Force Microscopy", ISTFA 2001. Pp. 199-208.
10. F. Du, G. Bridges, D.J. Thomson, R. Gorganthu, S. McBride and M. Santana Jr. "Enhancements of Non-contact Measurement of Electrical Waveforms on the Proximity of a Signal Surface Using Groups of Pulses", ISTFA 2002 Proceedings, pp. 483-492.
11. A. Majumdar, Tutorial on scanning thermal microscopy, 1998 Tutorial Notes, International Reliability Physics Symposium, pp. 2b 1-16 (March 1998).
12. EXW Wu, XH Zheng, JCH Phang, LJ Balk, JR Lloyd, "Characterization of Interconnect Defects due to Electromigration using Scanning Thermal Microscopy", ISTFA 2003, pp. 419-424.
13. A. Henning, T. Hochwitz, Scanning probe microscopy for 2-D semiconductor dopant profiling and device failure analysis, Mat. Sci. Eng. B42, 88-98 (1996).
14. T. Yamamoto, Y. Suzuki, M. Miyashita, H. Sugimura, N. Nakagiri, Scanning Capacitance Microscopy as a Characterization Tool for Semiconductor Devices, Jpn. J. Appl. Phys. 36, 1922-1926 (1997).
15. A. Erickson, D. Adderton, Y. Strausser, and R. Tench, Scanning capacitance microscopy for carrier profiling in semiconductors, Solid State Technology 40, 125-133 (June 1997).
16. P. De Wolf, Clarysse, W. Vandervorst, Hellemans, Niedermannn, Hanni, Cross-sectional nano-spreading resistance profiling, J.Vac. Sci. Technolog. B 16(1) 355-361 (1998).
17. J. Colvin, "Identification and Analysis of Parasitic Depletion Mode Leakage in a Memory Select Transistor", Proceedings of the International Symposium for Testing and Failure Analysis, 247-254 (2000).
18. K. Jarausch, J.F. Richards, L. Denney, A. Guichard, P.E. Russell, Site specific implant profiling using FIB assisted SCM, Proceedings of the International Symposium for Testing and Failure Analysis, 467-471 (2002).
19. Application Notes, Digital Instruments, Veeco Metrology Group.
20. A.Olbrich, B. Ebersberger, C. Boit, "Local electrical thicknessmapping of thin oxides with conducting atomic forcemicroscopy" ISTFA Proceedings (2000).
21. M. Porti, M. Nafria, X. Aymerich, A.Olbrich, B. Ebersberger,"Nanoscale observations of the electrical conduction of ultrathin SiO2 films with Conducting Atomic Force Microscopy",IRPS 2001 pp 156-162.
22. J. Colvin, "A New Technique to Rapidly Identify Gate Oxide Leakage in Field Effect Semiconductors Using a Scanning Electron Microscope", ISTFA 1990, pp331-336.
23. J. Colvin, "The Identification and Analysis of Latent ESD Damage on CMOS Input Gates", 1993 EOS/ESD Symposium Proceedings, EOS-15, pp 109-116.
24. Andrew N Erickson, "What Can a Failure Analysis Dowith an AFM", ISTFA Proceedings (2002) pp455-459.
25. J. Lee, J.H. Chuang, "Fault Localization in Contact Level by Using Conductive Atomic Force Microscopy", ISTFA 1993, pp413-418.
26. IBID p. 414.
27. IBID p. 416.
28. Andrew N Erickson, "What Can a Failure Analyst Do with an AFM", ISTFA Proceedings (2002) pp 455-459.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Energy Dispersive X-ray Analysis

W. Vanderlinde
Laboratory for Physical Sciences

Introduction

By far the most common micro-analytical technique in the failure analysis laboratory is energy dispersive x-ray spectroscopy, known as EDX or EDS. It is commonly available because it is relatively cheap and easy to add an x-ray detector to a scanning electron microscope. However, as seen in Table 1, EDS lacks the lateral spatial resolution, depth resolution and sensitivity of many other techniques. Nevertheless, with careful analysis EDS can measure many elements down to a composition of about 0.1%, or with spatial resolution and depth sensitivity of 1 micron or less, so it is a reasonably powerful technique for many failure analysis problems.

Table 1: Properties of analytical techniques including energy dispersive x-ray spectroscope (EDS) in the SEM or STEM (with thin samples), wavelength dispersive x-ray spectroscopy (WDS), Auger electron spectroscopy (AES), x-ray photoelectron spectroscopy (XPS, also knows as ESCA), secondary ion mass spectrometry (SIMS), Rutherford backscattering spectrometry (RBS), Fourier transform infrared spectroscopy (FTIR), and total-reflection x-ray fluorescence (TXRF). Each of these capabilities is given for "ideal" conditions, and optimum sensitivity and lateral resolution may not be achievable at the same time.

Technique	Sensitivity	Lateral Spatial Resolution	Depth Sensitivity	Depth Profiling?
SEM/EDS	0.1%	~ 1 micron	~ 1 μm	No
SEM/WDS	10 ppm	~ 1 micron	~ 1 μm	No
STEM/EDS	0.1%	10 nm	100 nm	No
AES	0.1%	20 nm	2 nm	Yes
XPS	0.1%	1 mm to 10 microns	2 nm	Yes
SIMS	1 ppb	30 nm	10 nm	Yes
RBS	100 ppm	1 mm	10 nm	Yes
FTIR	100 ppm	3 microns	1 micron	No
TXRF	10 ppb	1 mm	1 nm	No

EDS overview

X-rays are produced when a high energy electron beam enters a solid material. These x-rays are always created in the SEM regardless of whether an x-ray detector is present. To perform EDS, one can simply attach an x-ray detector to the SEM chamber.

Safety note: 30 keV electrons can produce a significant amount of hazardous radiation when they strike a sample in a SEM. Vacuum chamber walls will normally protect the user, but if a view port is added to the chamber then leaded glass should be used.

X-rays produced by electron bombardment of materials are of two types: characteristic and Bremsstrahlung. Characteristic x-rays occur when an incoming high energy electron collides with an atom and ejects an electron from an inner shell leaving a vacancy, see Figure 1. This vacancy or "inner shell ionization" will be filled when an outer shell electron drops down, thereby releasing energy. This energy may be released as an x-ray, or it may be carried away by another electron, known as an Auger electron, emitted from an outer shell. The Auger effect tends to dominate for low atomic number elements, and so relatively few x-rays are produced for light elements. The emitted x-rays and Auger electrons have energy that is characteristic of the electron transitions for that atom, and this information can be used for microanalysis. Auger electrons will not be further discussed in this article.

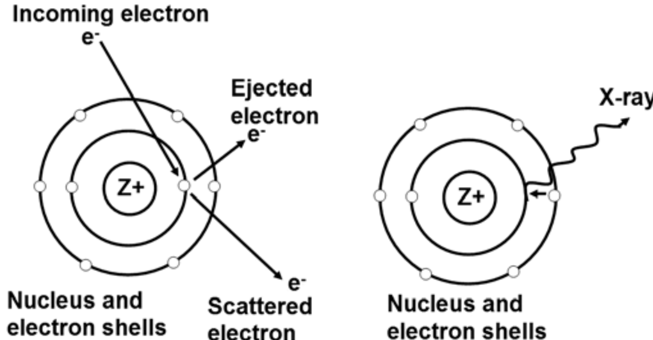


Figure 1: Characteristic x-ray production.

Physicists typically label electron shells by the letters K, L, M, N, and O. This may be less familiar to some readers than the more common notation used by chemists: $1S_{1/2}$, $2S_{1/2}$, $2P_{1/2}$, etc. However, they represent the same thing, i.e. the K shell is the $1S_{1/2}$ shell, the L_1 sub-shell is the $2S_{1/2}$ sub-shell, etc. See Figure 2.

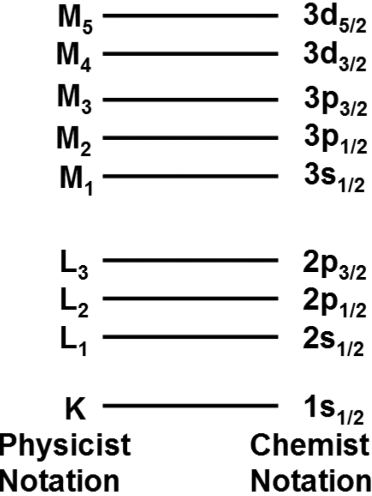


Figure 2: Comparison of two different notations for electron shells.

Using the physicist's notation, x-ray lines are labeled according to the shell that was originally ionized. Since more than one electron might drop into the vacancy, there will be two or more x-ray lines for each shell. Typically the strongest line with the largest number of x-ray counts is called the alpha line, the next strongest is called the beta line, etc. Since an x-ray carries away one quantum of angular momentum, transitions between sub-shells with the same orbital angular momentum state are forbidden for x-ray emission (but not for Auger emission). Thus there is no x-ray line for an $L_1 \rightarrow K$ transition.

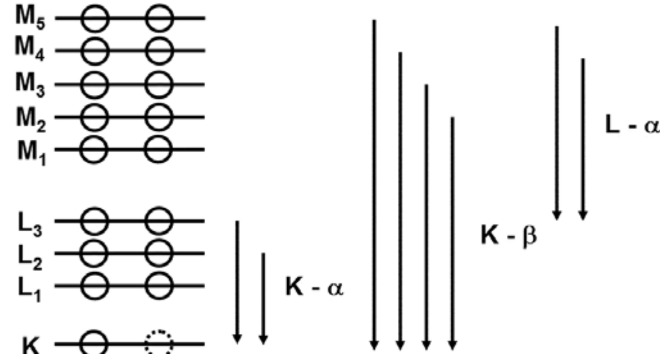


Figure 3: Some K-shell and L-shell electron transitions for EDS.

Consider Figure 3 where one of the K-shell electrons has been ejected leaving a vacancy. Either the L_2 or L_3 electron can drop into the K shell resulting in a K-α x-ray emission. The x-ray energy is given by:

$$E_{K-\alpha} = E_K - E_L$$

Since L_2 and L_3 have slightly different energies the K-α line is actually split into two lines, $K\alpha_1$ and $K\alpha_2$, but these lines are too close in energy (~ 1 eV for Si) for an EDS detector to resolve. In fact, for light elements like Si, even the K-α and K-β lines will not be resolved since the separation is only 90 eV. Alternately, if one of the M shell electrons drops to the K-shell vacancy this results in a K-β x-ray. If instead an L_3 electron were ejected, a M_4 or M_5 electron could drop down producing an L-α x-ray.

There must be enough electrons present in a given element for a particular x-ray line to occur. For example, H, He, and Li do not produce characteristic x-rays at all. Elements without at least some M electrons cannot produce L-series x-rays. In general, the higher the atomic number of an element, the more x-ray lines will be produced.

A typical x-ray spectrum is shown in Figure 4. In addition to the characteristic x-ray lines, there is also a broad x-ray background called "Bremsstrahlung" (German for "breaking") radiation. Bremsstrahlung radiation is caused by the deceleration of electrons as they pass through the sample. Bremsstrahlung will produce a broad background of x-rays from close to zero energy all the way up to the primary electron beam energy of the SEM, E_0. The amount of Bremsstrahlung is somewhat dependent on the average atomic number of the sample, however it is not a good source of quantitative information and simply acts as an annoying source of background noise for x-ray analysis.

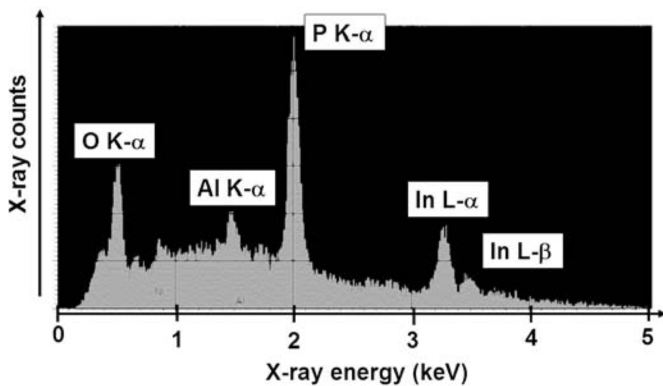


Figure 4: Energy dispersive x-ray spectrum for a sample with InP, Al and O.

Since each x-ray is produced by one incident electron, you will never see x-rays of higher energy than the primary electron beam energy. For example, an incident beam of 3 keV electrons could produce the P K-α x-rays at 2.01 keV but not the In L- α x-rays at 3.29 keV. In order to get a good x-ray signal one should use a primary electron beam with energy at least 2.5 times the energy of the desired x-ray peak.

Some x-ray peaks for common electronic materials are shown in Table 2. Note that light elements such as Al and Si produce only K-series x-rays because only the K and L shells are occupied. Intermediate weight elements like Ni and Cu will have both K-series and L-series lines. Heavier elements such as W, Au, and Pb will have K-series, L-series, and M-series lines, but their K-lines are around 50 to 70 keV and they can only be produced in a high energy TEM or STEM. Since SEMs are normally limited to 30 keV beam energy, x-ray peaks above that energy cannot be produced and are left off the table.

Since most elements have at least one x-ray line below 5 keV, it is possible to do x-ray analysis at relatively low primary beam energy (< 10 keV.) The presence of multiple x-ray lines for one element can be beneficial when peak overlap occurs. For example, the slight differences in peak shape and energy between the Si K-α peak at 1.74 keV and the W M-α peak at 1.77 keV can be difficult to distinguish. When doing x-ray mapping it is almost impossible to distinguish the two elements based on the Si K-α peak and the W M-α peak. The W L-α peak at 8.40 keV provides an alternative way to detect W, but it will require the use of much higher primary beam energy. Increasing the beam energy will have a significant effect on the depth sensitivity and lateral spatial resolution of the resulting spectrum or elemental map.

X-ray detectors

EDS

The most common x-ray detector is the lithium-drifted solid state detector, known as the EDS detector. This detector is essentially a sophisticated photo-diode. When an x-ray enters the detector, it will create electron-hole pairs. One electron-hole pair is created on average for every 3.8 eV of x-ray energy, so a sensitive amplifier can count the number of electron-hole pairs created by an x-ray and provide a measure of the x-ray energy. However, due to statistical variations in the number of electron-hole pairs created and electronic noise in the amplifier, the accuracy of the energy measurement is limited to about 140 eV for EDS detectors. (The resolution varies with energy but by convention is measured as the full-width-half-maximum for Mn K-α at 5.9 keV.) Characteristic x-ray lines appear much broader in EDS spectra than their natural line widths that are on the order of 1 to 10 eV.

Table 2: Characteristic x-ray energies for some common electronic materials. All energies are in keV.

Element	K-α	K-β	L- α	L- β	L-γ	M- α	M- β
Al	1.49	1.55	-	-	-	-	-
Si	1.74	1.83	-	-	-	-	-
P	2.01	2.14	-	-	-	-	-
Ti	4.51	4.93	0.45	0.46	-	-	-
Ni	7.47	8.26	0.85	0.87	-	-	-
Cu	8.04	8.90	0.93	0.95	-	-	-
Ga	9.24	10.26	1.10	1.12	-	-	-
As	10.53	11.73	1.28	1.32	-	-	-
In	24.21	27.27	3.29	3.49	3.92	-	-
Sn	25.27	28.48	3.44	3.66	4.13	-	-
W	>30	>30	8.40	9.67	11.28	1.77	2.04
Au	>30	>30	9.71	11.44	13.38	2.12	2.41
Pb	>30	>30	10.55	12.61	14.76	2.34	2.65

A major disadvantage of the EDS detector is that it must be kept at low temperature during operation. Most detectors are cooled by liquid nitrogen although some newer systems have Peltier thermoelectric coolers that do not require liquid nitrogen. Since the sample chamber must be vented periodically and we don't want moist room air to condense ice onto the detector, it is necessary to have a vacuum window in between the detector and the chamber. The various types of vacuum windows all tend to absorb low energy x-rays. This is particularly a problem for the light elements that have low energy characteristic x-rays are also very inefficient at creating x-rays. Elements lighter than fluorine (Z=9) cannot be detected using older beryllium windows. Some newer polymer windows can detect down to boron (Z=5). Some systems have a movable window that can allow the detector to operate "windowless" when at good vacuum, but the efficiency of the detector is such that boron is still the lightest element that can be detected.

WDS

A much more expensive, complex and difficult detection scheme is the wavelength dispersive x-ray detector (WDS). X-rays are diffracted by bent crystals into a detector so that a very precise measurement of the x-ray wavelength can be made. Since the wave properties of the x-rays are being used, the method is called wavelength rather than energy dispersive detection, but the spectra are generally displayed in energy (keV) like EDS spectra. WDS has about 10 eV resolution but several different crystals must be used since each can cover only a limited range of the x-ray energy spectrum. Whereas EDS simultaneously detects all energies at once, the WDS detector must slowly scan across the energy spectrum, a considerably drawback for mapping many elements at once.

Although the excellent energy resolution (10 eV) of WDS makes it better than EDS for resolving overlapping peaks, the real advantage is improved sensitivity. Recall that the characteristic x-ray lines lie on top of a broad Bremsstralung background. If there are N counts of Bremsstrahlung x-rays in a given energy window, then there will be random noise ~ $N^{1/2}$ that will limit our ability to detect trace elements at that energy. If we can narrow the window from 140 eV to 10 eV, then the number of Bremsstralung x-rays and their associated noise will be greatly reduced. This vastly improves the sensitivity of WDS compared to EDS for detecting trace elements in a sample.

Micro-calorimetry

Recently a new x-ray detection scheme was implemented known as micro-calorimetry. [1] A superconducting thermometer held near zero absolute absorbs x-rays and then measures the tiny of amount of thermal energy that is created by the x-ray in the detector. Energy resolution of 10 eV has been demonstrated, and 2 eV may be possible. This technique allows for high sensitivity to trace elements. It also measures the entire spectrum simultaneously unlike the WDS detectors that must be scanned. 2 eV energy resolution would be sufficient to study peak shifts from chemical bonding as is done with XPS.

X-ray artifacts

The EDS detector is not a perfect detector, and a number of artifacts must be taken into consideration when analyzing spectra. First and most obviously, the limited energy resolution of the spectrometer causes the characteristic peaks to appear much broader (~140 eV FWHM) than the true energy width of the peaks (~ 1 to 10 eV). This peak broadening is not perfectly Gaussian but has some asymmetrical distortion due to the way the electron-hole pairs are counted.

Second, EDS detectors have very limited detection efficiency at low energy as discussed above. Depending on the type of vacuum window, the detection efficiency may be a highly variable function at low energy (below 600 eV) due to absorption band edges for various materials in the window and the detector. The result is that Bremsstralung, which increases as energy decreases, will be "cut-off" by the detector at some energy in a way that could resemble a characteristic peak. Therefore one must take great care in determining whether low energy peaks are valid characteristic peaks like carbon (277 eV.)

Third, it is possible for two x-rays to arrive at the detector so close together in time that their energy is counted as if it were a single x-ray. These x-rays will appear as an extra peak at twice the correct energy, known as a "sum peak." EDS detectors will only give an accurate energy measurement if all the electron-hole pairs from one x-ray are completely counted before a second x-ray arrives. Circuitry is added to the detectors to avoid "pulse pile-up" by rejecting a signal if it appears to be from two different x-rays. The time that the detector is unavailable due to pulse pile-up rejection is known as the "dead-time." It is good to keep dead-time below 50% as this will minimize the presence of sum peaks. It is possible to run EDS detectors at 90% or more dead-time, but this is unwise since it results in a reduction in the overall x-ray count. To reduce the dead-time, simply use less beam current on the sample.

Fourth, it is possible that an incoming x-ray may fluoresce a silicon atom in the detector resulting in a Si K-α x-ray. If this x-ray is able to escape from the detector, it will carry away 1.74 keV of energy that should have been measured for the incoming x-ray. The result will be an x-ray peak at 1.74 keV below the proper energy. "Escape peaks" will have some small fraction of the number counts of the main peak they are associated with, so they are usually only noticeable for the largest peaks in the spectrum. It is important to consider the possibility of escape peaks when labeling spectrum. A careless analyst may falsely identify escape peaks as trace elements that are not actually present in the spectrum.

Finally, it is possible for an incoming x-ray to fluoresce a silicon atom in the inactive or "dead" layer on the EDS

detector. If this x-ray is then absorbed by the active part of the detector, it will appear as a silicon K-α x-ray at 1.74 keV in the spectrum. This "silicon internal fluorescent peak" can make a sample appear to have a trace amount of silicon when none is actually present.

Of the various x-ray artifacts, the silicon escape peak will most often lead the user astray. Even very experienced analysts have been known to publish data claiming the presence of trace elements in a sample when the trace element peaks were actually silicon escape peaks.

Qualitative Analysis

Peak identification

Characteristic x-ray peaks can be identified using tables such as Table 2. Most x-ray analysis programs will also have automated software for peak identification, although the user should be careful since these programs often assign peak identities that are not realistic, such as rare earth elements on samples which are unlikely to have such elements. One should always review peak identifications manually and consider the possibility of x-ray detector artifacts such as sum peaks, escape peaks, or a silicon internal fluorescence peak.

The analyst should also consider that once an element has been positively identified based on its strongest line, the secondary lines will also be present. For example, suppose a defect is identified to have the element W based on the L-α and L-β peaks at 8.40 keV and 9.67 keV. Another peak at about 1.75 keV could be plausibly identified as Si. However the W M-α peak at 1.77 keV must also be present in the spectrum, and this peak is easily confused for the Si K-α peak at 1.74. In this case, it might be good to compare the observed spectrum with a reference spectrum on pure W to see if the low energy peak is purely due to the W M-α peak or if there is also some Si K-α signal.

When considering a completely unfamiliar spectrum, it is useful to notice that K-series peaks tend to be sharper than other peaks and they always come in K-α / K-β pairs. The K-β peak is slightly higher in energy and about 10% the peak height of the K-α peak. For light elements like Si the K-α / K-β pair are not resolved, but the resulting combined peak appears somewhat lopsided toward the high energy side due to the K-β x-rays. L and M series peaks tend to be broader and the M series x-rays have many small peaks due to the large number of possible electron transitions.

The number of counts in an x-ray peak will *generally* be proportional to the amount of that element in the sample, so the large peaks will be the major constituents and the small peaks will be minor constituents. However, there are many factor affecting the size of a given peak including inner-shell ionization cross-sections, absorption corrections, fluorescence corrections, and detector sensitivity, and so one must be careful in assessing material composition based on peak size alone. This issue is further discussed in the section on Quantitative Analysis.

Depth and spatial resolution

When an electron beam enters a sample, the electrons are scattered laterally and gradually lose energy due to collisions with atoms in the sample. The overall result is that the electron energy is deposited in a tear-drop shaped figure known as the "excitation volume" (see Figure 5.) Unlike secondary electrons, many characteristic x-rays tend to have a long path length in the sample relative to the beam penetration. As a result, the depth sensitivity and lateral spatial resolution of EDS analysis will depend ***strongly*** on the electron beam energy.

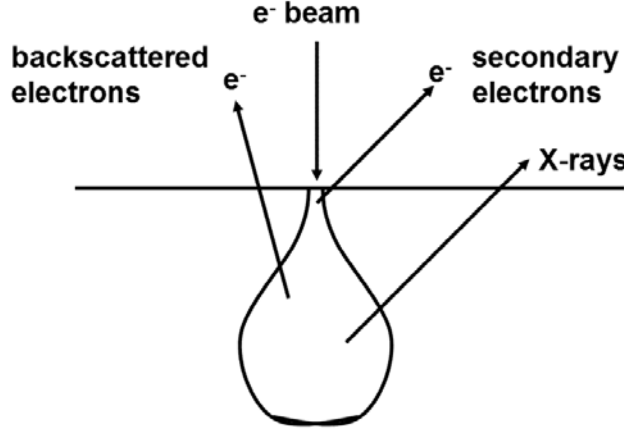


Figure 5: Electron beam – sample interaction volume and interaction products.

The electron depth penetration can be calculated from the Kanaya-Okayama (K-O) formula [2]:

$$R_{KO} = 0.0276 * A * E_o^{1.67} / Z^{0.89} * \rho$$

where:
R_{KO} = electron range (in microns)
A = atomic weight
E_o = incident electron beam energy (in keV)
Z = atomic number
R = density (in gm/cm^3)

Electron ranges calculated from the K-O formula for several materials commonly found in integrated circuits are shown in Table 3. Note that electron range increases rapidly with the incident electron beam energy. This fact should be considered when selecting the electron beam energy to be used for x-ray analysis. For example, if a defect is in the top 0.1 micron of the sample surface, it may be best to use low beam energy (≥ 5 keV) so that the x-rays are mostly from the area of the sample near the defect. This will also reduce the lateral electron scattering in the sample and improve the lateral spatial

resolution. Of course, it may be necessary to use higher electron beam energy in order to create higher energy x-ray lines.

Depth penetration also decreases with increasing material density. Dense metals like W and Au have limited electron ranges, and consequently EDS spectra are more surface sensitive in those materials.

Table 3: Electron beam range (in microns) as a function of energy for several materials commonly found in integrated circuits, calculated from the Kanaya-Okayama formula.

Electron beam energy (keV)	Al	Si	SiO_2	Cu	W
1	0.028	0.031	0.036	0.0097	0.0057
3.5	0.023	0.26	0.29	0.079	0.046
5	0.41	0.47	0.52	0.14	0.084
10	1.32	1.49	1.66	0.45	0.27
15	2.59	2.92	3.27	0.90	0.53
20	4.19	4.73	5.28	1.45	0.85
25	6.08	6.87	7.70	2.10	1.23
30	8.25	9.31	10.4	2.85	1.67

X-rays will not be created throughout the entire excitation volume, but will be limited to the depth for which the electrons still have sufficient energy to create the inner shell ionizations that result in a particular x-ray. Thus the maximum depth for creation of an x-ray produced by a transition with initial ionization energy E_i will be:

$$R_{max} = R(E_o) - R(E_i)$$

For example, $R(E_i)$ for the Si K-shell is $R_{KO}(E_i=1.84 \text{ keV})$ = 0.09 micron. Therefore a 5 keV electron beam will penetrate 0.41 micron into Si, but will only create Si K series x-rays for the first 0.32 micron of depth.

Some x-rays will be absorbed in the sample and thus prevented from reaching the detector. Absorption cross-sections vary strongly with energy due to the presence of sharp "absorption edges," and x-ray absorption can vary by orders of magnitude with a small change in energy. A full discussion of x-ray absorption and fluorescence lies outside the scope of this article, but it should be noted that for beam energies of 10 keV or less, a large fraction of most characteristic x-rays will have sufficient range to exit the sample from most of the penetration depth.

Monte-Carlo simulation programs provide the easiest way to estimate the depth and lateral distance over which x-rays are created in a sample. These programs use tabulated data and random numbers to project electron paths and x-ray creation in samples of any given composition. For example, Figure 6 shows a Monte-Carlo simulation of 15 keV electrons in InP using the "Electron Flight Simulator" program. [3] The depth at which P K-α x-rays are created, and the fraction that can escape are also shown. (This data corresponds to the spectrum shown in Figure 3.) In this case, P K- α x-rays are created about 1.25 microns into the sample, but most of the x-rays that are detected come from the top 0.85 micron of the sample. It can also be seen that the x-rays will be distributed about 0.5 micron laterally from the electron beam entry point, and this will limit the lateral spatial resolution of the technique.

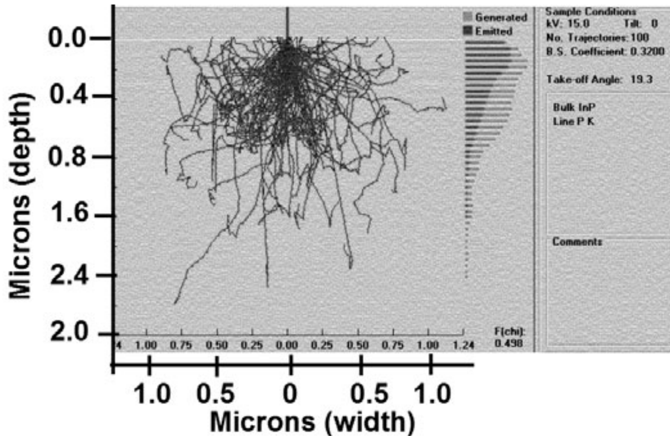


Figure 6: Monte-Carlo simulation of electron trajectories for 15 keV electrons in InP. The number of P K-α x-rays that are created and emitted from each depth are also shown.

X-rays will be created from the entire area of the sample surface that is illuminated by the electron beam. A small area raster box is often used to create a small area for x-ray analysis. This box may be moved onto specific defects or particles for analysis. But if the box is smaller than the lateral spread of the electron beam, then some x-rays may originate from outside the desired area of analysis. X-ray fluorescence (i.e. x-rays fluorescing atoms to produce more x-rays) can also result in a larger than expect analysis area.

Quantitative x-ray analysis

With a great deal of effort, EDS analysis can be made reasonably quantitative, and microanalysis results can be achieved that are accurate to better than 1%. This type of work requires careful analysis and should be done using reference standards for all the elements being analyzed. Although most x-ray analysis computer programs will run a "standard-less quantification" with the click of a single button, this is unlikely to result in a reliable quantitative result.

Two common methods for quantification are known as "ZAF" and "$\Phi(\rho Z)$." The ZAF method will be discussed briefly here. This method consists of measuring the x-ray signal I from a given element in an unknown sample and then comparing it to the signal for the same elemental peak in a reference sample. In the absence of corrections, the two will be proportional, i.e. 100 counts/sec from pure copper and 1 count/sec in an unknown sample suggest a 1% copper composition in the unknown sample. However, there are a number of "matrix corrections" that need to be made because the copper in the

unknown sample is surrounded by different atoms (i.e. a different matrix) than in the pure copper sample. The correction factors are:

Atomic number correction: electron trajectories depend strongly on the atomic number of the matrix. For a high Z matrix, the range will be shorter, and a large fraction of backscattered electrons may create additional x-rays as the electrons pass back through the sample a second time on their way out.

Absorption correction: x-rays for a given elemental peak may be strongly absorbed as they exit the sample. This results in substantial loss of x-ray signal depending on the matrix composition.

Fluorescence correction: x-rays from matrix elements may fluoresce a given element resulting in extra x-rays. For certain combinations of elements this can be a very large effect.

The quantification can be summarized as follows:

$C_s / C_r = [ZAF]\ I_s / I_r$

where:
C_s = weight fraction of an element in the sample
C_r = weight fraction of an element in the reference
Z = atomic number correction
A = absorption correction
F = fluorescence correction
I_s = x-ray signal from an element in the sample
I_r = x-ray signal from an element in the reference

The actual values of Z, A, and F for a given element and matrix can be calculated using mathematical programs based on electron trajectory, x-ray absorption, and x-ray fluorescence data. It is vital to have correct information including the detector working distance, tilt, and take-off angle, plus the detector window type. It is important to note that any quantification method assumes that the sample is flat and homogeneous within the volume excited by the x-ray beam. If the x-rays originate from an area with varying composition or an uneven surface, the absorption and fluorescence corrections may be highly inaccurate. One is often interested in determining the composition of a particle. Special versions of the ZAF correction have been used to model x-ray emission from spherical particles, but in general this is a very hard problem.

For a more detailed discussion of the ZAF method see reference [4].

When reporting elemental analysis results, it is important to distinguish between atomic percent composition, and weight percent composition. Atomic percent is based on the stoichiometry of the sample, i.e. how many atoms of each type are present. Metallurgists usually use weight percent, since they add so many grams of each metal to form an alloy. The difference can be significant. For example, 60-40 tin-lead solder has 60 wt.% tin and 40 wt.% lead. However, the atomic percent composition of the solder is 72.4 at.% tin and 27.6 at.% lead. In most cases, analysis techniques like EDS will count the number of atoms and therefore report atomic percent, but some computer programs may convert the results to weight percent. It is important to know which method is being used.

X-ray elemental mapping

EDS elemental maps

One of the most important uses of EDS analysis is to map the elemental composition in an sample. Although a simple comparison of spectra between a defect and a reference area may have as much useful information, an x-ray map produces a striking image useful for presentations and reports. As discussed above, the surface sensitivity and lateral spatial resolution of the map will depend strongly on the electron beam energy. The analyst must also be aware of certain artifacts that can occur during mapping.

EDS mapping will be illustrated by discussing one particular example in extensive detail. The sample consisted of 1.25 micron wide 0.5 micron thick Al metal lines in a SiO_2 dielectric. This sample was prepared by mechanically polishing a finished circuit so that the top level of Al metal was exposed. A beam energy of 5 keV was selected as this will produce sufficient x-rays from Si K- α (1.74 keV), Al K-α (1.49 keV) and O K-α (0.53 keV) while minimizing the electron beam penetration. If we wish to map for other elements or survey the spectrum for unknown elements a higher energy electron beam would be appropriate, but for mapping these elements 5 keV is adequate. A survey spectrum taken over a broad area of the sample is shown in Figure 7.

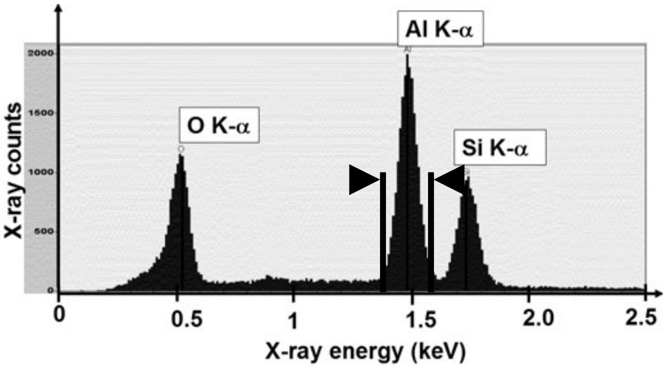
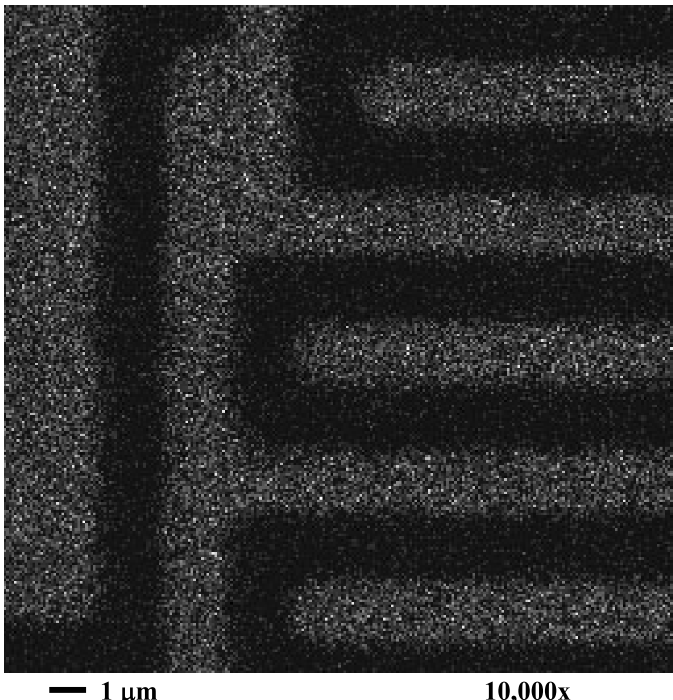


Figure 7: EDS spectrum of Al metal lines in SiO_2 dielectric at 5 keV. An energy window for mapping is shown around the Al peak.

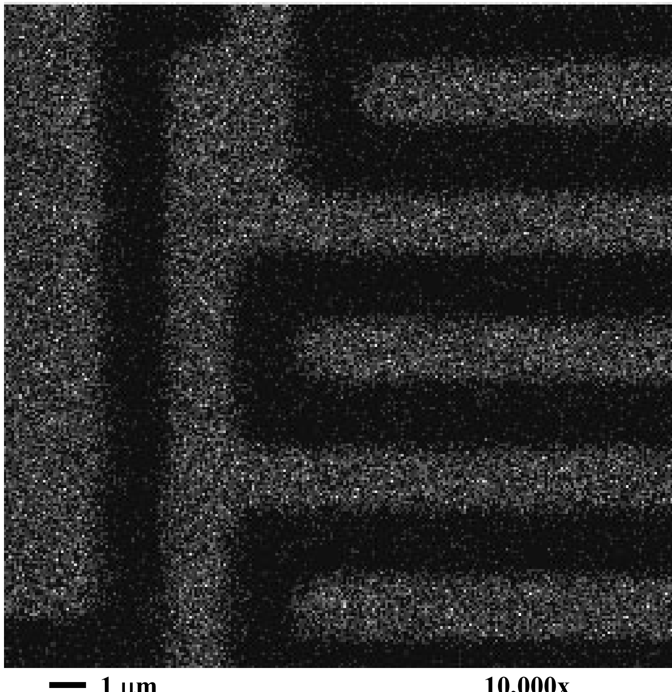
The user typically will select an energy window around each peak indicating that all x-rays within that window are to be counted as part of that elemental map. (Note that some Bremsstralung x-rays will always lie within the window, and that can lead to artifacts in the x-ray map.) The electron beam

is then scanned across the map area, and an x-ray spectrum is acquired at each location. A map is then generated for each element. Since all x-ray energies are acquired in parallel by EDS detectors, many elements can be mapped at one time. Elemental maps for O, Si, and Al are shown in Figures 8(a), 8(b) and 8(c). A SEM image of the same area using the same beam current is shown in Figure 8(d). From the SEM image it can be seen that substantial sample charging has occurred, but since x-rays not charged particles the x-ray map will not be affected unless the charge becomes so large that the primary electron beam is deflected.



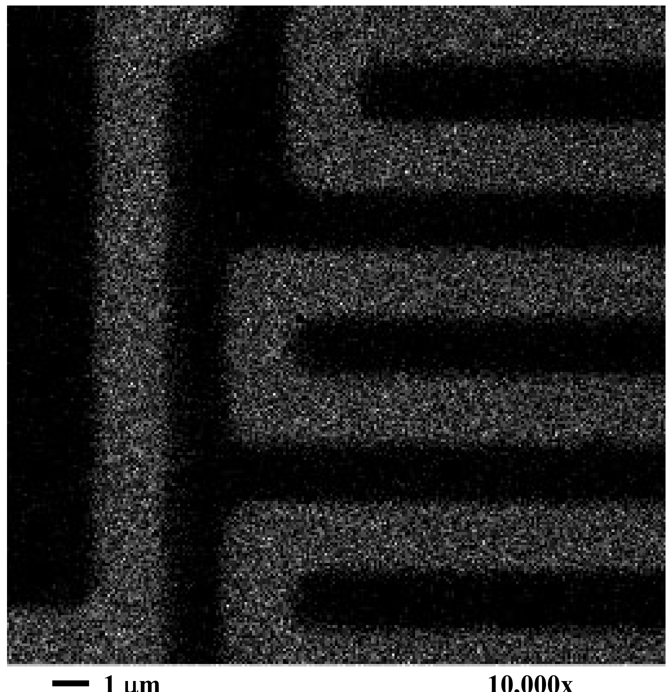
1 μm **10,000x**

Figure 8(a): O K- α elemental map. Electron beam energy is 5 keV, magnification is 10,000x.



1 μm **10,000x**

Figure 8(b): Si K- α elemental map. Electron beam energy is 5 keV, magnification is 10,000x.



1 μm **10,000x**

Figure 8(c): Al K- α elemental map. Electron beam energy is 5 keV, magnification is 10,000x.

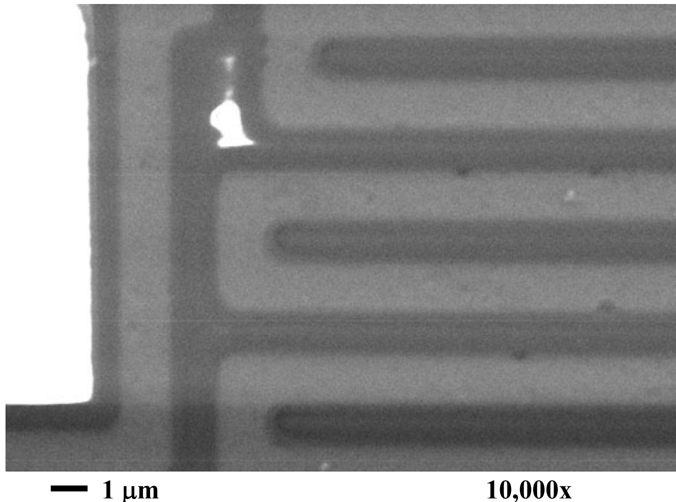


Figure 8(d) SEM image of the same area as the elemental maps.

All three elemental maps are well-defined, although clearly at much lower resolution than the SEM image. The Monte-Carlo simulation for Si K-α emission in Figure 9 suggests that the excitation volume is about 0.2 microns in diameter, which appears consistent with the observed spatial resolution of Figure 8(b). The Monte-Carlo simulation also shows that Si x-rays are created in the top 0.25 micron of the sample and most of the x-rays are able to exit the sample and reach the detector. Since the Al metal line is about 0.5 micron thick this should be adequate depth sensitivity for mapping the Al line pattern and any associated defects.

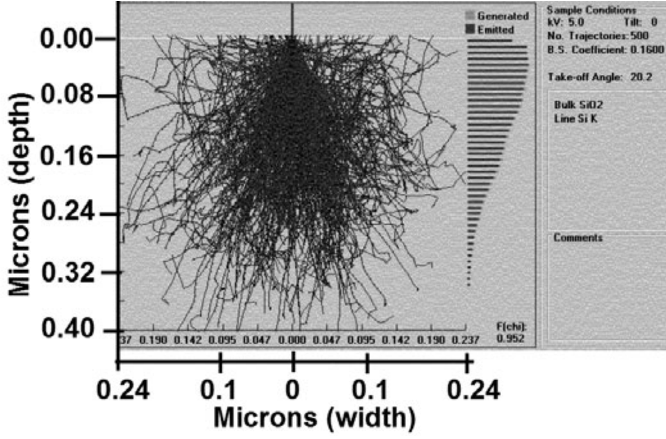


Figure 9: Monte-Carlo simulation for Si K-α x-rays in a SiO_2 matrix at 5 keV.

Now consider the same analysis conducted with 15 keV electrons. The resulting spectrum is shown in Figure 10. The increased electron beam energy has greatly increased the number of x-ray counts (15,000 peak counts for Si vs. 2,000 at 5 keV), and the signal to noise ratio is much improved. However, the Al and O peaks are much weaker relative to the Si peak than for the 5 keV case. At 15 keV the beam is penetrating more than 2 microns into the sample. As a result, the Al signal from 0.5 micron thick lines is diminished relative to the Si signal which comes from the thick dielectric layer. O x-rays are also created deep in the sample but the O K-α x-rays at 0.525 keV are very short range and can only escape from near the surface. Thus the Si K-α signal dominates the spectrum. Lateral resolution is compromised as well. Monte-Carlo simulations for O and Si in SiO_2 at 15 keV illustrate these issues, see Figures 12 and 13.

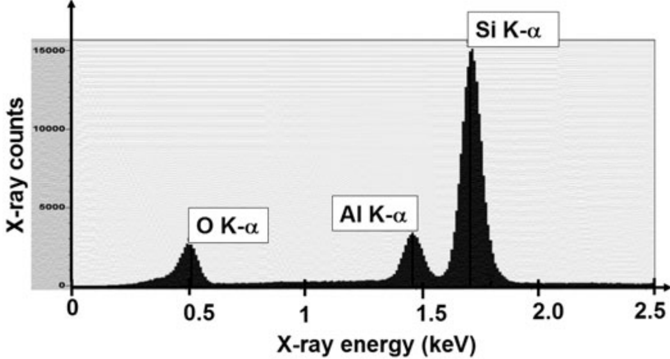


Figure 10: EDS spectrum of Al metal lines in SiO_2 dielectric at 15 keV.

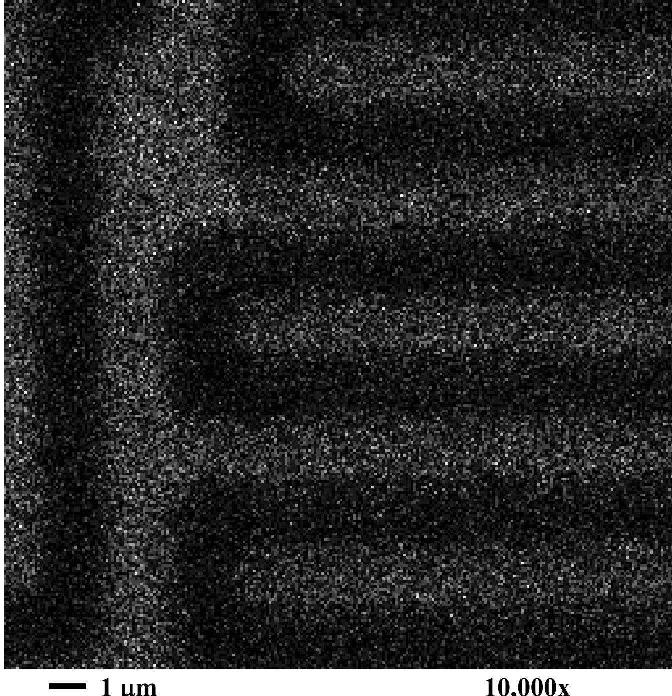
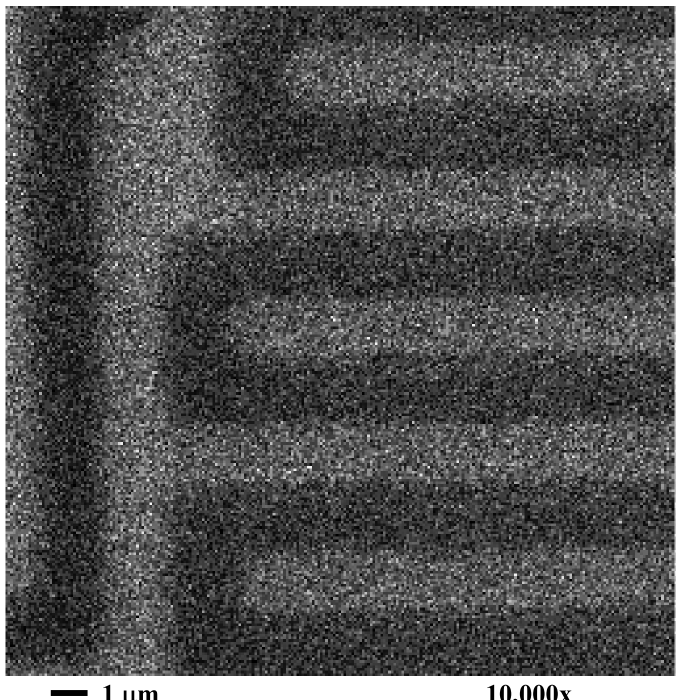
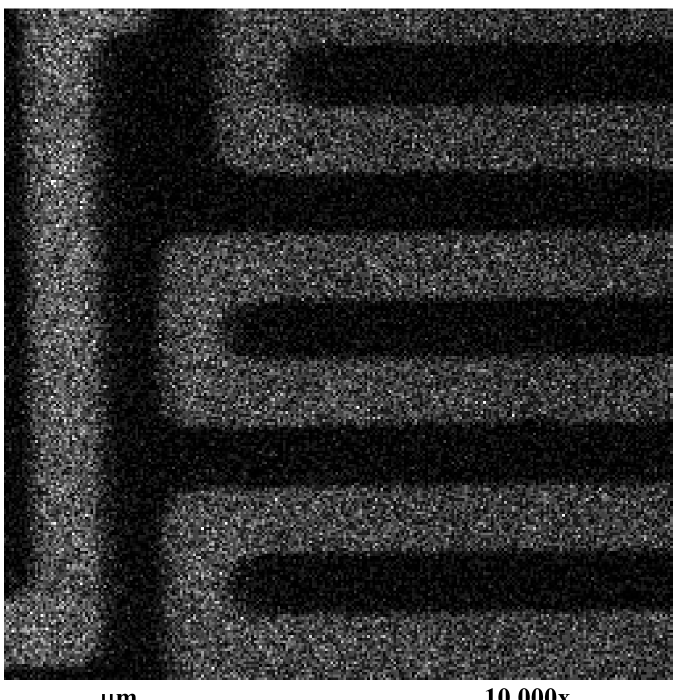


Figure 11(a): O K- α elemental map. Electron beam energy is 15 keV, magnification is 10,000x.



1 μm 10,000x

Figure 11(b): Si K- α elemental map. Electron beam energy is 15 keV, magnification is 10,000x.



μm 10,000x

Figure 11(c): Al K- α elemental map. Electron beam energy is 15 keV, magnification is 10,000x.

The 15 keV x-ray maps in Figure 11(a), 11(b) and 11(c) are generally worse than the ones at 5 keV. The oxygen map is noticeable fuzzier and has more noise. The Si K-α x-rays are created deep in the sample and can escape though the Al layer, resulting in Si x-ray counts from Al metal lines areas where the 5 keV map properly showed a gap. If the beam energy were a bit higher, we might see Al x-rays from the next lower Al metal level as well. This example illustrates the value of choosing low beam energy for EDS mapping. Furthermore, if one has the option of choosing among several x-ray lines to map a given element, the lower energy line will tend to give more surface sensitivity and better spatial resolution.

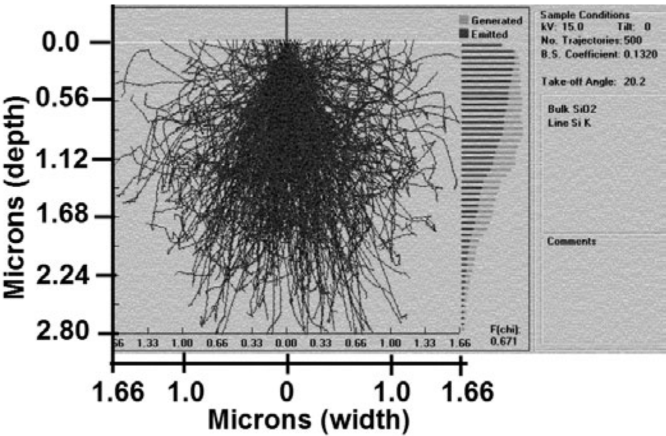


Figure 12: Monte-Carlo simulation for Si K-α x-rays in a SiO_2 matrix at 15 keV.

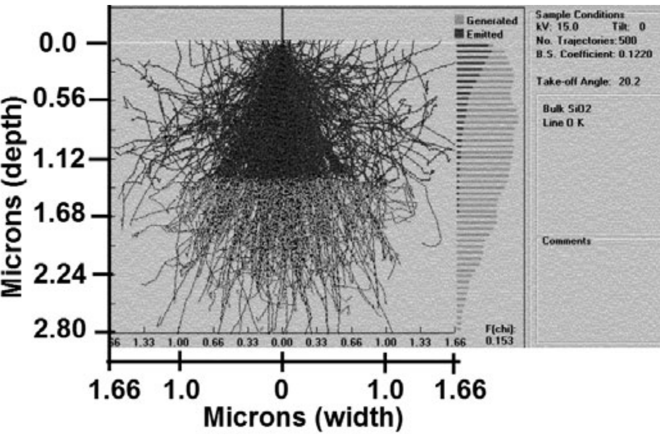


Figure 13: Monte-Carlo simulation for Si K-α x-rays in a SiO_2 matrix at 15 keV.

STEM-EDS elemental mapping

If a sample is thinned to 100 nm or less, little lateral scattering will occur and EDS mapping can be done with very high spatial resolution, see Figure 14. Thin samples are routinely created by FIB or other methods for viewing in a TEM (transmission electron microscope) or STEM (scanning transmission electron microscope.) Thin samples can also be viewed in an ordinary SEM by using a STEM detector [5] or a special sample holder [6,7]. For EDS elemental mapping of thin samples in a SEM, one could simply mount the thin sample over a narrow hole drilled in a carbon sample stub.

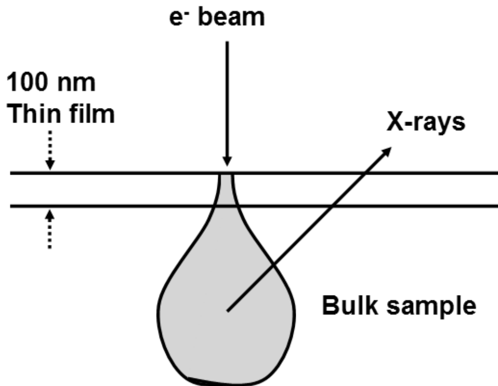


Figure 14: Comparison of x-ray emission from a bulk sample and from a thin film sample.

The small excitation volume in the thin film will not produce as many x-rays as a bulk sample, but one can compensate by increasing the beam energy. For an SEM used in STEM mode, 30 keV is usually the maximum available beam energy. Fortuitously, the characteristic x-ray production increases more rapidly with beam energy than the Bremsstrahlung background, and so the peak to background ratio improves with higher beam energy, see Figure 15.

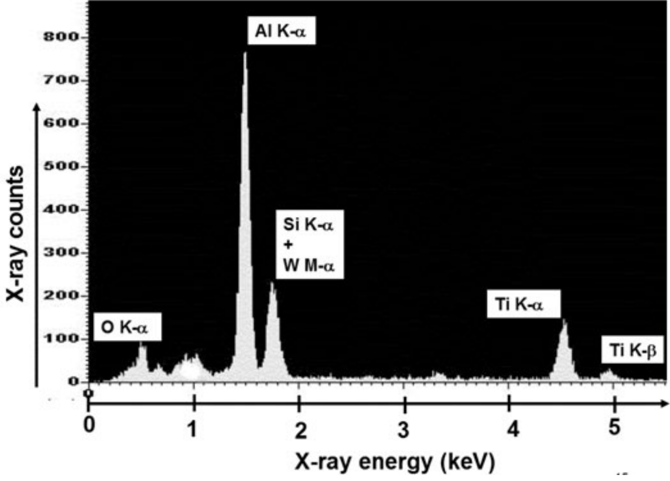
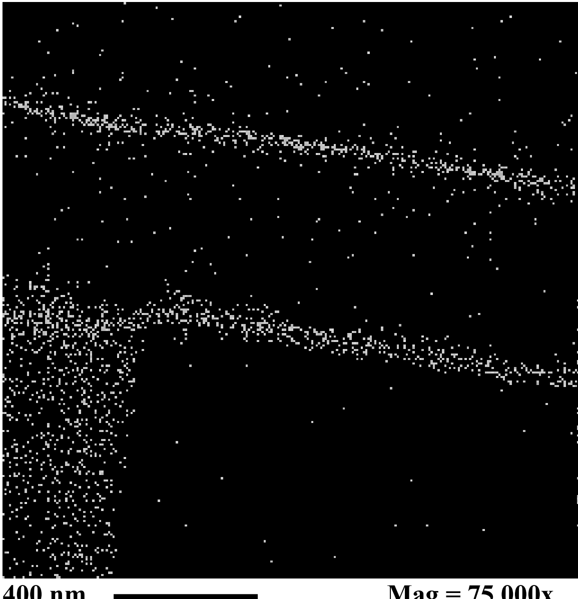


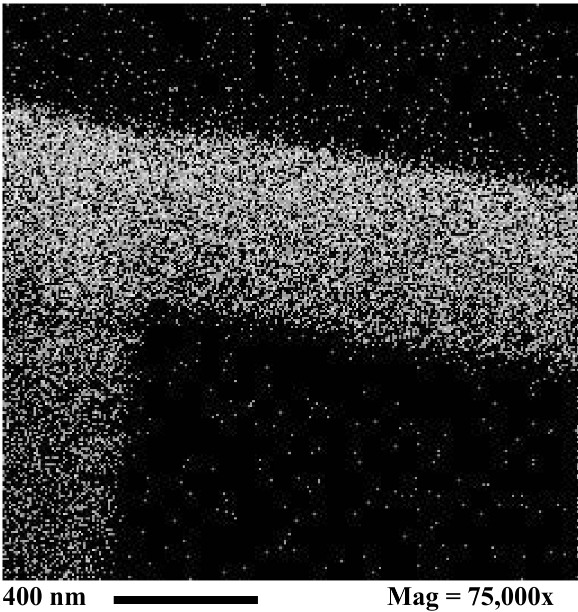
Figure 15: EDS spectrum of 100 nm thin section at 30 keV.

A 100 nm thin cross-section of an Al line with Ti barriers on top of a W plug was mounted to a STEM-in-SEM holder and analyzed in a SEM. EDS maps of Ti, Al, and W are shown in Figure 16(a) through 16(c) and a SEM taken with the same beam current is shown in Figure 16(d). Note that these maps are sharper than the bulk EDS maps shown in Figures 8 and 11, even though the magnification is much higher. The lateral spatial resolution of the EDS map on the thin section is about 10 nm compared with about 200 nm for the bulk EDS map. The use of thin sections is a powerful method for high resolution EDS mapping, even in a SEM.



400 nm **Mag = 75,000x**

Figure 16(a): Ti K-α elemental x-ray map on a 100 nm thin sample.



400 nm **Mag = 75,000x**

Figure 16(b): Al K-α elemental x-ray map on a 100 nm thin sample.

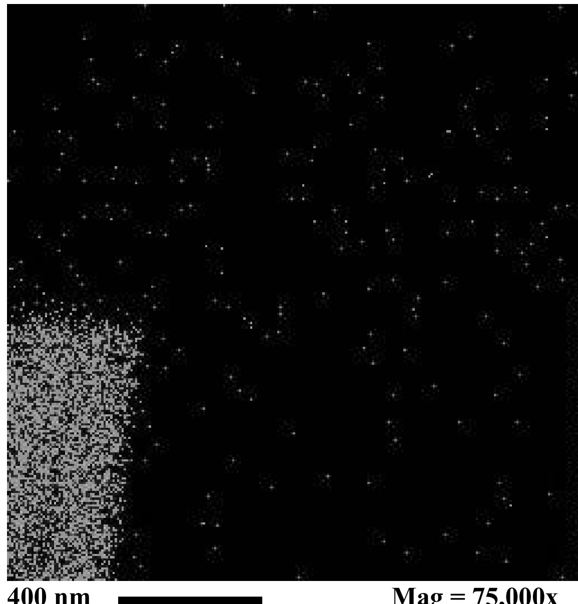


Figure 16(c): W L-α elemental x-ray map on a 100 nm thin sample.

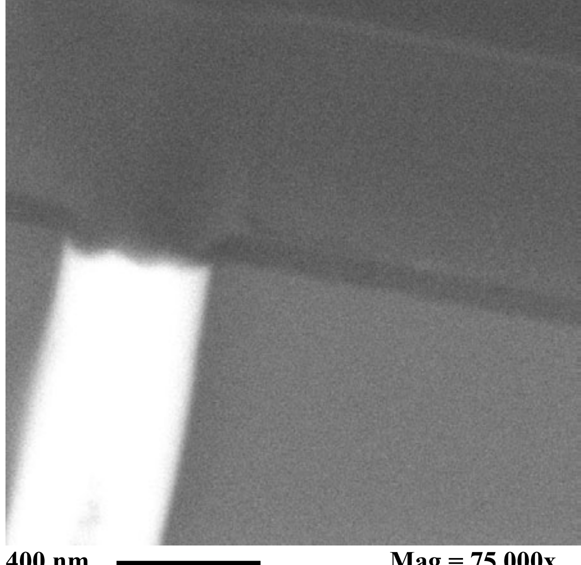


Figure 16(d): SEM image of approximately the same area as the elemental x-ray maps in Figures 16(a) to 16(c).

X-ray map artifacts

A significant background artifact in x-ray mapping is due to the fact that the energy window used for each element cannot distinguish between characteristic x-rays and Bremsstralung x-rays at the same energy. Thus, there will always be some noise in the elemental maps and some counts will appear to come from areas where a given element is not present. Furthermore, the intensity of the Bremsstralung radiation is proportional to the average atomic number in the sample. Therefore, when the beam rasters over a high atomic number material, extra Bremsstralung x-rays will be produced. The result is that areas of high atomic number materials will often show up in elemental maps for other elements. This effect can be seen in Figure 16(a) and 16(b). The W plug shown in Figure 16(c) appears faintly in the Ti and Al maps although there is no Ti or Al in the plugs. This effect can be reduced by performing background subtraction in the analysis of the characteristic peaks used for the map.

X-ray mapping will also show artifacts if the surface is not adequately smooth. Surface topography on a scale large enough (~ 1 micron) to block x-rays from getting to the detector will shadow the x-ray elemental maps, and produce contrast that is difficult to distinguish from actual elemental variation in the sample.

References

1. P. B. Simmnacher, R. Weiland, E. Langer, M. Bühler, J. Höhne, C. Hollerith, Proceedings of ISTFA 2002, pp. 87-92 (2002).

2. Joseph I. Goldstein, Dale E. Newbury, Patrick Echlin, David C. Joy, A. D. Romig, Jr., Charles E. Lyman, Charles Fiori, and Eric Lifshin, *Scanning Electron Microscopy and X-ray Microanalysis, A Textbook for Biologists, Materials Scientists, and Geologists*, 2nd Edition, Plenum, New York, 1992, p. 89

3. Available from Small World LLC, 2226 Chestertown Drive, Vienna, VA 22182.

4. Goldstein, et al., pp. 395-522.

5. B. Tracy, Proceedings of ISTFA 2002, pp. 69-76 (2002).

6. E. Coyne, Proceedings of ISTFA 2002, pp. 93-100 (2002).

7. W. E. Vanderlinde, Proceedings of ISTFA 2002, pp. 77-86 (2002).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Analysis of Submicron Defects by Auger Electron Spectroscopy (AES)

Juergen Scherer
Evans PHI

Patrick Schnabel
Charles Evans & Associates

Kenton Childs
Physical Electronics

Introduction

The competitive semiconductor marketplace, rapid development of new semiconductor designs, and shrinking design rules drive the need for continuous yield enhancement. As design rules shrink the critical defect size becomes smaller and the identification of defects becomes more challenging. For particulate contamination the National Technology Roadmap for Semiconductors[1] outlines the requirements that compositional analysis techniques need to meet in the future. These requirements are derived by relating the minimum analyzable particle size to device design rules (Table 1). More generally, it is imperative that similar requirements will also apply for non-particulate defects, such as thin residues, in the future.

The surface sensitivity and small analysis volume of Auger Electron Spectroscopy (AES) makes AES the ideal analytical technique for the compositional characterization of submicron defects. Various examples for the application of AES to the identification of defects that are difficult or impossible to characterize by Energy Dispersive X-Ray Spectroscopy (EDS) are presented. These examples include small defects, complex particles, thin residues, defects containing low atomic number (Z) elements, and buried defects. For the latter the combination of AES with Focused Ion Beam (FIB) cross sectioning has proven to be particularly powerful.

Table 1 Device design rule for memory and logic, and the corresponding requirement for minimum particle size for compositional analysis, as a function of time.

	2001	2002	2003	2004	2005	2006	2007
DRAM ½ pitch (nm)	130	115	100	90	80	70	60
MPU Physical Gate Length (nm)	65	53	45	37	32	28	25
Minimum particle size for Compositional analysis (nm)	43	35	30	24	21	20	17

Historical Developments

Auger Electron Spectroscopy is an analytical technique used to determine the elemental composition in the near-surface region of a solid material. AES is based on a process first described by Pierre Auger in 1925 to explain the radiationless relaxation of excited ions observed in cloud chambers[2]. Following a core level ionization, an atom will relax to a lower energy state through a two-electron coulombic rearrangement that leaves the atom in a doubly ionized state. The energy difference between these two states is carried away by the ejected Auger electron, which has a kinetic energy characteristic of the parent ion (Fig 1). An Auger spectrometer measures the energy distribution of the emitted electrons. By plotting the electron intensity versus the kinetic energy of emitted electrons, an Auger spectrum is obtained. Elements can be identified by the peak position of their corresponding transitions in the Auger spectrum (Fig. 2).

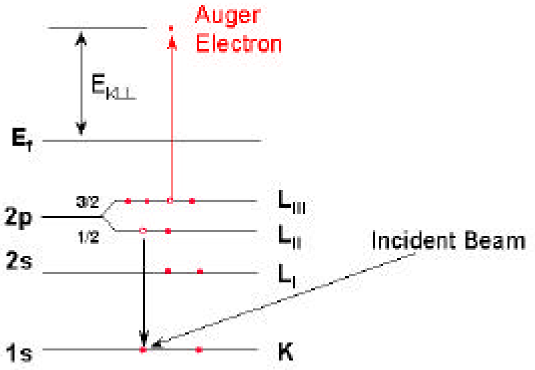


Figure 1 Schematic diagram of Auger electron emission

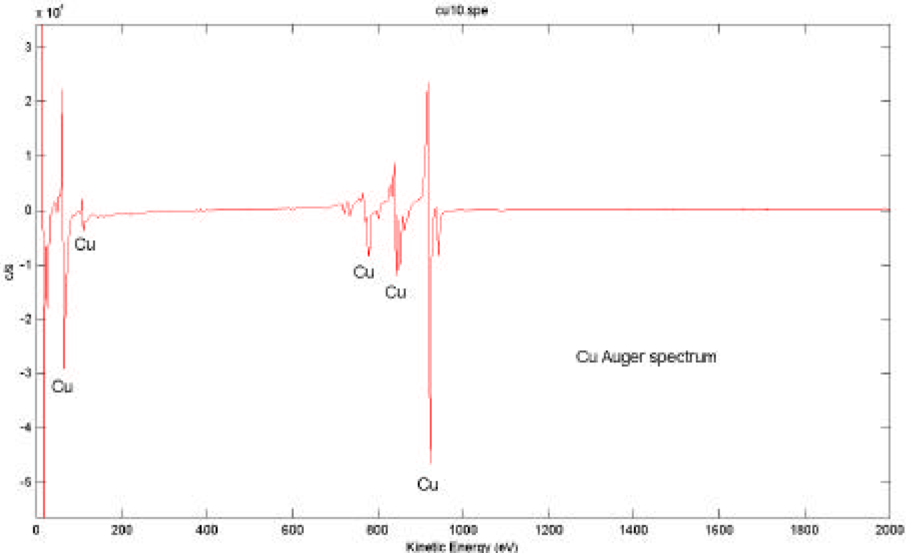


Figure 2 Auger spectrum of Cu

Every element except H and He emits Auger electrons, and can thus be identified by its Auger spectrum. In 1953 Lander proposed the use of electron bombardment in high vacuum as a means of inducing and observing Auger emission from solids[3]. The first practical equipment was a Low Energy Electron Diffraction (LEED) apparatus, reported by Palmberg and separately by Tharp and Scheibner[4]. Since the first implementation of an Auger apparatus there have been many improvements in the design of these instruments. Nowadays, Auger instruments feature high transmission or high resolution energy analyzers and multichannel detectors for higher sensitivity, as well as finely focused electron beams with beam sizes below 10nm. These high performance instruments have been around for about a decade and are found in most materials characterization laboratories. Due to their high spatial resolution and high surface sensitivity, they have become an established technique for materials characterization, failure analysis, and defect identification in the semiconductor industry. Although during recent years most developmental efforts have been spent on increasing the analytical capabilities of Auger instruments generally, there is now clearly a trend towards tailoring the instrumentation to the needs and requirements of the semiconductor manufacturing industry. Like many other predecessor analytical instruments, AES has moved from the off-line materials characterization lab to the near-line 'FabLab". This transition was enabled by a new generation of Auger instruments that have full-wafer capabilities, precision laser interferometer controlled stages, and robotic wafer handling systems in a clean room compatible enclosure. These instruments can read defect files from optical defect detection (ODD) tools and allow quick navigation to any desired defect for analysis. This enhancement in analysis speed has cost benefits in that it reduces potential process tool down time and wafer scrap.

Requirements for Submicron Defect Analysis

General Considerations

Ideally, any technique used for submicron defect analysis would provide microscopic ("visual") and compositional information about the defect. Microscopic inspection of the defect allows categorization of the defect e.g. as a particle, flake, residue or other type. Furthermore, it allows correlation of the defect location with process related information (for instance, if a given defect class is always observed at the same site). A comprehensive compositional analysis would provide the elemental composition, chemical information (i.e. oxidation states), or molecular information (i.e. if the defect consists of an organic material of a certain type). Sometimes even compositional differences within the defect might be of interest.

No single analysis technique exists that could provide all of the information listed above. Therefore it is

necessary to prioritize the kind of information sought. The dimensions of "submicron" defects constrain the analysis choices and require the focus to be on techniques providing the appropriate lateral resolution and sensitivity (due to the small amount of material present). Additionally, sample-related issues need to be taken into account, such as sample geometry, sample conductivity, and potential artifacts generated by the measurement itself (i.e. beam damage of the defect). Lastly, a high level of confidence in the reproducibility of the measurement and a short analysis time are desirable.

The possibility of narrowly focusing electron beams makes techniques utilizing electrons as a probe highly suitable for submicron defect analysis (compared, for instance, to techniques utilizing X-rays, ion or laser beams). Using a finely focused electron beam as a probe also has the advantage that the surface can be imaged by collecting the signal of secondary or backscattered electrons. However, although different electron beam based techniques such as AES and EDS are very similar with regard to their imaging capabilities, there are significant differences in the compositional information they can provide about the sample. Such differences are inherent to the physical process of signal generation. In the following section, the interaction between energetic electrons and solids and the potential information to be gained from this interaction are discussed in more detail.

Analysis Volume

AES instruments use an electron beam to create the ionization in the near-surface region that is followed by the emission of Auger electrons. However, besides Auger electrons, the electron impact also leads to the emission of secondary electrons, backscattered electrons and characteristic X-rays. Each of these secondary particles provides different information about the sample analyzed. Since all four secondary particles are generated by the same incident probing beam, it is very common to find detectors for several of them in the same instrument. Although it is more common to find either SEM/EDS or SEM/AES tools, SEM/AES/EDS tools that combine the advantages of both techniques are already commercially available.

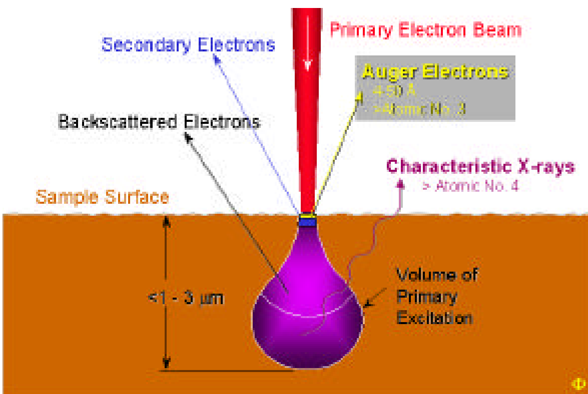


Figure 3 The interaction between an incident electron beam and a solid, showing the analysis volumes for Auger electrons, secondary electrons, backscattered electrons and X-ray fluorescence.

Fig 3 illustrates the interaction of the electron beam with a solid sample, showing the analysis volumes for Auger electrons, secondary electrons, backscattered electrons, and X-ray fluorescence. Although the primary electron beam has a very small diameter, the actual excitation volume can be very large. The size of the excitation volume is determined by scattering of the electron beam in the solid and depends strongly on beam energy and material composition. The excitation volume extends down to a depth of 1-3µm with a comparable lateral spread. Anywhere within this volume Auger electrons and characteristic X-rays can be generated. However, only Auger electrons that are generated within the first few atomic layers can leave the surface without energy loss, and are therefore available to be detected as characteristic Auger peaks. Most X-ray photons that are generated within the excitation volume can leave the solid and will be detected as characteristic X-ray peaks. As a result, the analysis volume in AES is much smaller than in EDS, and is confined to the very surface, typically a few tens of Angstroms or less.

The primary analysis volume in AES can be described as a cylinder with a diameter that is given by the beam size, to a first approximation, and a depth that is determined by the inelastic mean free path (IMFP) of the Auger electron, which ranges from ~0.4-5nm. The AES analysis depth is determined by the kinetic energy of the Auger electron and is independent of the beam energy. In EDS, on the other hand, the analysis volume is comparable to the excitation volume, which can be up to several microns in size and is strongly dependent on the beam energy. The high surface sensitivity of AES classifies it as a surface analysis technique, while EDS, due to its large analysis volume, is considered a bulk analysis technique.

This difference in analysis depth has consequences for the application of these techniques for the analysis of submicron defects. Submicron defects can be 3-dimensional particles, thin flakes, or residues, in which case the thickness represents the submicron dimension. In order to determine the usefulness of a technique for the identification of submicron defects, the ratio between the signal intensity that can be obtained from the defect and the signal intensity obtained from the surrounding area needs to be taken into consideration. Fig.4 illustrates this situation for a large (>1 µm) and a small (<0.2µm) defect.

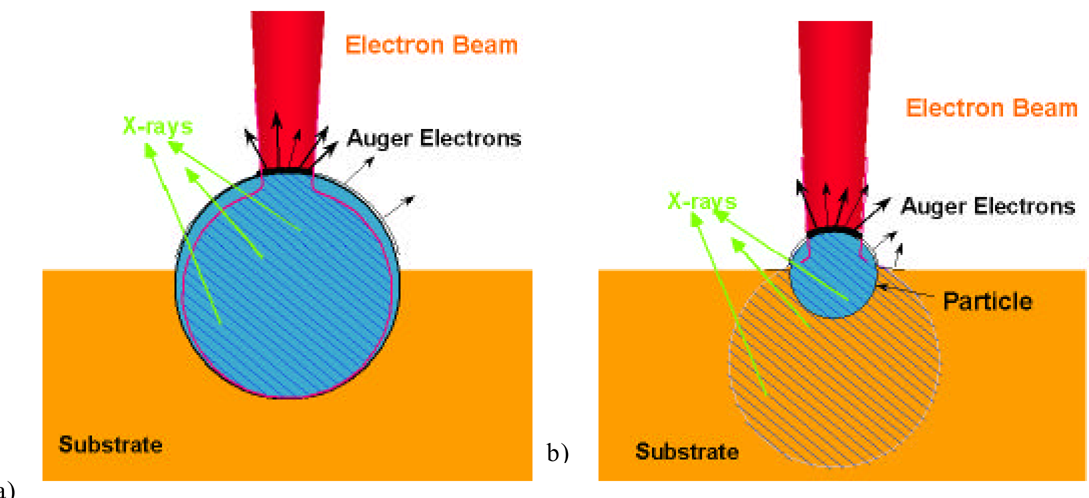


Figure 4 Analysis volumes for large (a) and small (b) particles.

In the case of large defects, the entire excitation volume is contained in the particle (Fig. 4a). Since the X-ray photons detected in EDS originate from this volume, EDS measures the overall composition of the defect, however, it would be difficult to detect contaminants on the surface of the defect, since the surface of the defect constitutes only a small fraction of the analysis volume. AES, on the other hand, analyzes the surface of the defect. If the surface composition of the defect is identical to the bulk composition, surface analysis by AES is sufficient to characterize the defect. However, if the defect is contaminated on the surface or has a "shell" of different composition, a spectrum of the surface will not yield the correct identification of the bulk of the defect. For AES to be able to analyze the bulk of a large defect, the core of the particle has to be exposed. This can be achieved by removing material by ion beam sputtering (ion milling) or using a focused ion beam (FIB) to cross-section the particles and subsequently analyze the cross-sectioned surface. For large particles, AES and EDS are truly complementary techniques. If only the overall composition of the defect is of interest, EDS is the easiest and fastest way to analyze the defect. However, if more information about the lateral and in-depth distribution of the constituents is required, Auger can provide additional insight into the makeup of the particle.

With decreasing particle size, the excitation volume moves into the substrate (Fig. 4b). At particle sizes of about 0.2µm, most of the EDS signal originates from the substrate, which makes it increasingly difficult to identify the composition of the defect. However, the primary AES analysis volume is still located in the surface of the defect. In order to meet the requirements of the Technology Roadmaps for the minimum particle size for compositional analysis, an analysis technique has to be able to analyze defects much smaller than 0.2µm. According to the ITRS, the minimum particle size for compositional analysis for 2002 is 35nm.

As the defect size decreases below 0.2µm the contribution of EDS signal from the particle shrinks as well, becoming negligible for very small or thin defects. On the other hand, the primary AES signal from the particle remains the same as long as the analytical beam size is smaller than the particle size, which is typically the case. Although the primary AES analysis volume is completely contained in the surface of the defect, even down to a size of 35nm, the AES spectrum of a small defect also contains a substrate signal, which is related to scattering of the electron beam (Fig 5). Primary electrons can be scattered out of the particle and onto the substrate (lateral scattering), where they can excite Auger electrons. Furthermore, electrons can be scattered back and excite Auger electrons on their way out of the solid (backscattering).

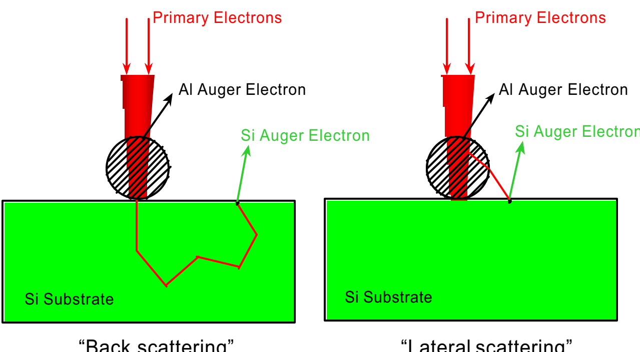


Figure 5 Lateral and backscattering (i.e. Al particle)

Both lateral scattering and backscattering contribute to the Auger signal. The amount of this contribution depends on beam energy and particle size[5]. In the case of large particles, backscattering mainly takes place within the particle, thus contributing to the overall signal intensity originating from the surface of the defect. Lateral scattering, however, leads to the emission of Auger electrons from the surrounding surface. A lower beam energy may be used to minimize this contribution for certain sample geometries. This approach is effective for large particles, in which the scattering is contained within the particle (~1 micron or greater). It is also effective for planar structures, where the particle sits within the substrate rather than on top of it, and for which there is no lateral scattering out of the sidewall of the particle.

Lateral scattering and backscattering result in a greater substrate contribution to the Auger signal as the size of the defect is reduced. Since the primary AES analysis volume does not change as a function of particle size, variations in the ratio of signal intensity from the defect to signal intensity from the substrate are mainly determined by the amount of lateral scattering and backscattering. Very small particles sitting on top of the substrate are typically analyzed at high beam energies of 20keV or higher to minimize the lateral scattering. Since high energy electrons are barely deflected as they penetrate a very small particle, the amount of lateral scattering is greatly reduced and the scattering contribution is mainly determined by backscattering. As the particle size shrinks from 200nm down to 35nm, the AES signal from the primary analysis volume remains nearly the same, and is largely confined to the particle. However, the backscatter contribution includes an increasing amount of substrate area (and less particle area) within the backscatter radius. Thus, the decrease in defect intensity is fairly moderate. As a result, the identification of a 35nm particle is almost as straightforward as the identification of a defect that is 10 times bigger in size.

Compositional Identification

So far, the capabilities of AES for the analysis of submicron defects have been discussed independent of the composition of the defect. In terms of detectability, the elemental sensitivity of a technique has to be taken into consideration. Since most elements have several Auger transitions, the transition with the highest sensitivity can be selected for analysis. As a result, for the majority of elements, the sensitivity varies by less than a factor of ten across the periodic table. This uniform sensitivity enables the detection of light (low Z) as well as heavy (high Z) elements with comparable sensitivity. EDS, on the other hand, is less sensitive to lighter elements. Carbon (C), for example, is a fairly common component of particulate or residual contamination, but it is difficult to detect by EDS. This becomes even more of a problem when the particle is small compared to the EDS analysis volume. Auger however, has a very good sensitivity for light elements, especially at low beam energies.

Chemical Information

Although AES is generally considered a technique for elemental rather than chemical analysis, there are a few cases where chemical information can easily be obtained from the Auger spectrum. An Auger transition is a three-electron process. If the binding energies of the electrons involved in the process change due to a change in the chemical environment surrounding the atom, this change may be reflected by a shift in energy of the Auger peak. This "chemical shift" is usually fairly subtle, but in some elements that are used in semiconductor processing such as aluminum (Al) and silicon (Si), the differences in peak energy between the oxide and the "pure" metal are very easy to detect.

An example for a peak shape change is provided by silicon carbide, (SiC). Fig. 6 shows the comparison of the C peak in SiC, compared to the C peak for sp^2-bonded C (i.e. graphite). At first glance, the spectrum generated by a SiC particle that is sitting on top of a Si surface looks very similar to the spectrum generated by an "organic" (hydrocarbon) particle sitting on the same surface, since the latter shows a Si peak due to scattering of the electron beam. However, due to the difference in peak shapes, the analyst can distinguish a SiC particle on a Si surface from a generic hydrocarbon particle.

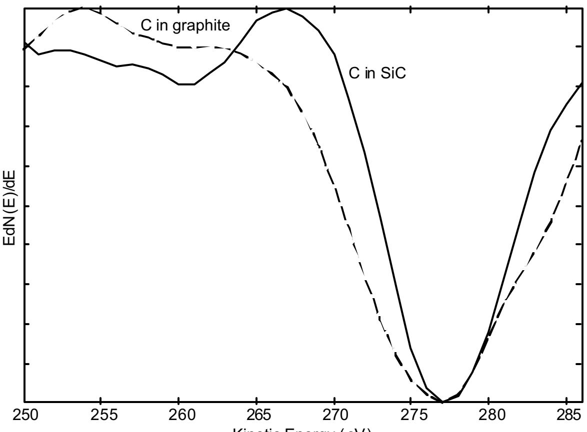
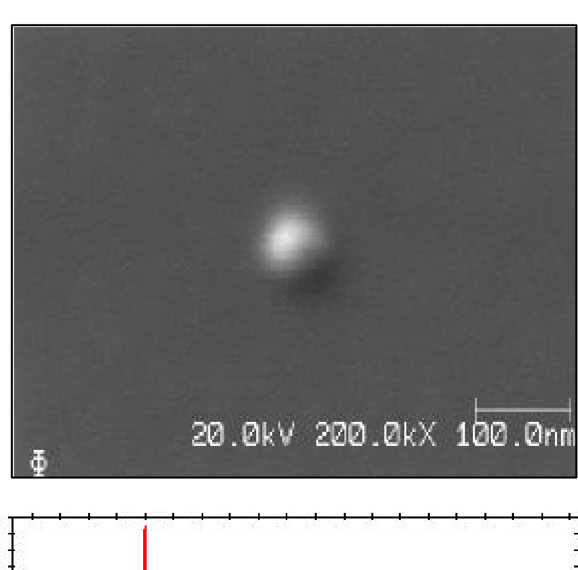


Figure 6 Comparison of C KLL peak for graphite and SiC.

Applications

Small Particles

Fig.7 shows the SEM image and Auger spectrum of a 50nm oxidized Al particle on a Si surface. Despite the small size of the defect, a strong Auger signal of the particle can be obtained. The Si peaks in the spectrum are due to backscattering from the Si substrate. The presence of C is related to air exposure. It also needs to be pointed out that defects on wafers are usually analyzed on a Defect Review Tool (DRT), which requires data from an Optical Defect Detection Tool (ODD) in order to locate a defect on a wafer. Therefore, the ODD has to be able to locate a defect before an analytical technique can identify it. Since Auger can analyze defects that are smaller than the current ODD detection limit, AES is a technique capable of compositional identification of defects that are still too small to be detected in optical scanners[6].



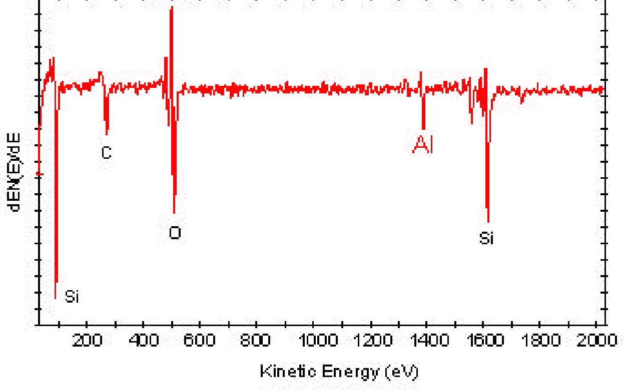
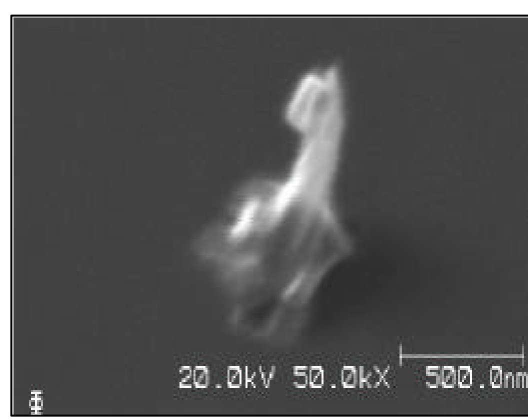


Figure 7 SEM image and Auger spectrum of a 50nm Al oxide particle.

Complex Defects

Fig. 8 shows the image of a large defect that was detected on a monitor wafer that was transferred through an idle tungsten (W) etch back tool.[7] The W etch back process uses a sulfur hexafluoride (SF_6) plasma treatment to remove a blanket W film with a thin titanium nitride (TiN) adhesion layer beneath it. AES survey spectra acquired at two different locations on the defect reveal that the particle is composed of physically-segregated phases of Ti and Al. It was concluded that the particle originated from the interaction between an etch by-product of the TiN adhesion layer and the process chamber hardware. Low level Ti contamination was found over the entire surface of the wafer. The lack of Ti on the Al particle indicates that the low level Ti contamination occurred prior to the particle falling onto the wafer. This suggests a gaseous source of Ti in the process chamber from previous etch runs. These detailed conclusions were not available from EDS analysis.



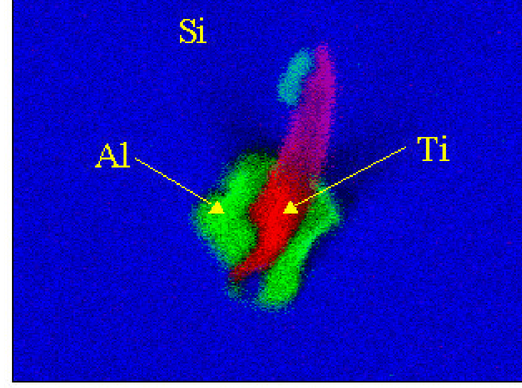


Figure 8a SEM image and Auger elemental map of a complex particle detected after W etch back process. Al: green, Ti: red, Si: blue

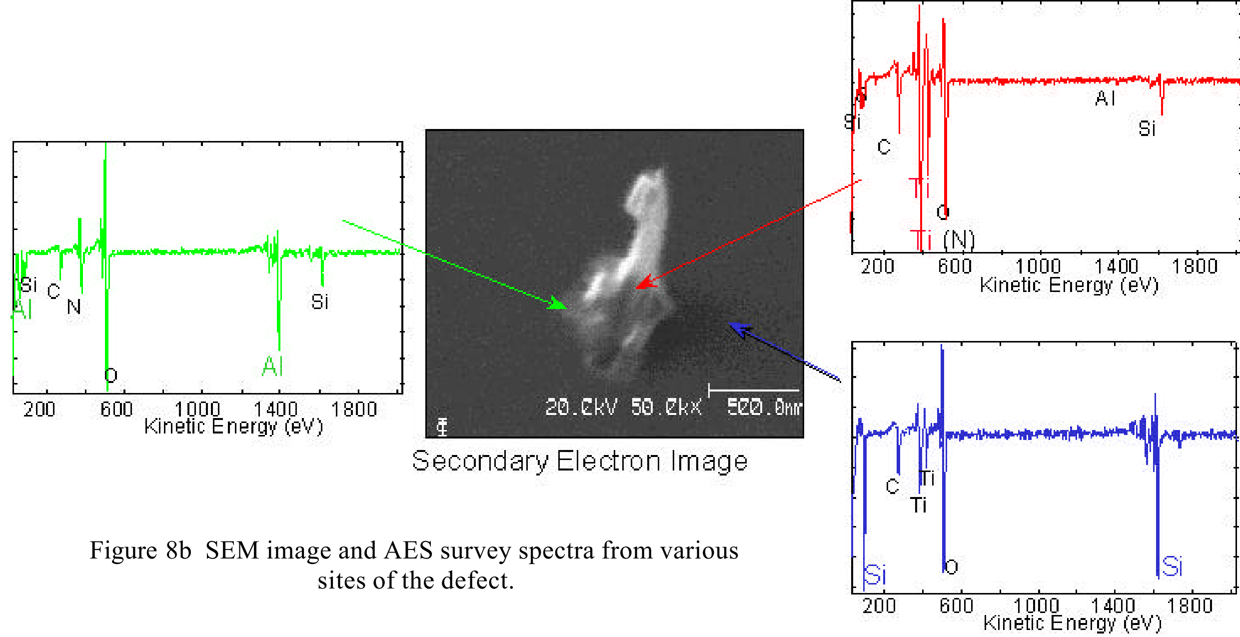


Figure 8b SEM image and AES survey spectra from various sites of the defect.

Low Z defects

Fig. 9 shows an example of a defect detected after poly-Si etch and photoresist strip. The defect is approximately 200nm in size and was located on top of a poly-Si line. The AES spectrum that was acquired on the defect reveals the presence of C, sulfur (S) and oxygen (O), while only Si and O (from "native" silicon oxide) were detected off the defect. Auger maps for C and S show that both elements were present only in the defect. The presence of C, S and O is a general "fingerprint" signature for photoresist residue, indicating an incomplete strip and clean.

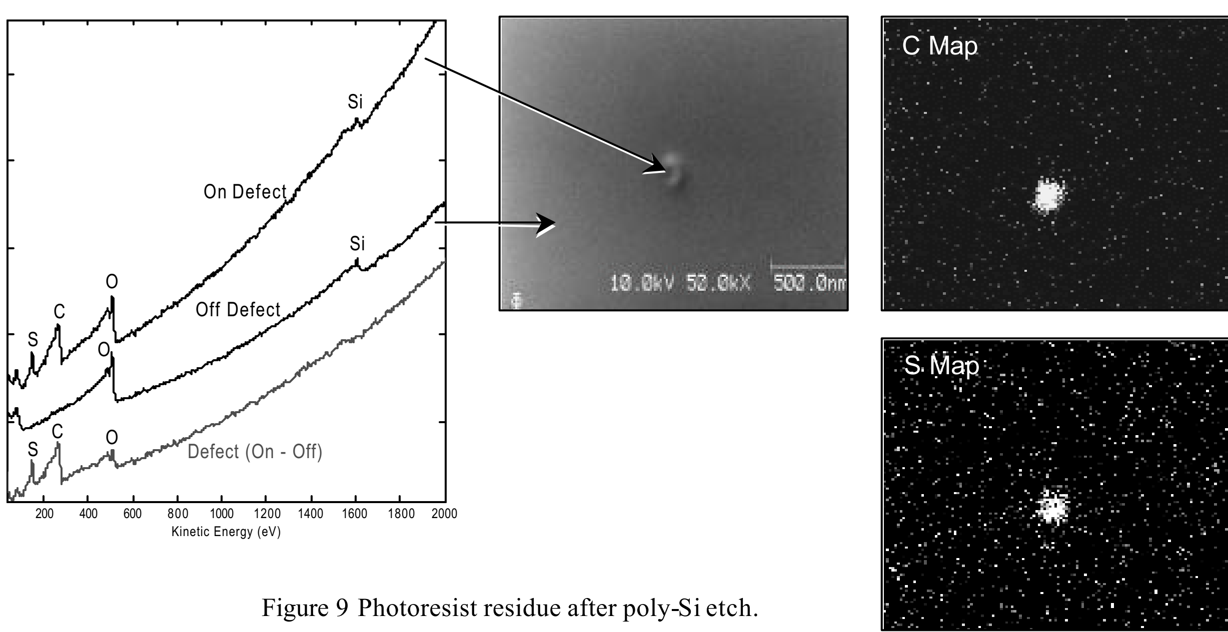


Figure 9 Photoresist residue after poly-Si etch.

Thin Flakes and Residues

Thin flakes and residues are other types of submicron defects. Since these kinds of defects are very often large in area but thin, the depth of the analysis volume is more important than its lateral extent. Flakes can be as thin as a few hundred Å and residues can be as thin as a monolayer. A technique like EDS that has a very large sampling depth is literally looking "through" the defect and will not be able to detect it if the defect is sufficiently thin. With an analysis depth of 1-10 atomic layers, AES can detect even a sub-monolayer coverage, which makes it an ideal tool for the identification of thin defects and residues.

Fig. 10 shows the SEM image of a thin flake that is sitting on top of TiN capped Al lines on SiO_2. The defect is semi-transparent in the SEM image, which indicates that it is very thin. The AES spectrum that was acquired on the defect shows Al, O and some fluorine (F), while the spectrum that was obtained on the metal line shows Ti and nitrogen (N) (from the TiN cap). The Al Auger map confirmed that the entire flake consisted of Al. The spectrum that was acquired on the defect shows neither TiN nor SiO_2, which again demonstrates the high surface sensitivity of this technique. The presence of F on the defect suggests that it could be a flake from the etch chamber, in which Al reacted with F and fell onto the wafer after patterning. The presence of such defects may suggest the need to clean the etch chamber. This defect could result in an electrical short and a circuit failure. Since the primary beam penetrates to the SiO_2, EDS would have detected all elements present at every analysis location. Thus, it would have been very difficult to distinguish the Al in the flake from the Al in the metal line, and results would not have been conclusive.

Another example of thin defects is shown in Figure 11. Particle adders from an etch tool were collected with a Si monitor wafer. Subsequent optical inspection located 105 defects. Ninety percent of the defects were "flower" defects, which have a central particle, surrounded by a thin film contamination.

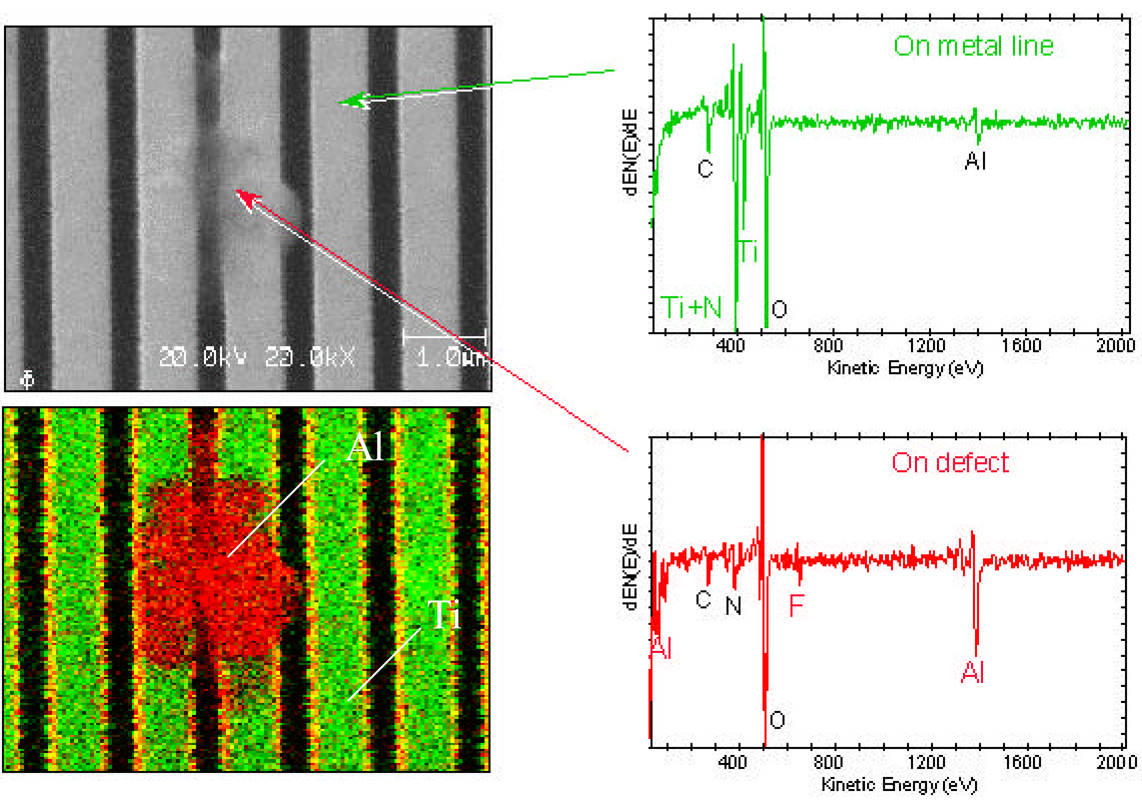
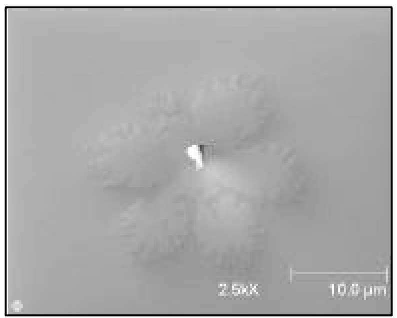
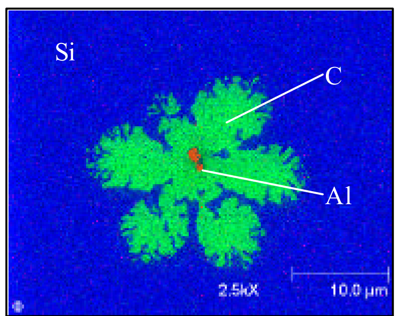


Figure 10 SEM image, survey spectra and elemental map of thin Al flake on metal line.





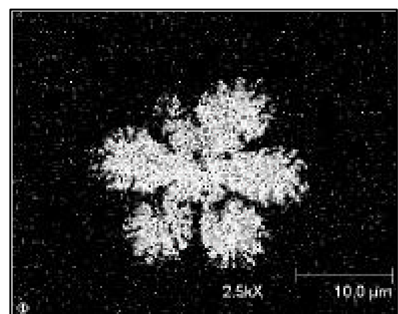


Figure 11 SEM image and elemental maps of flower defects with Al center particle.

Auger analysis showed that most of the central particles in flower defects were Al and F, while the remainder consisted of Si and nickel-iron-chromium (Ni-Fe-Cr) (from stainless steel). Auger analysis also showed that the thin petal portion of the flower defects is composed of C and F, indicating a thin fluorocarbon residue. The contamination was traced to condensed fluorocarbon, originating from pump oil in the etch tool wafer pre-pump chamber. The central particles may act as an initiation site for condensation. In order to determine the thickness of the fluorocarbon petals, an Auger sputter depth profile through the CF petal was measured. In this measurement, the surface was eroded by an ion beam, while the surface composition was simultaneously monitored by AES. Plotting the AES intensities as a function of sputter time yields a representation of the depth distribution of the elements measured (Fig.12).

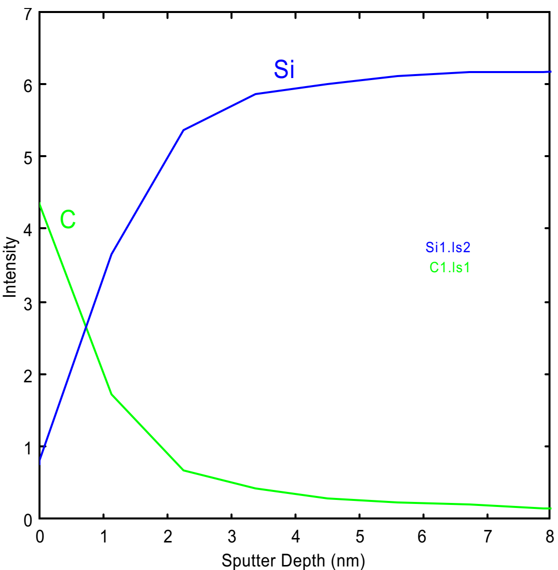


Figure 12 Depth profile of petal.

The depth profile reveals that the fluorocarbon petal was only ~1-2nm thick. EDS measurements were not able to analyze the composition of the petals (Fig.13)

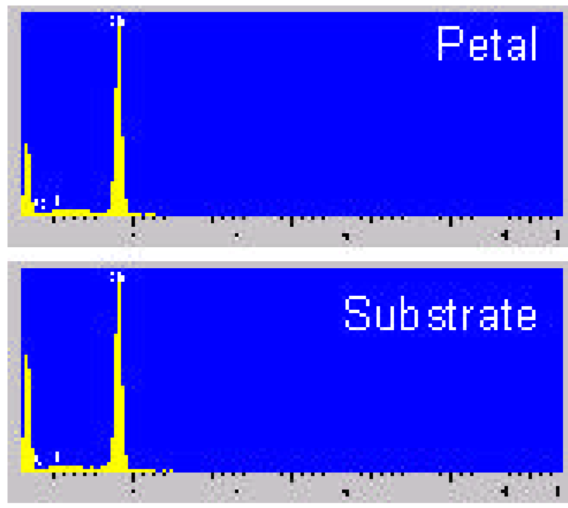


Figure 13. EDS spectrum of petal.

Buried Defects

Defects are often not discovered immediately after they occur but only after subsequent processing of the wafer. Therefore, an analytical technique for characterizing submicron defects has to be able to analyze defects that are located on the surface *and* defects that are buried under a coating or inside of a complicated patterned structure. As already discussed in the case of large surface defects, AES as a surface analysis technique can only analyze the very surface of a sample. For the technique to be able to analyze defects that are buried under a coating or incorporated within a device structure, the defect has to be exposed either by ion milling or by FIB cross-sectioning. Figure 14 shows the SEM image and EDS spectrum of a particle on a poly-Si line that was buried beneath Si nitride. The EDS spectrum shows only Si and N, both on and off the defect.

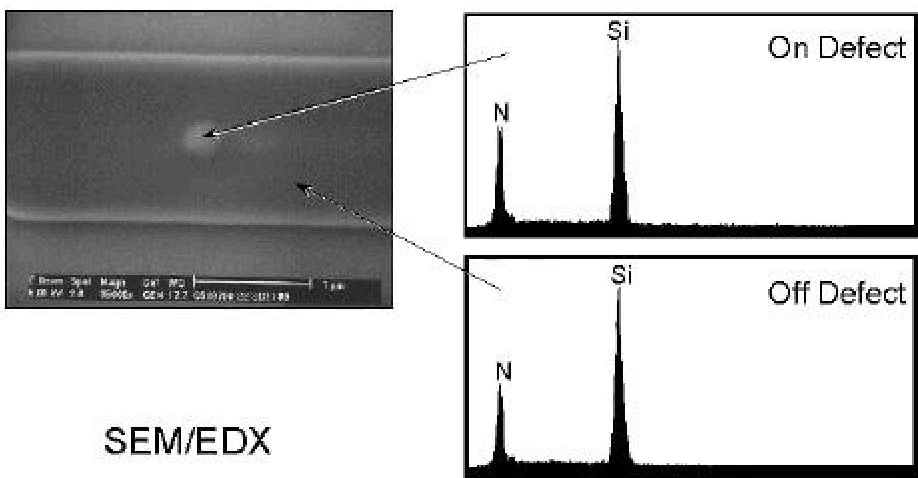


Figure 14 SEM image and EDS spectrum of a buried particle on poly-Si line.

In order to analyze this particle by AES, the Si nitride overlayer was removed by ion milling (Fig. 15). AES survey spectra that were acquired on the defect revealed that the buried particle consisted of silicon oxynitride (SiON), while only Si was detected off the defect (poly-Si line).

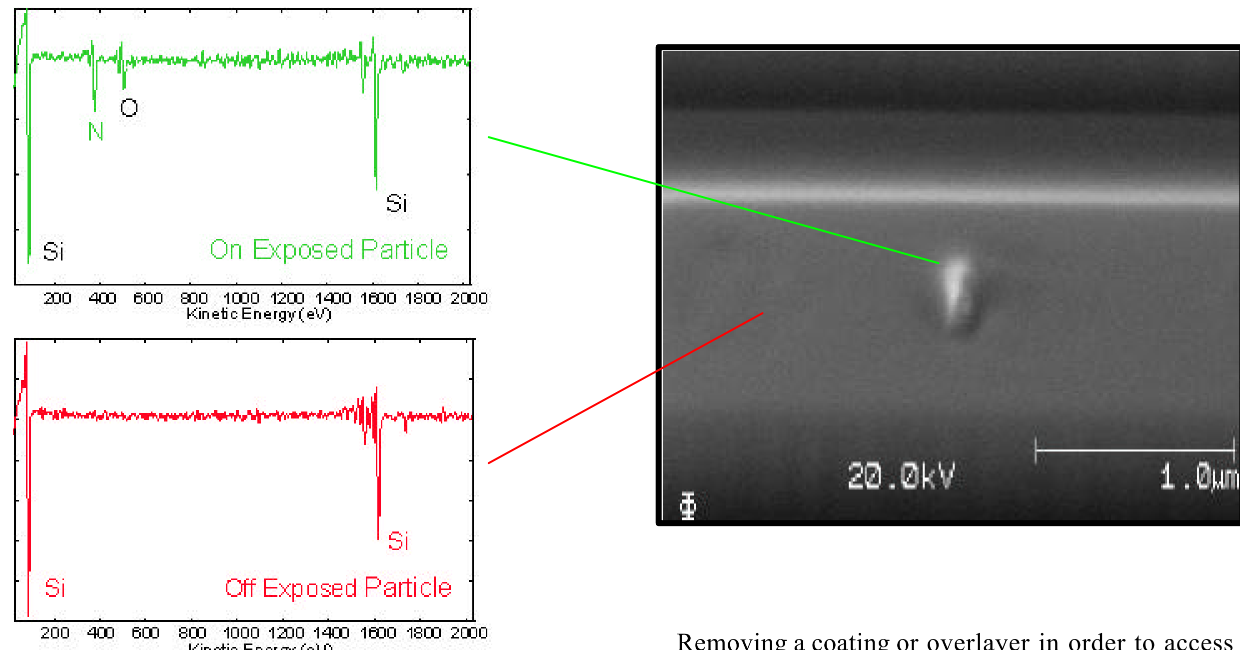


Figure15 SEM image and AES survey spectra of buried particle after removal of Si nitride.

Removing a coating or overlayer in order to access a buried defect is a reasonable approach as long as the coating is relatively thin and the structure of the pattern is simple. If the coating is too thick, ion milling would take up a considerable amount of time. Furthermore, after prolonged ion milling, sputtering artifacts can develop. These artifacts include non-uniform removal of material due to different grain orientation, shadowing during sputtering, etc. Especially in complex patterned structures, data interpretation can become very difficult. In these

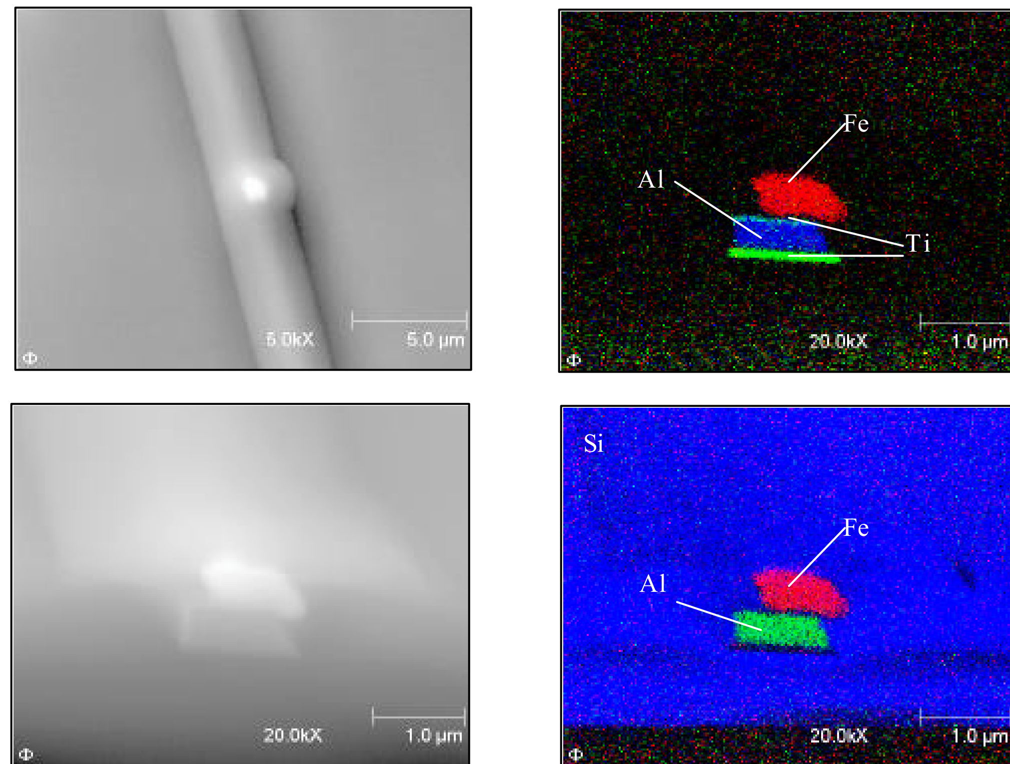


Figure 16 SEM images before and after cross-sectioning and elemental maps of buried stainless steel particle after cross-sectioning with FIB.

cases it is recommended to expose the defect by cross-sectioning it with a focused ion beam (FIB). Figure 16 shows the SEM image and elemental maps of a cross-sectioned defect. AES survey spectra showed Fe and Cr, indicating that the defect is a stainless steel particle. Auger maps for Fe, Ti, Al and O show that the particle is sitting on the TiN cap of the metal 1 Al line and is surrounded by Si oxide. The defect hangs over the edge of the metal line, indicating that it arrived after M1 etch. The M1 line and the particle were subsequently covered by the dielectric deposition. The source of these types of defects may be the plumbing or wafer transfer mechanism used after the metal etch process and before the dielectric deposition.

Summary

The underlying physics of Auger Electron Spectroscopy (AES) offer two distinct advantages, making it a superior choice for the analysis of sub-micron defects over Energy Dispersive X-Ray Spectroscopy (EDS). These two advantages are its very small analysis volume and its higher sensitivity for light elements. The small analysis volume ensures that the ratio of signal generated by the defect to signals originating from the surrounding materials is favorable even down to defect sizes of 30nm and less. In EDS, the signal from a sub-micron defect may be negligible relative to the much larger "spectral background". This is especially true for very thin defects. Additionally, the large analysis volume of EDS makes measurements of defects on patterned wafers particularly difficult.

The higher sensitivity for light elements ensures that no elements will be missed by Auger (with the exception of hydrogen or helium). Modern instruments fully utilize these advantages, pushing the limits for an elemental characterization of defects well beyond what can be done by EDS. For larger defects, which could alternatively still be analyzed by EDS, AES provides the option to obtain more detailed insight for instance by investigating compositional differences at the surface and within a defect or by mapping the lateral distribution of elements across the defect. The combination of AES with Focused Ion Beam (FIB) allows for exposure of buried defects for analysis and has proven to be a particularly powerful combination. FIB/AES can also

be used to cross-section defects to investigate the interior of a defect.

In recent years, AES has continuously been tailored for the specific needs of the semiconductor industry, and moved from the off-line materials characterization lab to the near-line "FabLab". The new generation of AES "tools" has full wafer capabilities, precision laser interferometer controlled stages, and robotic wafer handling systems in a clean room compatible enclosure. In addition, these instruments can read defect files from optical defect detection systems (ODDs) to quickly locate and analyze defects. AES is positioned extremely well to meet future requirements for elemental characterization of sub-micron defects, as outlined in the semiconductor roadmap.

References

1. International Technology Roadmap for Semiconductors – 2001 Edition, International SEMATECH, http://public.itrs.net/Files/2001ITRS/Home.htm, (2001).
2. Pierre Auger, J. Physique Radium **6**, 205 (1925).
3. J.J. Lander, Phys. Rev. **91**, 1382 (1953).
4. L.N. Tharp and E.J. Scheibner, J. Appl. Phys. **38**, 13320 (1967)
5. Kenton D. Childs, David Narum, Lori A. LaVanier, Patricia M. Lindley, Bruno W. Schueler, George Mulholland, Alain Diebold, "Comparison of submicron particle analysis by Auger electron spectroscopy, time-of-flight secondary ion mass spectrometry, and secondary electron microscopy with energy dispersive X-ray spectroscopy," *J. Vac. Sci. Technol.* A **14**(4), pp. 2392-2404, (1996).
6. Seong-Ho Yoo, James Weygand, Juergen Scherer, Lawrence Davis, Benjamin Liu, Kurt Christenson, Jeffery Butterbaugh, Natraj Narayanswami, "Identification and sizing of particle defects in semiconductor-wafer processing," *J. Vac. Sci. Technol.* B **19**(2), pp. 344-353, (2001).
7. Y. Uritsky, L. Chen, S. Zhang, S. Wilson, A. Mak, C. R. Brundle, "Root cause determination of particle contamination in the tungsten etch back process," J. Vac. Sci. Technol. A **15**(3), pp. 1319-1327 (1997).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

SIMS Solutions for Next Generation IC Processes and Devices

Gary Mount
Charles Evans and Associates, EAG, Sunnyvale, California, USA

Yung Liou and Han-Chung Chien
Evans Taiwan, EAG, Hsinchu, Taiwan

Introduction

Secondary Ion Mass Spectrometry (SIMS) can be a powerful tool for trace element failure analysis investigations. Dynamic SIMS is a depth profiling technique that features parts-per-billion (ppb) detection limits for many elements and at least parts-per-million (ppm) detection limits for every element in the periodic table including hydrogen. Dynamic SIMS can be used to detect contamination distributed throughout films or at interfaces. Dopant profiling is a common application and stoichiometry can also be measured for some materials systems.

Dynamic SIMS cannot directly make measurements on deep sub-micron devices. Lateral analysis dimensions are typically 800 μm^2 to 100 μm^2. With special effort, 10 μm^2 areas can be depth profiled, a much larger area than that of a single sub-micron device. Special test structures are often incorporated into wafer patterns in order to accommodate SIMS measurements. Shrinking lateral dimensions have been accompanied by ultra-thin layer structures. Recent advances in SIMS depth profiling have made it possible to resolve structures as thin as 2 nm thick.

With excellent detection sensitivity for all elements, SIMS can be used to detect contaminants that may be interfering with device performance, dopants that may not have intended concentrations, or dopants that may have diffused into undesirable locations. Layer thickness can be measured directly because SIMS is a depth profiling technique. Stoichiometry can be measured for some materials systems and compared with expectations.

Failure analysis using SIMS often takes the form of a “good vs. bad” study. SIMS, with its tremendous sensitivity, will frequently find contamination through films and at interfaces in both “good” and “bad” devices. Some contaminants may be irrelevant to device failure. Dopant levels can be measured, but cannot independently determine if the dopant level causing a problem. The “good vs. bad” comparison is necessary in order to determine the significance of the findings.

A successful outcome to SIMS failure analysis experiments is a product of good experimental design and planning. In this article, we outline the considerations that are part of the experiment planning process so that the chances of solving problems can be maximized.

SIMS Basics

In a SIMS analytical system, ions are extracted from an ion source, and then accelerated and focused down a column in order to form an ion beam. The ion beam is then directed onto a sample surface in a square raster pattern that results in sputter removal of the sample surface. During the sputtering process, some of the sample material is ionized. This ionized sample material is accelerated by an electric field away from the sample forming a second (or secondary) ion beam. The secondary ion beam is passed through a mass spectrometer that identifies the mass of the ionized particles. The mass filtered secondary ion beam is then directed to a detector that measures the intensity of the secondary beam. By careful calibration to standards, the secondary ion intensities are quantified.

Secondary Ion Generation

The SIMS process of sputtering a sample surface is a very inefficient generator of secondary ions. Only a few percent of the sputtered sample material is ionized. Ionization probabilities can be increased through the use of reactive primary ion beams. An oxygen primary ion beam can be used and will increase the probability of positive secondary ion formation. Alternatively, a cesium

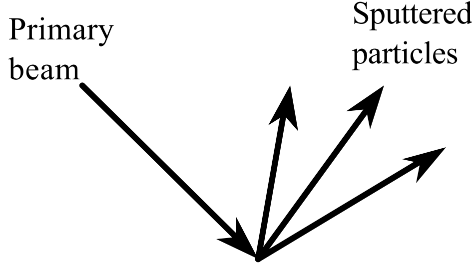


Figure 1: A cross-section of a sample showing the SIMS sputtering process and crater formation.

primary ion beam will increase the probability of negative secondary ion formation. The choice of primary beam is therefore often determined by the elements that are to be analyzed.

Vacuum Compatibility

SIMS primary and secondary ion beams need to travel through vacuum in order to have a mean free path that is long enough to reach the intended destinations. Therefore samples need to be vacuum compatible.

Depth Profiles

The SIMS sputtering process etches a crater into the sample surface (Figure 1). Secondary ions are collected continuously as the primary ion beam continues to erode the sample surface. An intensity versus time depth profile is the result. Layers and contaminants at interfaces can be seen (Figure 2).

Dynamic and Static SIMS

The terms "static" and "dynamic" SIMS are relatively well known. The names suggest that these are two different SIMS techniques. They are not. It really refers to the damage that is done to the sample surface. If the primary ion beam is very low current, then only a fraction of the atoms and molecules on the sample surface will be disturbed. The "static SIMS limit" is defined as less than 1 in 100 surface particles are disturbed by the primary ion beam. Below the static SIMS limit, the elements and molecules that are sputter etched are representative of the elements and molecules that are on the sample surface. Static SIMS is an effective tool for the examination of surface contamination.

Dynamic SIMS occurs when the primary beam disturbs greater than 1 in 100 of the surface particles. Usually the goal is to sputter etch the surface causing erosion with the result being a depth profile. Sputter etching is a very violent process on an atomic scale. Surface elements are mixed into the substrate. Molecules are broken apart. Mixing of the substrate material can cause new molecules to form. This breaking of existing molecules into their component parts and the formation of new molecules from atomic species

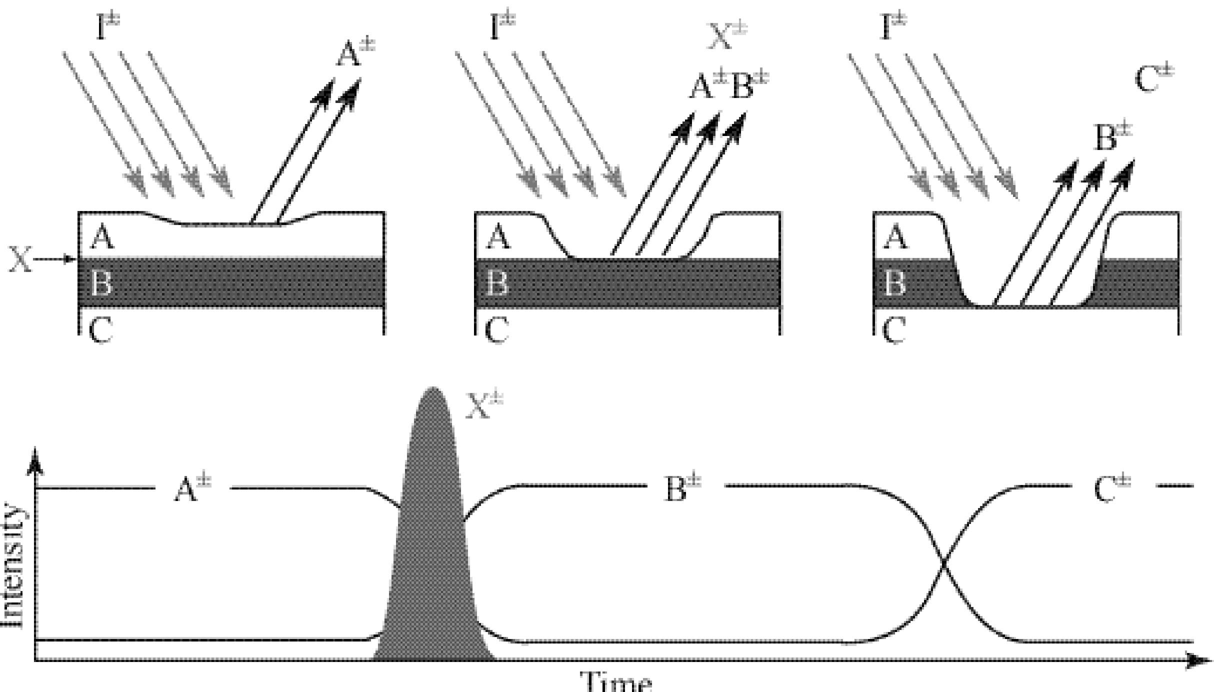


Figure 2: SIMS depth profiles are formed by collecting secondary ion intensity versus. time information as the primary ion beam sputters through a multi-layer structure.

makes analysis of existing molecules on surfaces and in thin films impossible. It cannot be know if the molecule was pre-existing or was manufactured by the sputtering process.

Dynamic SIMS is a highly effective depth-profiling tool. Static SIMS is very effective in the evaluation of surface contaminants and molecules. A full examination of the capabilities of both techniques is beyond the scope of the article. The rest of this discussion will focus on dynamic SIMS and the profiling of ultra-thin films.

Dynamic Sims

Instrumentation

There are two main types of instrumentation in common use today. These are called “magnetic sector” and “quadrupole”. Both of these names refer to the method of mass separation that occurs in the secondary ion beam (see Figure 1).

As the name suggests, the magnetic sector instrument uses a magnet to separate secondary ions based on their mass-to-charge ratio. The magnetic sector instrument is capable of high mass resolution that allows for the separation of elements from molecules that occur at the same nominal mass. Elemental identification is therefore less ambiguous. Good mass separation requires secondary ions with relatively high velocity, which in the case of magnetic sector instruments, is achieved with high extraction fields over the sample surface. High extraction fields make it difficult to bombard the sample surface with low energy primary ions, important for depth profiling thin layers.

Quadrupole SIMS uses a radio frequency field to filter out unwanted secondary ions such that molecules of only one mass can reach the detector. The radio frequency field can be changed in amplitude very quickly in order to monitor another mass. Rapid mass switching makes it possible to monitor many masses together in one profile while maximizing the number of data points that are collected. The radio frequency filter is more effective if secondary ions spend more time in the field. Therefore, low velocity ions are preferred which requires low voltage extraction fields over the sample. Low voltage secondary ion extraction fields also make it much easier to bombard the surface with a low energy primary ion beam. The primary ion beam must progress through the secondary ion extraction field to reach the sample surface. A low energy field will cause less deviation of the primary ion beam making it possible to sputter the sample surface at the desired angle of incidence.

Quantification

SIMS ion yields are dependant on the element of interest, and on the matrix material from which the element is sputtered. Ion yields for various element/matrix combinations can vary by orders of magnitude, making quantification extremely

Magnetic Sector

MagneticA nalyzer
Energy Slit
Projector Lens
Electron Multiplier
Faraday Cup
Electrostatic Analyzer
Ion Source

Quadrupole Mass Filter

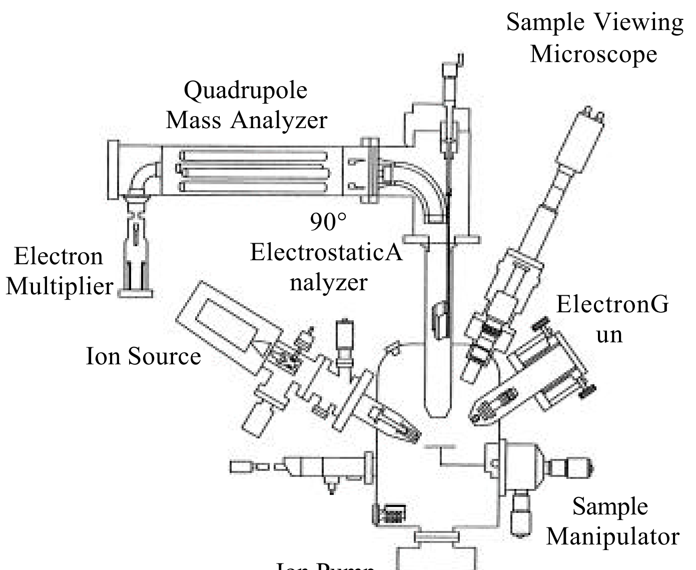


Figure 3: Diagrams of magnetic sector and quadrupole SIMS instruments in common use for dynamic SIMS depth profiling.

difficult. However, SIMS claims accuracy of better than 15% for most materials systems and better than 5% for some very specific applications. This accuracy is achieved by closely matching standards to the element/matrix material combination being examined.

The conversion of time to depth can also be done with a calibration standard, or the SIMS craters can be directly measured for depth.

Accuracy

SIMS quantification accuracy is dependant on the standard. There are a few standards for SIMS available from NIST and for these elements, accuracy is considered to be within 5% or better of the real value. Most other elements use ion-implanted standards that match the element/matrix combination that is being measured. Depth is most often determined by measuring craters with a stylus profilometer. The profilometer is calibrated using NIST traceable reference standards. Depth accuracy determined by a well-calibrated stylus profilometry is within 2%.

Precision

Reproducibility of the measurement is often more important than the absolute value of concentrations. Studies of "good vs. bad" and the effect of subtle variations in process parameters can only be seen if the measurements are very stable and repeatable. Precision is very much a product of instrument stability but careful sample placement and judicious selection of analysis location can also play a role. Typical precision using careful analysis is ± 5%. Repeat measurements on the same sample to generate some statistics can provide precision of ± 2% or better.

References

A good practical handbook explaining most of the important considerations for SIMS analysis in more detail can be found in Reference [1]. The latest in SIMS research and development can be found in the bi-annual SIMS conferences [2].

Ultra-Thin Film Analysis

Next generation IC devices have small lateral dimensions and very thin films. Auger analysis can provide measurements in small lateral dimensions and SIMS can provide measurements in very thin layers. With recent improvements to SIMS depth resolution, measurements on layers as thin as 1.5 nm are being made.

SIMS Depth Resolution

When an ion beam sputter etches a surface, damage is caused not only to the surface, but also to some distance under the surface. The primary beam is implanted into the sample, and a collision cascade ensues that disrupts the original positions of the atoms. It is the depth of this damage that is a prime determinant of depth resolution. To reduce the depth of this damage, the primary beam energy can be reduced so that the primary beam does not penetrate as deeply, and the ensuing collision cascade is not as energetic and does less damage. A cross-sectional diagram of this behavior is shown in Fig. 4.

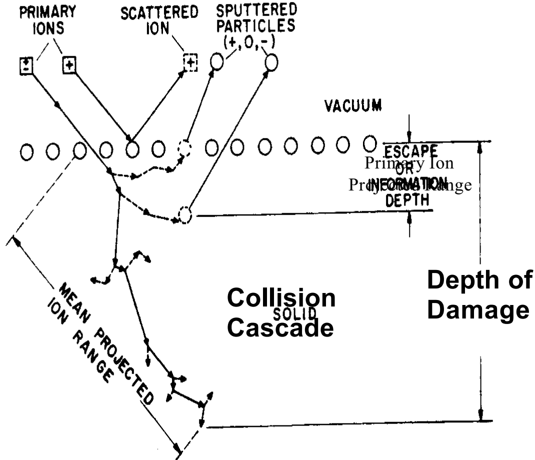


Figure 4: Damage done underneath the surface by a SIMS primary ion beam. Reduction of primary beam energy reduces this damage and improves depth resolution.

Low primary beam energy can improve depth resolution but the original surface must also be flat. Any sample surface irregularities or roughness becomes superimposed onto the crater bottom and will degrade depth resolution. If a layer of interest is particularly thin, it must be close to the surface in order to get the best depth resolution. The SIMS sputtered surface will roughen under the influence of the primary beam and depth resolution will degrade as depth increases.

Applications

The diagnostic capabilities of SIMS are generally used to evaluate the performance of fabrication tools used to make devices. That is not to say that SIMS cannot be valuable in evaluating materials performance failures. Examples will be shown.

Oxynitride

Oxynitrides or SiON is now seeing wide use as a gate dielectric material in CMOS devices. It has moderately higher capacitance than SiO_2 alone, and reduces interfacial-trapped charge. SiON thickness ranges from 5 nm down to 1.2 nm. SIMS profiles are done to measure nitrogen dose, and distribution, and to get a measure of oxide thickness. Very low primary beam energy must be used in order to get sufficient depth resolution to measure the nitrogen distribution. Quantification is done with standards that are closely matched to the concentrations and distributions of the samples being analyzed. Oxide thickness is calibrated using reference samples that are closely matched to the thickness of the measured samples. A profile from a relatively thick oxynitride film is shown in Figure 5.

Oxynitride analysis is often done to measure differences in nitrogen distribution and concentration caused by adjustments to various process parameters. High precision is therefore important in order to see the effects of these changes.

Even with very low primary beam energy, the SIMS profile still does not fully resolve the nitrogen depth distribution. For some of these oxides the total thickness consists of 5 or 6 atomic layers! The value of the SIMS profile is the ability to see the effect on distribution and concentration of small process changes.

Ultra-Shallow Implants

Another major driver for improvement of SIMS depth resolution has been the need to characterize ultra-shallow implants, particularly boron. Much has been written and published on this subject and a full discussion is beyond the scope of this article [3]. By using an appropriate beam, energy and angle of impact, accurate information can be obtain from very shallow implants with very high depth resolution.

Ultra shallow implants are characterized for various reasons. One of the most common applications is implanter evaluation. Many shallow implants are performed by ion implanters operating in “decel mode”. Ions are extracted from a source at relatively high energy in order to get good beam current, and then decelerated to get a low energy implant. Implant systems are not able to decelerate the entire primary beam leaving a percentage that is implanted at higher energy. The SIMS depth profile is done to characterize the implant for “energy contamination”, or the amount of implant that is not implanted at the intended energy.

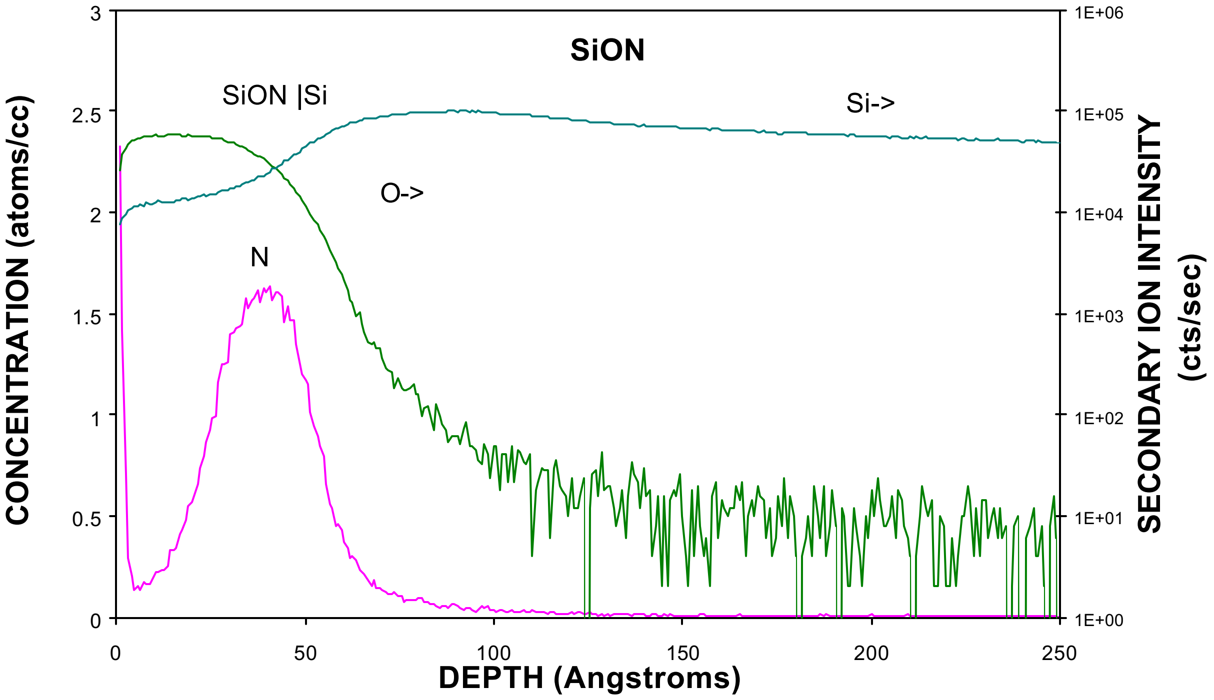


Figure 5: High depth resolution oxynitride profile showing nitrogen concentration and the position of most of the nitrogen at the oxide/substrate interface.

A profile showing energy contamination in shown in Figure 6. Data is shown on a semi-log scale, which is typical for SIMS data presentation. The energy contamination is a dominant visual feature in the profile, but this aberration is due to the use of the log scale concentration axis. The energy contamination portion of the implanted dose is really only a few percent of the total.

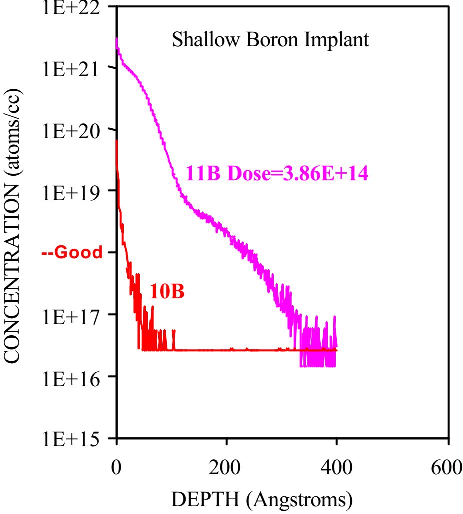


Figure 6: A boron ion implant showing energy contamination from a "decel" implanter. The curved boron distribution starting at 120 Angstroms and extending to 350 Angstroms is an energy contamination signature.

Another common application is the evaluation of activation anneals. In order to make the ion implanted dopant electrically active, the sample needs to be annealed. Unfortunately, there are rapid thermal diffusion effects that make it junctions. One idea that has shown promise is laser annealing. Figure 7 shows an overlay of the original "as-implanted" profile, and the profile after laser anneal. The overlay format of data presentation allow for direct comparison of changes.

SiGe

Silicon germanium is seeing increased use in high speed, high power transistors. SiGe devices can be used instead of GaAs.. Switching speed is still slower, but silicon process technology can be used to build the devices and they are easily integrated into silicon chips. SIMS plays a vital role in the development and diagnostics of deposition processes. SIMS profiles can show SiGe stoichiometry, dopant distribution, and contaminant incorporation. Thermal treatments can cause dopant diffusion and contamination can occur during epitaxial growth.

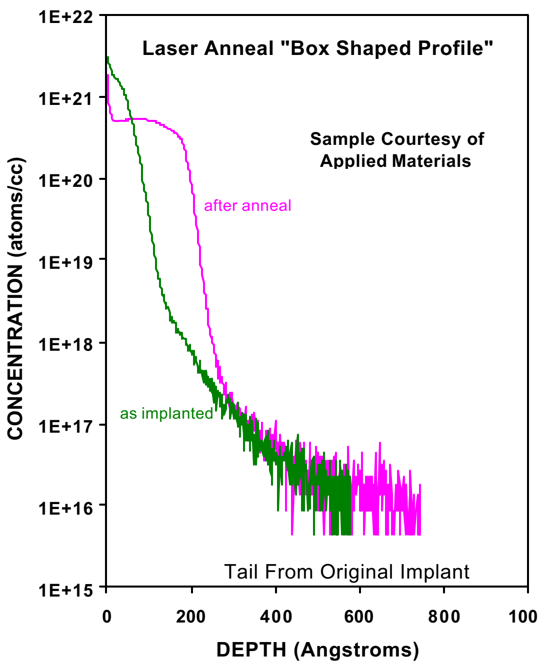


Figure 7: A shallow boron implant shown as implanted and after laser annealing. The tail is from the original implant and not from the annealing process.

Figure 8 shows a SiGe HBT structure. High depth resolution is necessary to resolve the thin layer structure. A sharp SiGe gradient is seen which is important for device speed. Dopants in the cap layer are seen and it is important to determine their concentration, and to determine if they have diffused into the SiGe layer. Oxygen contamination can be measured at the same time, which must not exceed 1E+18 at/cm3 or device speed is reduced Good information about the sample layer structure can be obtained in one profile. However, better profiles in terms of depth resolution or detection limits can be obtained using individual profiles with set-ups optimized for particular dopants or particular contaminants.

For instance, the oxygen contaminant and boron dopant can be measured along with SiGe stoichiometry in one profile using a cesium primary beam. However, the depth resolution for boron will be compromised by a strong cesium primary ion beam mixing effect. Much better depth resolution for boron can be attained using an oxygen primary ion beam, but then oxygen contamination cannot be measured at the same time. A choice needs to be made about what parts of the analysis are most important.

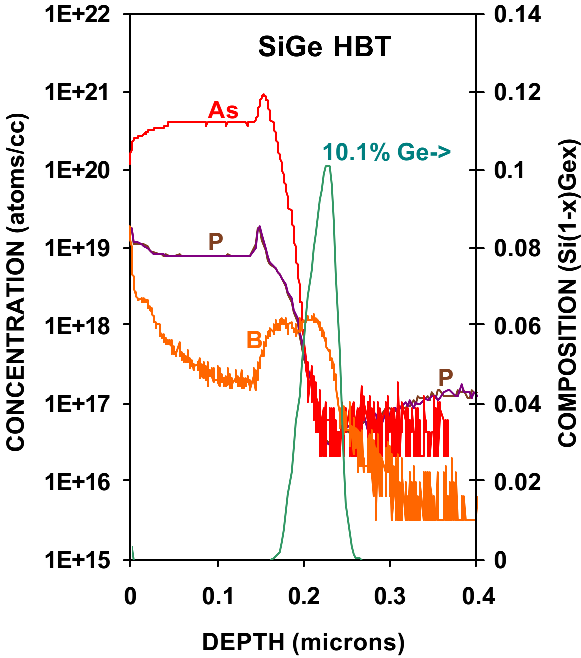


Figure 8: A high depth resolution profile showing dopant distribution and stoichiometry determination in SiGe.

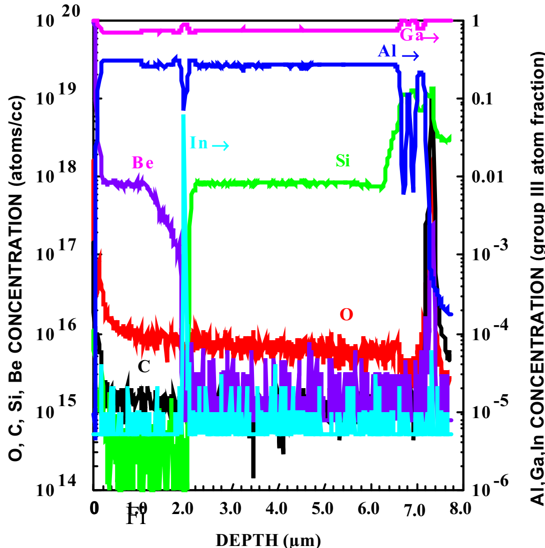


Figure 9: A depth profile of a laser structure showing the layer composition, layer thickness, dopant distributions, dopant concentrations, contaminant concentrations, and contaminant locations.

III-V Materials

High-speed communications are highly reliant on III-V materials such as GaAs. Depth profiling using SIMS is an ideal way to characterize these multi-layer materials that often have atomically sharp interfaces. Many of the same issues that are important for SiGe are also important for III-Vs. Carbon and oxygen can act as dopants and if present in unintended concentrations, can degrade device performance. Intentional dopants need to be present in desired concentrations and distributions. Sometimes intentional dopants will diffuse into other unintended locations. The SIMS depth profile can characterize these dopant distributions. Composition of the III-V materials can be measured, provided appropriate standards are available. III-V devices often have very complex layer structures. These devices can be characterized and the layer structure resolved with a depth profile. A depth profile of a III-V material can be seen in Figure 9.

Much of the SIMS work on III-V materials is as a diagnostic tool for evaluating the performance of layer growth processes and equipment, but sometimes in support of device failure analyses. Figure 10 shows an overlay of profiles taken from devices that exhibited "good" and "bad" performance. In this case a clear difference could be seen in the zinc dopant distribution.

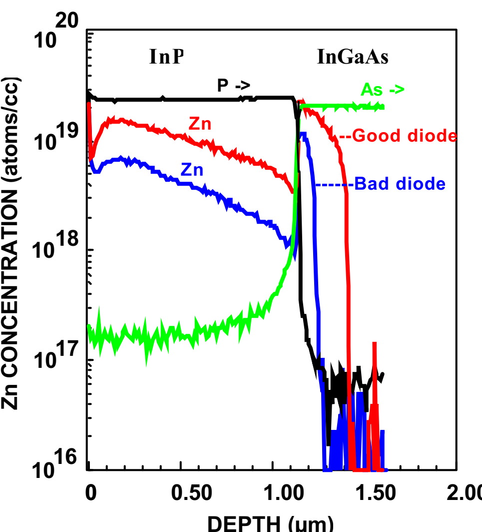


Figure 10: Overlay profiles from "good" and "bad" diodes showing large differences in the zinc dopant distribution.

Dielectric Materials

Dielectric materials can be evaluated with SIMS depth profiling. Electrically insulating materials can be difficult to analyze, but by using a combination of low energy secondary ion extraction fields and charge compensation provided by an electron gun, stable and reproducible depth profiling is possible. Passivation layers can be evaluated using SIMS depth profiling. In Figure 11, the layer structure of the passivation stack, aluminum contamination associated with

photoresist residuals, boron and phosphorous associated with a BPSG layer, and the presence of mobile ion contaminants can all be seen. The efficacy of the BPSG mobile ion gettering layer can be evaluated. This measurement can be used to evaluate process failure, but more frequently is used as a quality control measurement. The quality control measurement can provide proof that the gate oxide next to the substrate is free from mobile ion contamination.

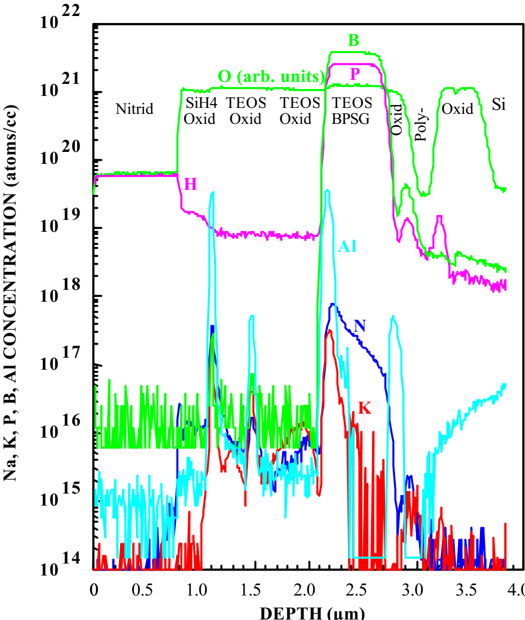


Figure 11: Profile of passivation layers showing layer structure, contaminants, and BPSG stoichiometry.

Mobile ion analysis is particularly difficult because as the name implies, elements such as Li, Na and K are mobile in SiO_2 under the influence of an electrical field. Charge compensation conditions that minimize the transport of mobile ions in SiO_2 are necessary before we can draw any conclusions about the meaning of mobile ion distributions. In the case of Figure 11, no mobile ion contamination was seen in the gate oxide providing a measure of confidence that the BPSG layer was acting as an effective getter for these elements.

Gate Oxide Breakdown

SIMS can be useful for device failure analysis. Figure 12 shows a "good vs. bad" depth profile study of gate oxides. The failed device clearly showed evidence of copper within the gate oxide that lead to breakdown, while the good device showed no evidence of copper. In this case the experiments produced a successful outcome, but the experiment must be well designed in order to be useful.

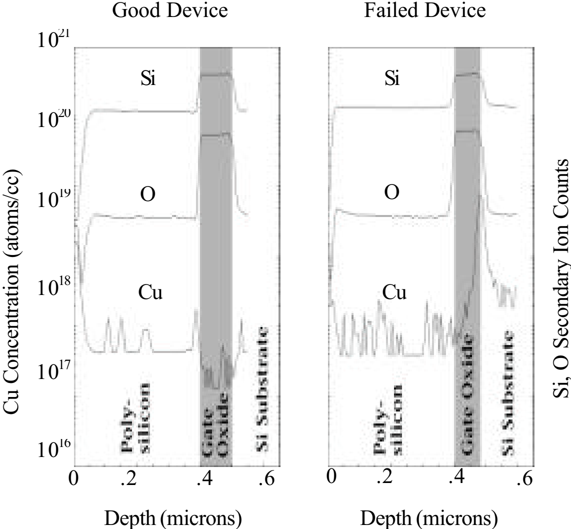


Figure 12: Depth profile failure analysis study revealing copper in the gate oxide of a failed device compared with a good device

Failure Analysis Experimental Design

Dynamic SIMS for failure analysis studies is rarely successful without some clues in advance as to what to investigate. For example, the search for an unknown contaminant at an interface will probably fail without a fairly short list of elements to investigate. The low chance of success is due to the nature of magnetic sector and quadrupole instruments. During a depth profile, elements are monitored sequentially, not simultaneously. The chances of sampling the specific element causing the problem are very low while profiling through thin films and interfaces, particularly if a long list of elements were being monitored.

A much more successful strategy for contaminant identification involves picking a relatively short list of elements to monitor during the profile. That will result in good sampling density for all of the monitored elements through thin films and interfaces. The list of chosen elements should include some matrix markers so that the layer structure can be identified, and a small number of suspect contaminants. The list of elements needs to be chosen by manufacturing process experts working closely with the SIMS analyst. With intimate knowledge of the process, a theory of failure can be developed, and a list of elements can be identified as most likely candidates. Working alone, it is almost impossible for the SIMS analyst to pick an appropriate list of elements. SIMS depth profiling is then done to confirm or refute various failure theories.

"Good vs. bad" experiments are an essential part of a failure analysis study. The SIMS technique is very sensitive and low levels of contaminants are found in many device structures. The fact that a contaminant was found in a layer or at an interface does not prove that it is causing the failure. Only when compared with a "good" can it be concluded that a root cause has been identified. Failure might also be due to differences in doping levels or layer structures. Again, only a "good vs. bad" comparison study will positively identify these differences.

Failure analysis profiles often result in a null result, e.g., suspect contaminant, dopant or layer structure differences are not found when comparing "good vs. bad". The null result holds value, however, since a theory of failure can be disproved and another theory can be investigated.

Surface and Interface Contamination

Many manufacturing processes have specifications that specify allowable levels of contaminants that may be introduced by a process. Contaminants are often introduced at the beginning or end of a process step, or during product transfer in and out of the processing vessel.

The high sensitivity of SIMS can be valuable for investigating contaminants, but there are difficulties in accurate quantification on surfaces. At the initiation of the SIMS analysis procedure, ion yields are low until the implantation of the primary ion beam element into the surface reaches high enough concentrations. During the time between the start of profiling and the maximization of secondary ion yields, called the "transient region", quantification is difficult.

One strategy for minimizing quantification difficulties during the equilibration transient is to use an oxygen primary beam and backfill the analysis chamber with oxygen. Oxide-like ion yields are then provided from the beginning of profiling on a surface, and accurate measures of surface contamination are possible. This analysis technique is called "SurfaceSIMS" and has proven to be valuable for the study of surface contaminants, particularly from ion implanters. Figure 13 shows a SurfaceSIMS profile of aluminum contamination that was introduced onto a silicon surface during the implantation of phosphorous. The distribution shows that the implanter added aluminum, not just to the surface, but energetically deeper into the sample which indicated that the implanter beam contained not only phosphorous, but also aluminum. By noting the peak of the aluminum distribution, the energy of the implantation can be determined and the point in the beam line where the aluminum contamination is introduced can be deduced.

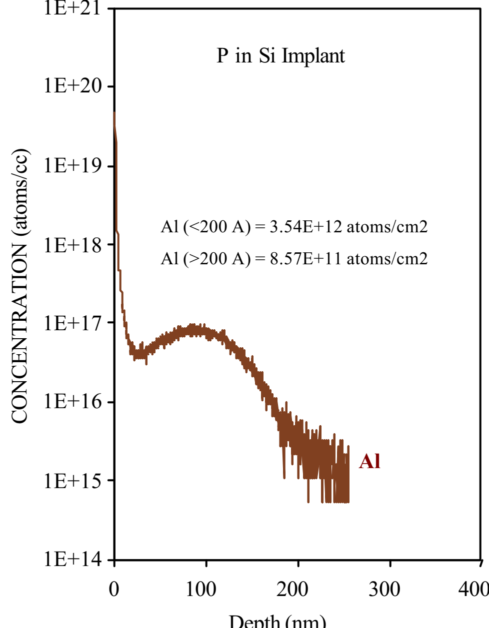


Figure 13: Aluminum contamination deposited on the surface and implanted energetically during a phosphorous implantation process.

Quantification at interfaces can also be difficult. The matrix material plays a strong role in the ionization probability of sputtered particles. If the matrix material changes from one layer to the next, then ion yields will change, sometimes by orders of magnitude from one layer to the next. Quantification of contaminants at the interface between layers of different materials can therefore be very complicated.

One way to check, especially for start-of-process and end-of-process contaminants, is to deposit several layers of material using the same process step being investigated. Using this technique, interfaces will be between layers of the same material and quantification will be straightforward, provided a standard is available. Figure 14 shows multiple layers of tantalum deposited by repeatedly using the same process steps. Contamination associated with the start and stop of the process can be seen accumulated at the tantalum layer interfaces. Contamination during deposition appears to be relatively low. Because each interface is between the same materials, no large changes in ion yields are seen and the quantification of the contamination will be accurate.

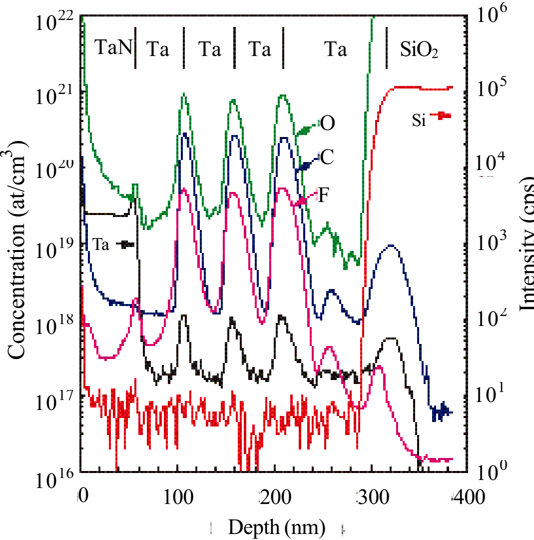


Figure 14: Carbon, oxygen and fluorine contamination are seen at interfaces between layers generated by cycling the sample through the same process. Most contamination is introduced at the start and end of the deposition process.

Conclusion

SIMS is a powerful tool for the investigation of dopant distribution, contamination, and layer structures, and can be an effective tool for failure analysis, especially when performing "good vs. bad" studies where high sensitivity and depth profiling capabilities are required.

Successful outcomes to SIMS investigations are much more likely with careful planning before the analysis.

References

1. C.W. Magee, F.A. Stevie, Robert G. Wilson, "Secondary Ion Mass Spectrometry: A Practical Handbook for Depth Profiling and Bulk Impurity Analysis", Wiley-Interscience, (1989).
2. "Proceedings of the 12th International Conference on Secondary Ion Mass Spectrometry, SIMS XII", A. Benninghoven, P. Bertrand, H.N. Migeon, H.W. Werner eds, Elsevier Science, (2000).
3. C.W. Magee, D. Jacobson and H.J. Gossmann, J. Vac. Sci. Technol. B, 18(1), (2000).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Focused Ion Beam (FIB) Systems: A Brief Overview

Kultaransingh (Bobby) Hooghan
Weatherfordlabs, Houston, TX
***Richard J Young**,*
FEI, Hillsboro, Oregon, USA

Abstract

In the past couple of decades, Focused Ion Beam (FIB) systems have become firmly established in the Microelectronics industry, transforming from being a "nice to have" to a "must have" toolset, for design houses, Failure Analysis (FA) labs and Production facilities (FAB's). This transformation has come about primarily due to the versatility of this tool and breadth of applications that are supported. One can go from circuit edit/debug, to analytical applications as quickly as one can change samples in the chamber.

In this article we will try to explain the basic operation of a FIB system and its main applications. These will be divided into: FIB basics; Electrical applications; and FA and Analytical applications. The theory and applications will be presented from a practical standpoint, with emphasis on understanding the systems and concepts.

There are many excellent reviews and overviews of FIB systems and applications [see for example, 1,2]. In addition, many use cases and application techniques are described in the conference proceedings from ISTFA, EFUG, and similar meetings.

Introduction

A FIB system is very similar to a scanning electron microscope (SEM). The primary difference is the use of ions (typically gallium) instead of electrons for imaging, sputtering etc. FIB systems can do many of the same operations as a SEM can do, but because the heavy ions are inherently destructive a wide range of other application possibilities are opened up by using the ion beam as a nanoscale machining (and deposition) tool.

The "Focused" part of the name can be interpreted in a couple of ways: (1) Focusing the field ionized Gallium ions into a "spot" and (2) focusing the beam energy within patterns drawn by software on sample surfaces, in order to accomplish milling (via physical sputtering) and deposition.

The destructive nature of the beam is attributed to a couple of basic facts. The Gallium ions are about 128,000 times the mass of electrons, and for similar beam energies, the momentum of the Gallium ions is about 360 times that of electrons [3].

FIB basics

The basics can be divided into: a) Ion beam generation, b) collimation and focusing, c) raster and signal collection, d) contrast, imaging and resolution.

Ion beam generation
The basis of the FIB in wide use today is the Liquid Metal Ion source (LMIS). These are field emission type ion sources. LMIS's typically have a small source size (~50nm) and relatively high brightness. The LMIS typically consists of a Tungsten needle wetted with the liquid metal, as shown in Fig. 1 below. Taylor showed that when a strong electric field is applied to the wetted needle, competing forces of electric field stress and surface tension result in a "Taylor cone" being formed at the end. When a high voltage is applied close to the liquid coated field emitter, the electric field at the end of the cone (which is essentially a point) will become high enough to initiate field evaporation and ion emission is started. Surface tension ensures that the tip is replenished with material from the wetted tip as ionization continues. For a detailed description of the source and column optics readers are encouraged to consult reference [2].

Figure 1: Actual gallium LMIS used in contemporary FIB systems, consisting of a W tip and a reservoir mounted on a metal/ceramic substrate. Ceramic diameter ~ 1cm.

The most common material for the LMIS is gallium, as it is: a) liquid at room temperature and b) has low vapor pressure at room temperature. The Ga LMIS is well suited for high-resolution applications because of its basic characteristics, namely: 1) good angular intensity, 2) high brightness, 3) narrow energy distribution of ions, 4) good lifetime and

operating stability. Due to the high intensity, into a small solid angle, only modest apertures are needed to achieve currents in the nA range. Corresponding spot sizes are in the 5-500nm range. High brightness corresponds to a smaller optical source size; hence smaller spot sizes can be achieved, in an optical column, with only two lenses [2].

Collimation and focusing

As the ions are considerably heavier than electrons, magneto static lenses lack the necessary "muscle power" to deflect/steer ions beams. For comparative beam energies the magneto static lenses would have to be about 30 times stronger, which is simply not feasible. Also the steering and focusing effects of electrostatic optics are essentially the same independent of the charge to mass ratio of the particles used. Hence electrostatic lenses are used.

Similar to electron optics, the beam spot shape/size on the target is a result of the following factors: a) Spherical aberration coefficient, b) Chromatic aberration coefficient, 3) finite source size, and 4) diffraction coefficient [2].

Mathematically it has been shown that a) at low values of current, the physical size of the source dominates the subsequent spot size, b) at intermediate currents, chromatic aberration is the dominant effect on ultimate spot size, and c) at higher currents spherical aberration is the dominant affect on spot size.

In earlier designs, FIB systems were made with a single lens system, with conveniently long working distances. However, two lens systems are the norm in contemporary FIB systems. Pre-lens deflection systems help reduce the working distance between final lens and sample, while dual-deflection systems enable the beam to be pivoted through the center of the lens as it is scanned over the sample – both affects improving optical performance of the ion column.

Figure 2 shows a typical FIB schematic for a two lens system. The column contain a beam acceptance aperture in the source region that sets the maximum beam current, while the beam defining aperture is normally a strip of different sized holes to define the lower beam current values.

Raster and Imaging

After the beam is generated and collimated it is focused on the sample and rastered in a given pattern. Ion beam interaction with the surface gives rise to secondary electrons and ions, ion implantation, and sputtering damage. Secondary's (both electrons and ions) are used for imaging. Secondary electron emission is typically 10-100-fold more than secondary ion emission; hence imaging using secondary electrons is the generally preferred method of imaging, unless there is a good reason for using secondary ions, such as minimizing sample charging, or looking for particular materials contrast. The penetration depth for the ion beam, while imaging is typically 200-400 Å, depending on the material [1]. Software is available that can simulate the interaction of ions into different materials [SRIM, 4].

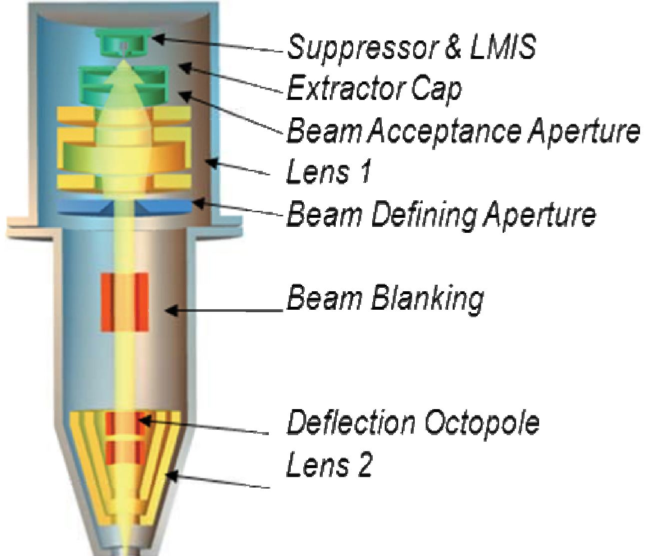


Figure 2: Schematic of two lens FIB system.

To form an image, charged particles emitted from the sample surface are directed toward a biased grid. They are then collected by one of several types of detectors mentioned below. The pixel-by-pixel signal is then amplified and displayed on a monitor as an image.

An overview of the different detectors that may be used for imaging is as follows:

Micro Channel plate (MCP): is typically located around the pole and is used to collect electrons or ions. It consists of many small diameter hollow tubes forming a honeycomb structure, and a cascade effect from a primary particle amplifies the signal. Grid bias determines whether the electrons or ions are detected. The physical size of the MCP means they are normally just used with "single beam" FIB systems (see later for more discussion about "single" v. "dual beam" systems).

Channel Electron Multiplier (CEM) or Continuous Dynode Electron Multiplier (CDEM): This is in the proximity of the sample surface, and is effectively a single large area electron/ion multiplier.

Scintillator ("Everhart Thornley" type, ETD): This is typically an electron detector only, typically being used as an off axis detector. Also, on FIB-SEM dual beam systems, there may be additional detectors available, such as "through-the-lens" types that give additional options for image formation.

Contrast, Imaging and resolution

There are four basic contrast mechanisms in the FIB system: 1) Topographic contrast; 2) Materials contrast 3) Ion Channeling contrast and 4) Passive voltage contrast

1) Topographic contrast depends on surface topology and is the basic contrast and imaging mechanism, as it is in the SEM (Fig. 3a).

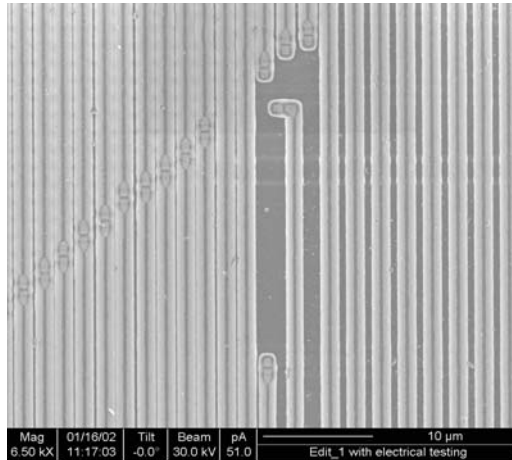


Figure 3a: Picture showing topographic contrast in FIB systems

2) Materials contrast is dependent on the compositional differences in the area being rastered with the ion beam. Different materials can have significantly different secondary electron or ion yields which creates varying image contrast (Fig. 3b).

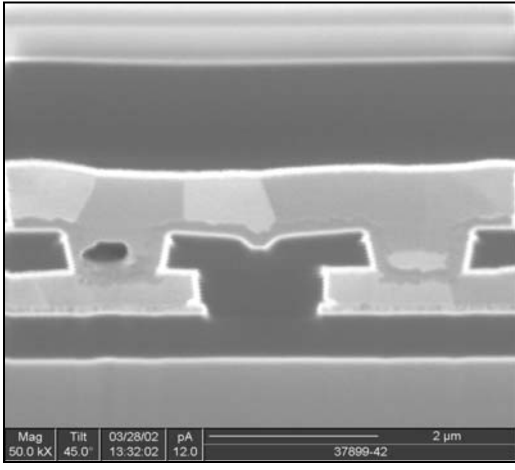
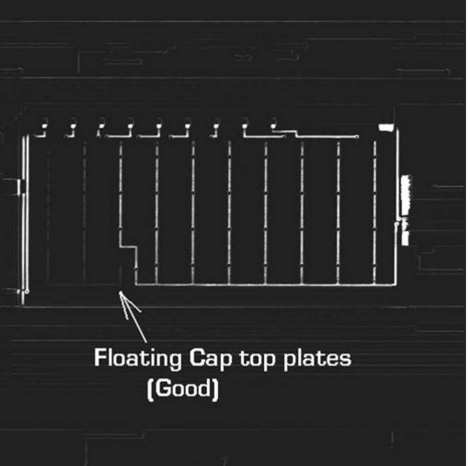


Figure 3b: Picture showing combination of materials contrast (compare barrier layer with main metal interconnect directly above); ion channeling contrast (in the main metal lines); and passive voltage contrast (the inter-layer dielectric areas appear dark)

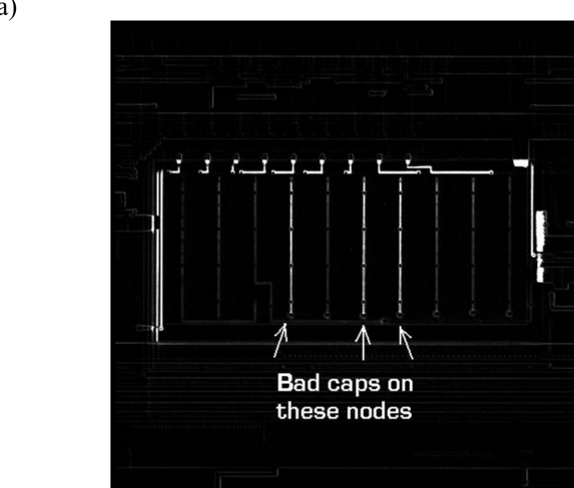
3) Channeling contrast depends on variation in secondary electron yield with grain orientation, and grain orientation with respect to beam direction. These yield variations produce changes in intensity, which develop contrast in the image related to grain structure.

4) Passive voltage contrast (PVC) results from the charged nature of the beam. If the surface/component is electrically floating (not grounded), the beam charge is not dissipated and hence collects on the surface. The positively charged surface does not release the secondary electrons needed to generate an image. On grounded surfaces the charge does not accumulate on the surface, and hence secondaries are released in order to form an image. Given that the ion beam is positive and the secondary electrons are negative, then the net surface charge will always be positive. This differs from the SEM where either negative or positive charging is possible, depending on whether the insulating region has a net accumulation or net emission of negative charge.

There are two main areas that FIB-PVC tends to show up: firstly on insulator layers, where strong contrast is seen between such layers and grounded metal lines or the substrate (e.g. see Fig. 3b); and secondly also when a conductor is electrically isolated. This is extremely useful in knowing which nodes in a given circuit are open or shorted. Figure 4 gives a nice working example of voltage contrast imaging. Such comparisons between a "known good" and a "possibly bad" region are fast ways to localize a defect site. By using the ion beam to cut or ground different parts of the array you can quickly isolate a break or short.



(a)



(b)

Figure 4: Capacitor structure with top plate floating, demonstrating passive voltage contrast with FIB. On the reference "good" sample (a), the bright tracks are grounded, while the dark lines are floating. In the "bad" sample (b) more of the tracks are dark, indicating that the lines are not correctly grounded.

Resolution: In the conventional sense resolution is the ability of the imaging beam to resolve two nearby objects. The focused ion beam throws an additional twist in that being an

inherently destructive beam, means the specimen is constantly being sputtered away while being imaged. Therefore ultimate resolution is also dependent on the resilience of the sample to the sputtering caused by the beam, as the sample will be ultimately sputtered away [5]. Imaging resolution in contemporary FIB systems is sub-5nm.

FIB Applications

These can be divided broadly into two areas namely Electrical and Physical applications. Electrical applications can be further divided into basic device edits, electrical probing assistance and voltage contrast.

Physical applications can be further divided into several cross-section techniques for SEM and TEM sample preparation. These can be in support of defect origin and identification (including, EDS, SIMS etc.), process monitoring, metrology, imaging, materials contrast, grain size determination etc.

Before one gets into the applications there are a few overall system configurations and capabilities which need to be highlighted.

FIB Configurations:
There are a wide range of FIB based systems available on the market, but they generally fall into one of two main types – the so called "single beam" FIB and the "dual beam" FIB-SEM system, where the ion beam is combined with a scanning electron microscope. These dual-beams have an ion and an electron column in the same system, usually with an angle of 45-60 degrees between the columns (Fig. 5).

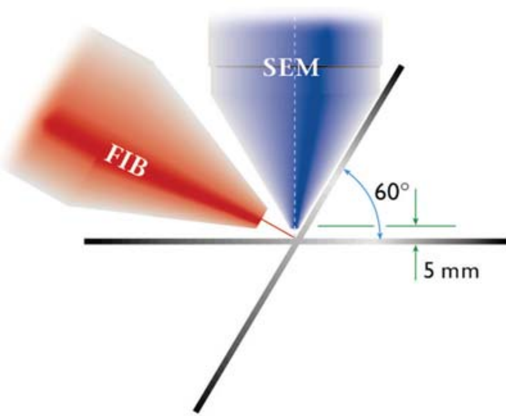


Figure 5: Example of a dual-beam FIB-SEM configuration, showing the ability to tilt a large sample (e.g. a wafer) to 60 deg tilt while maintaining a 5 mm SEM working distance at the beam coincident point.

The SEM column is typically a high resolution (field emission type) electron column and may also have analytical capability, such as energy dispersive X-ray (EDS, EDX). One can navigate using the SEM column and hence impart less surface damage to the sample surface. Also there is no need to tilt the sample to look at the cut surface of a cross-section (as there would be on a single beam FIB), as the SEM column is be looking at it constantly, enabling a "slice and view" operation to more accurately locate the exact area of interest with a section.

Within semiconductor labs dual beam FIB-SEM systems are now in the majority compared to single beam systems, although the single beam still has a key role to play in some important applications, in particular for circuit edit.

Beyond the question of one beam or two, the other major differentiation between systems types is the size of the sample stage. Full wafer systems, with 200mm and 300mm stages, whether in the FAB or the lab still have an important role to play as they enable the wafer to be kept intact for the analysis, (for example, allowing for easier defect navigation), but many systems have stages with 50-150mm travel capability, dealing with wafer pieces, or else bare/packaged/repackaged die. The wafer systems can often also handle such smaller samples too, enabling them to fulfill a dual role of wafer and small sample analysis.

FIB Capabilities:
These can be broadly broken up into imaging (discussed earlier), material removal, and material addition (deposition).

Material removal can be achieved by straight sputtering and by gas assisted etching (GAE). Even while imaging a surface, the surface is being milled away a few mono-layers at a time. This is just the result of the significantly higher momentum of the gallium ions that impinge upon the surface, compared to electrons. Unfortunately with regard to FIB systems, this is a way of life, and the "damage" due to milling can be minimized to a certain extent but cannot be eliminated. Reducing the beam current is one way to reduce the sputtering damage during imaging, but is to be balanced by the secondary electron/ion yield required for imaging.

Software available on FIB systems allows patterns to be drawn on sample surfaces and the sample is "milled or sputtered" within those boundaries. The Operator can control the raster area and the key points like dwell and overlap (or step size). A FIB raster schematic is shown below.

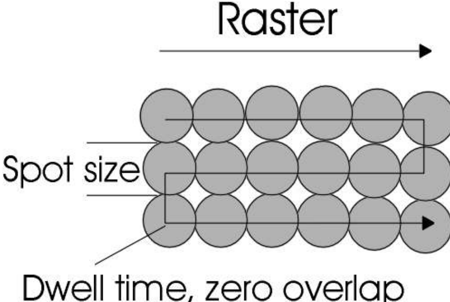


Figure 6: Raster schematic for typical FIB system. Spot size depends upon the aperture used; dwell and overlap can be varied for different outcomes.

This milling or sputtering is used in FIB systems to remove material. This may be to expose certain faces as in a cross-section, or to expose underlying metal runners in order to connect them elsewhere, or disconnect them.

Straight sputter or milling has limits in terms of the aspect ratio of milled depth to pattern size. As aspect ratio increases so does the effect of redeposition, with tapering of the milled hole occurring, ultimately limiting the aspect ratio to perhaps 5:1 or less. Whereas this is not a big issue as far as cross-sections are concerned, with device edits it is indeed a show stopper in many cases. This is especially true when one has to expose metal lines other than top level metal as the redeposited material can act as a short between layers or across a track that you are attempting to cut. Redeposition is a constant battle in that case and it is over come by Gas Assisted Etching (GAE).

GAE is achieved by providing a localized pressure of an etch gas in the vicinity of the beam scanned area using a needle in close proximity of the sample surface. The needle is usually about 50-100µm from the surface, the local flux being several orders of magnitude higher than the pressure rise seen in the overall specimen chamber.

The gas flowing from the needle converts the sputtered material into a volatile compound for many substrates which is pumped away by the vacuum system, resulting in very little redeposition. Also, different gases react with different materials and hence milling selectivity between materials can be enhanced. GAE has certain distinct advantages over straight sputtering: etch rate enhancement for the same beam current, better selectivity between materials, and absence of redeposited material [6]. GAE gases are typically halogen based, such as I_2, XeF_2, Cl_2, Br_2, but more complex chemicals have been developed for use in specific applications, and even water is beneficial for carbon based materials such as polyimide and resist.

FIB induced deposition is used to add material to produce conductors and insulators. Conductors are deposited to reroute a circuit, or make probe points. In the analytical realm, conductors and insulators are used as a sacrificial layer to protect the top surface of the sample while milling. Insulators are also deposited to prevent leakage between exposed conductors. Deposition is achieved by the injection of a precursor gas into the chamber near the sample surface, similar to the GAE procedure described earlier.

Conductor deposition: The two most common and popular metals used for conductor deposition are platinum and tungsten. The metals are delivered using an organometallic gas, which adsorbs on the sample surface. Beam parameters are adjusted such that the resulting beam dose is sufficient to dissociate the adsorbed organometallic gas on the surface, but not to remove the underlying substrate by sputtering. The byproducts are then pumped away by the vacuum system. The reacted gas is desorbed on the surface with every raster of the beam and the metal layer is build up with repeated rastering. Though the metal film is by no means pure, containing a large proportion of organic components and gallium, the electrical characteristics are useful for most FIB applications. Platinum generally deposits much faster than W, but W has better electrical properties (lower bulk resistivity) than Pt. Carbon deposition is also used, but this is mainly for sample protective purposes, as the resistivity is far higher than W or Pt based deposition.

Insulator deposition: With a switch in the precursors, insulators can be deposited on the surface just like metals. Tetraethylorthosilicate (TEOS) or similar compounds are used as a precursor in FIB systems (often in combination with oxygen or water) and the end product is an SiO_2 like material. Good isolation of cuts in metal lines can be achieved by filling the cut with an insulator. Insulator deposition is very useful in accessing metal lines buried under power buses, where a part of the bus is taken out, backfilled with insulator and then a hole is drilled through the insulator to access and connect to the lower level metal. It can also be used to protect TEM samples as it provides a different contrast in the TEM than other conductor depositions.

With a dual beam FIB-SEM system the electron beam can also be used to induce beam chemistry. Within semiconductors the most usual application of this is for TEM sample protection, where an electron beam deposited layer 50-100nm thick can be used to prevent ion beam implantation and amorphization in the top surface of a sample [1]. In general, electron beam deposition is slower than the equivalent ion beam process, but going to lower kV (e.g. 1-2 kV) and higher currents (> 1 nA) can help boost the deposition rate.

Electron beam deposition can also be useful for planarizing samples that have high aspect ratio structures exposed on the surface – the planarization helping to minimize "curtaining" during the subsequent sample preparation.

FIB Electrical Applications:

As mentioned earlier these can be divided into Device edits, electrical probe assistance and voltage contrast imaging.

Device edits is the ability to carry out physical changes on the chip, as verification to design fixes, or changes in design, requested by the customers. Changes carried out on the chip, using a FIB system, are usually sufficient to do board level tests for verifications. Design changes could be to fix mismatches between layouts and schematics or accommodation of some last minute design requirements by clients. Traditionally, edits were carried out from the front side, but with contemporary chips having 9 or more metal layers and with the prevalence of flip chip packaging for high performance devices, editing from the back side of the chip is continuing to increase (see for example [7], for a recent ISTFA paper on this subject, with an extensive reference list) . A full discussion of front and backside editing is beyond the scope of this article, so just the main principles (focused on front side edits) will be covered.

The outcome of a device edit is critically dependent on the co-operation between the person submitting the request and the FIB engineer carrying it out. Once one understands the exact request, whether it is for probing, making a few cuts/connects etc. one needs to keep in mind the feasibility and the practicality of the edits. For this reason the device edit project can be divided into the Planning and Execution phases [8].

It is during the planning phase that the edit is effectively carried out "on paper", knowing what exactly needs to be done, getting a sufficient number of devices decapsulated (and tested good, post decap), checking for accessibility (whether one needs to approach the edit from the front or the backside of the chip), gathering the relevant targeting information and studying the FIB deposition layout.

Time spent on this stage of the project is fruitful later on, as good planning makes execution a breeze. It is really worthwhile for the two engineers to check the layout thoroughly, so that one can do the edits on metal levels as close as possible to top level metal (or lower levels if dealing with a backside edit). When accessing metals in the lower levels from the front (e.g. metals 2 or 3, in 6 to 7 metal levels) the aspect ratio increases, so end point detection can be a challenge. Also, the likelihood of nicking adjacent lines increases and then filling the metals also becomes very tricky. Points to be noted are that connects are made at the topmost level possible, in an area free of adjoining lines, and away from active circuitry. These conditions may not always be met, but are good baseline guiding principles. Engineers and layout personnel really get a feel for what is being talked about, when they actually sit in front of a FIB system and see the difficulty of what they are asking for. One usually has to do this exercise once or twice and people are convinced and are better prepared next time.

The next step is the gathering of targeting information, in order to get to the fixes. Depending on the technology (interconnect line width) one would gather targeting information to get to the proposed fixes. These would be info for navigating by visual references (dead reckoning), 3-pt alignment, or CAD navigation. No matter what technology one is dealing with, hard copies of plots are always useful and welcome. Black and White 8 ½" X 11" plots usually work fine, but colored plots are better. Depending on the fixes involved, one could get plots showing all or few metal levels, so that one can anticipate trouble spots. CAD navigation/viewing programs may also enable "virtual layers" to be added the CAD layout, allowing the proposed edits to be better visualized and documented.

Also to be considered at this point is the deposition layout. Design Engineers are usually geometrical i.e. "the shortest distance between two points is a straight line". That's how most of the connections get laid out. Considering the inherently destructive nature of the imaging/milling beam one has to make sure of avoiding sensitive underlying circuitry, when planning depositions. Parametric changes in transistors have been documented, with some electrical changes being recoverable but not all [9-11]. For example, a hydrogen anneal has been used to recover much of the threshold voltage shifts associated with FIB edits.

Once the edit has been carried out "on paper" and fine tuned it is time to move on to the Implementation/execution phase. So far the device undergoing repair was just being optimized for success rate and FIB time. The Implementation/Execution phase can be broken up into the following components: A) Sample mounting/grounding, B) Imaging/navigation, C) Vias (milling/filling) D) Connections, E) Disconnecting lines, and F) Cleanup [9].

A) Sample mounting and grounding: devices come in a variety of packages and they need to be mounted so that they do not move in the FIB chamber while being worked on. FIB engineers use a variety of methods such as metal plates (with different size hole cut outs), double sided tapes, Al or Cu metal tapes to hold the sample down and effectively ground it. Effective grounding is extremely important, because of the inherently destructive nature of the beam. If the surface charge is not dissipated effectively, it usually results in blown gates or pop ups in the protective oxides. Besides, accumulation of surface charge also results in bad image quality. When a chip is well grounded, the bond pads light up very well as seen in the figure below.

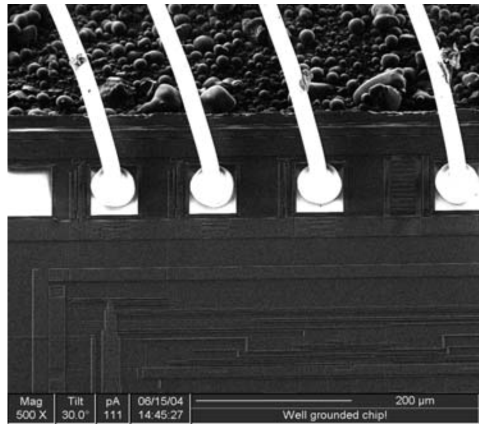


Figure 7: Well grounded chip

B) Imaging/Navigation: Effective grounding is a primary condition for good imaging. Imaging is also a tradeoff between the amount of charge a device can handle and the secondary ion/electron yield. Beam current is what governs the charge and image quality. One can only go so low in beam current before the amount of secondaries generated are not enough to obtain an effective image. When this happens there are a couple of options available:

1) Apply a local Pt coat: this is usually achieved by extending the Pt needle and manually opening the valve so that the Pt precursor gets deposited on the surface as the ion beam scans and helps dissipate some of the charge.
2) Apply a global carbon coat.

The only caveat with (2) is that it entails an additional step of burning off the surface carbon using an oxygen plasma. One has to be careful though when using a plasma asher, to make sure that the leads are well grounded in the asher.
These options work really well for improving imaging as the grounded surface gives much better secondary electron imaging even at lower beam currents. Also the lower currents and the conductive surface make it less likely that electrostatic damage will harm the sample.

An alternative to coating is to used a charge neutralizing flood gun to dissipate the charge [1,8] These units provide a low energy flood of electrons to neutralize the surface charge and so must be used in secondary ion mode for the FIB to be able to image. It should also be noted that secondary ion mode alone can also be a beneficial method of minimizing charging effects when navigating – by biasing the detector to collect positive secondary ions you are naturally pushing any secondary electrons back to the sample.

Navigation: On chip navigation is broadly categorized into the following: a) Dead reckoning, b) 3-pt alignment, and c) CAD navigation.

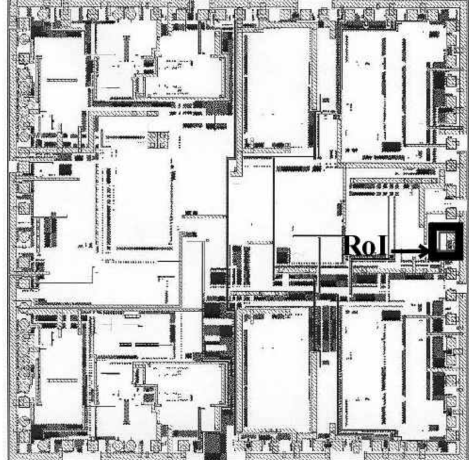


Figure 8: Overall chip plot showing RoI

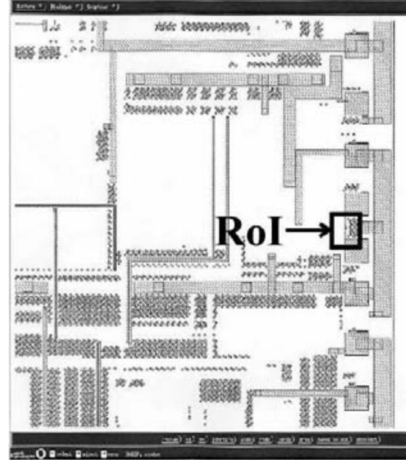


Figure 9: RoI zoomed in

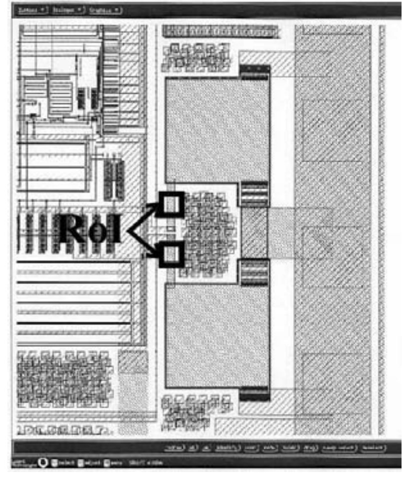


Figure 10: Regions of Interest zoomed in further

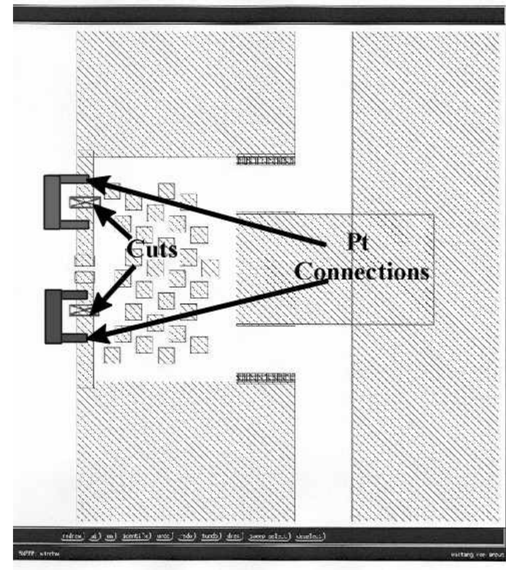


Figure 11: Edit schematic

a) Dead reckoning: This is just what the name implies, navigating by visual references. This is true typically for the relatively bigger geometries > 0.25 μ processed with non-planarized metal levels. The following four figures (Figs. 8-11) give a typical flow for the type of drawings/plots needed for navigation to the Regions of Interest (RoI)

b) 3-pt alignment: is the next progression for navigation. In this method the FIB stage is aligned to the circuit layout using a minimum of three points, usually taken on the corners of the chip in either clockwise or counter clockwise direction. This ensures that the stage and the chip are aligned and one can go to the region(s) of interest by simply entering the x, y co-ordinate of the target location(s). The (x, y) co-ordinates can be obtained from the layout tool directly by maneuvering the cursor to the desired locations, and then the stage is aligned to those locations. In the figures below, the same chip is shown with an overall chip plot and the lower right corner. The feature presented there is a mask alignment marker and is quite visible most of the time.

The alignment would typically proceed in the clockwise direction, with the alignment markers pointed right side up. Once the alignment is carried out, the zoomed plots are used to go to the regions of interest and the edits can be performed.

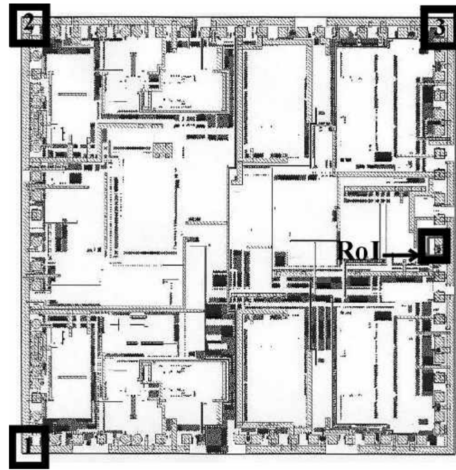


Figure 12: 3-pt alignment overall

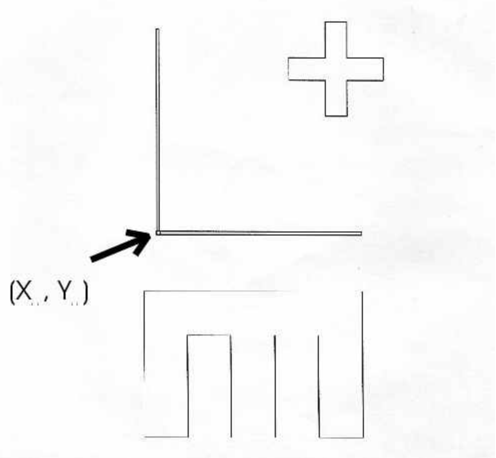


Figure 13: Corner 1 zoomed in,

c) CAD Navigation. This is the next step up in navigation. There are certain similarities between 3-pt alignment and CAD alignment, as the overall procedure is essentially the same. Three points along the periphery of the chip are used to align the database to the FIB image of the device. The difference between the two is that one has to convert the raw data into a format which can be read by the particular FIB system. The conversion has to be carried out usually on a standalone computer workstation and then transferred to the FIB system's computers. Once the layout is aligned and locked to the FIB system, one can carry out on-chip navigation using either the layout or the FIB system. On some systems one can draw FIB boxes for cuts/depositions etc and transfer them to the FIB to carry them out.

C) Vias: Vias for connection to top level metal lines can be just about a couple of microns on the side. Vias on lower levels are trickier. One has to take into account the fact that aspect ratios get larger the deeper we go into the chip. One has to make sure of using GAE in order to expose metals and also use endpoint detection to know when metal is exposed.

Another challenge is to make sure that the vias are filled up all the way. This means one has to avoid "bridging" of the via, where metal deposits only on the top of the via and does not contact the metal below. Also one has to make sure that the contact area is sufficient to minimize contact resistance.

This can be achieved by increasing the area of the mill boxes used; rectangular instead of square would give more area. Secondly, when filling up the contact holes thus made, one has to make sure that the fill boxes used are smaller in both x and y dimensions than the original box. This normally ensures that the subsequent deposition will go all the way down to metal.

D) Connections: Once the vias have been milled and filled up to the passivation, the next step is to make the connections. As mentioned earlier one needs to make sure of avoiding making depositions over active circuitry. A typical deposition line is 75µm x 1µm x 1µm. Typical beam currents used would be about 300-350 pA. One can use more current but then the material files would have to be altered to get a net deposition instead of milling. FIB software usually adjusts the dwell and overlap for different beam currents. Lines can be as long as needed, with the one of the author's (BH) record being one which was 2300µm long and the FIB edit was successful!

E) Disconnecting lines: All connections are to be completed prior to disconnecting lines as if we cut lines before deposition, the overspray, from deposition will short the disconnected lines. The lines are to be exposed just like opening for vias, using GAE and end point detection. One can then switch gases to a metal selective gas and remove the metal. Barrier metals may need to be tackled by a different gas in order to get a clean cut [1,8].

F) Overspray and clean up: Overspray is an unfortunate side effect of deposition wherein metal gets deposited outside the pattern that is drawn. As the beam is of a Gaussian shape, the beam tails are what give rise to depositions outside the drawn patterns. The overspray is of concern only when two adjoining FIB edits have overlapping over sprays. The two edits are then isolated from each other using sputtering or GAE.

Electrical Probe assistance: The FIB system can be used to probe signals using physical probes or e-beam probing. The probe pads are put down as if exposing the underlying metal for a FIB edit. Typical probe pads for external probing are shown in the figure 14 below. They are shown isolated from each other.

In-situ probing with probe needles inside the FIB vacuum chamber is also possible. For such probing, the FIB (or SEM on a dual-beam) enables fine positioning of the probes so that much smaller probe pads are required, with even individual exposed lines being able to be directly probed if needed.

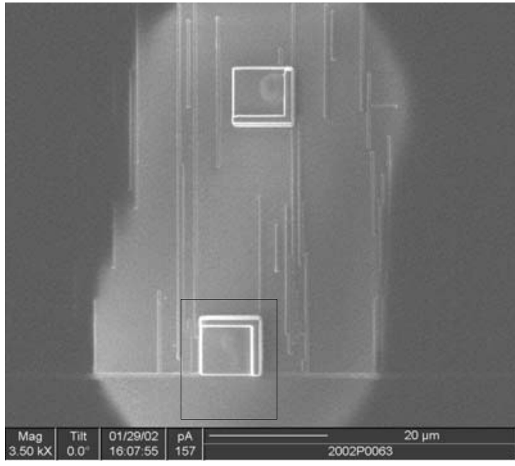


Figure 14: Probe pads

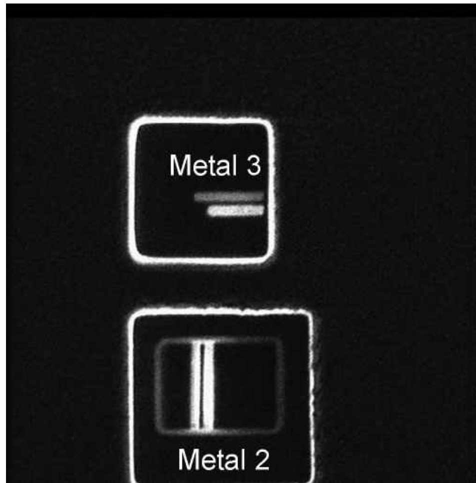


Figure 15: Voltage contrast imaging for Fault isolation

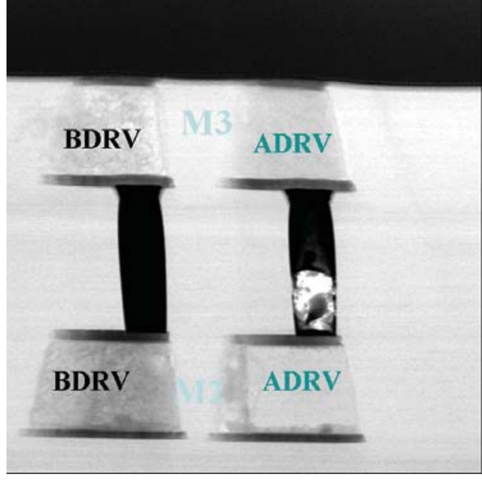


Figure 16: Physical evidence of open, follow up on Fig 15

Voltage contrast imaging: is used for failure analysis, in order to isolate floating nodes as opposed to nodes which are grounded. Figure 5 shows excellent examples of voltage contrast imaging for isolating bad capacitors. Figure 15 is an example of voltage contrast used for FA to isolate a bad via, prior to doing physical FA. In this particular example, the metal 2 and 3 lines were connected by vias, but as one can see, three of the four metal lines in question are bright and one of them is darker. That would be indicative of a bad via (high resistance) between the two lines, otherwise all lines should have lighted to the same intensity. This image was obtained by exposing the metal lines and then imaging them using the FIB system. This anomaly was subsequently confirmed by physical FA (Fig. 16) by making a TEM sample through the area of interest.

Analytical/Failure Analysis (FA) Applications:

The key analytical/FA application of FIB-based systems (and especially the dual-beam FIB/SEM combination) is site specific sectioning for either imaging in the prep tool (e.g. by FIB or SEM), or for analysis in a TEM (Fig. 17). This is because the FIB can make localized openings through widely differing material types, enabling analysis on areas that would not be possible by mechanical polishing due to both the localization required (e.g. a 20 nm thick TEM sample must be located to within a few nanometers of the correct location), and due to sample preparation challenges, such as material smearing/tearing due to differences in hardness.

In addition, multiple samples from nearby locations can also be obtained (Fig. 18), which could even be in different orientations if needed.

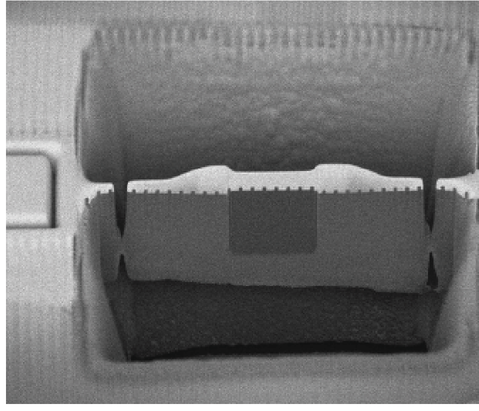


Figure 17: Automated TEM sample ready for extraction (in an external "lift-out" system) and placing on a TEM grid.

Over time, the advances in Moore's law, and the ever shrinking dimensions combined with new materials and structures, have pushed analysis systems to provide improved imaging in the sectioned results. This has been partially accommodated by improvements in the SEM resolution and contrast in the dual-beam, but has also resulted in many more TEM samples being made by dual-beams too, Fig. 19 shows two similar areas imaged as a dual-beam cross-section and in a TEM. The SEM view shows important details of the structure, but the TEM image provides additional clarity, although at the expense of longer prep and analysis time.

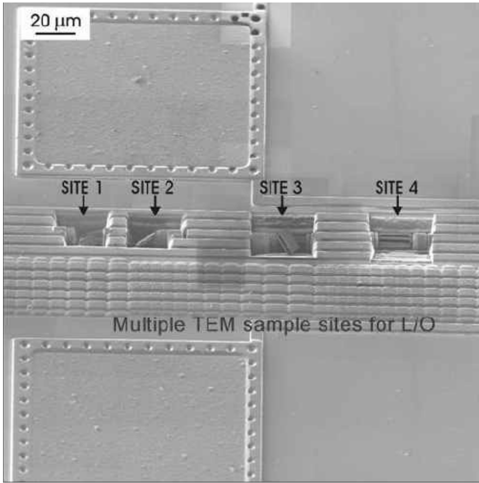


Figure 18: Multiple TEM samples obtained from nearby locations.

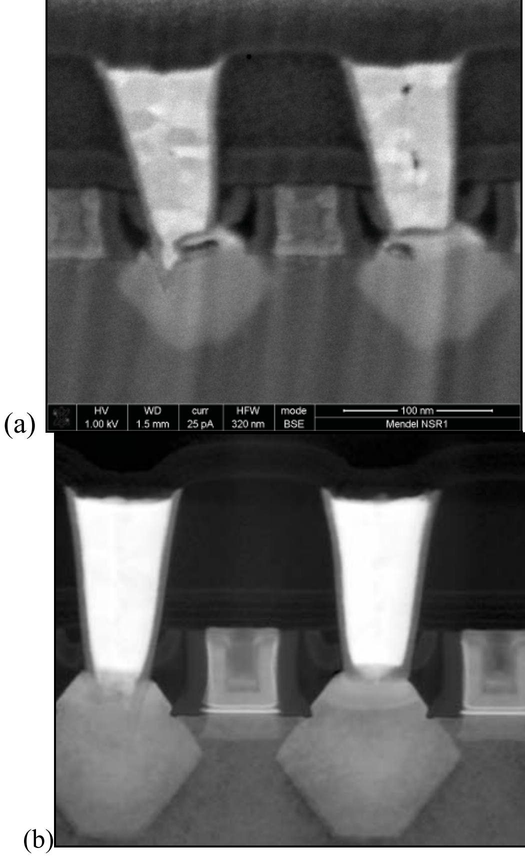


Figure 19: Comparison of SEM cross-section and TEM view of similar structures.

The increased need for TEM ample preparation has resulted in the use of automation (either full or partial) to improve overall throughput, and to enable sample preparation to be accomplished unattended (e.g. overnight when the system would otherwise be unused). An example is shown in Fig. 17. See [12,13] for two recent papers discussing the use of automation on wafer based dual-beam systems.

In addition to these sectioning applications the dual-beam has been applied to a wide range of other applications. For example, SEM sectioning can be extended into 3D by saving each image slice as the sectioning progresses. The resulting data set can be reconstructed to give a 3D volume of the region of interest, which can then be "virtually" sectioned in any direction.

In addition, the attachments associated with SEMs can also be added. These have included analytical techniques, such as electron backscatter diffraction (EBSD), cathodoluminescence (CL), and electron beam induced current (EBIC), and a wide variety of different sample stages. For example, cryo and heating stages enable experiments and sample preparation to be carried out in novel ways – a cryo stage may reduce some types of beam induced damage, while a heating stage allows the viewing of melting or other thermally induced processes directly in the FIB or FIB-SEM.

Summary

In the last twenty years FIB–based systems have become ubiquitous in semiconductor labs. Whether being used to support process development on the next node, debug a brand new design, find defects on the production line, or analyze a customer return, the focused ion beam's attributes of localized nanomachining and imaging has proven invaluable. As the challenges provided by shrinking design rules and shortening design cycles have increased, so has the variety of configurations and specializations of systems using FIBs.

Acknowledgements

The authors would like to thank colleagues at FEI and in the wider FIB community for many useful and illuminating discussions over the years, and acknowledge FEI applications for providing several of the images shown.

BH would also like to acknowledge all his co-workers at Beam-It Inc, while his tenure there, for all the help, encouragement and experimentation we were given. Also at Lucent/Agere the author would like to gratefully acknowledge the encouragement and support by Jim Cargo, Frank Biochhi and John Delucca. Also acknowledged are R Pajak, R. Ashmore, Lynette Ansorge, and Bruce Griffiths (Agere Systems) for all the help and several fruitful discussions. Special mention must also go to Peter Carleson, FEI Company, and Prof. (Emeritus) Jon Orloff, University of Maryland and FEI, for all their help and support for the past several years.

References

1. *Introduction to Focused Ion Beams: Instrumentation, Theory, Techniques, and Practice,* eds. L.A. Giannuzzi and F.A. Stevie, Springer, New York, 2005.

2. J. Orloff, L. Swanson, Lynwood, and M. Utlaut, *High Resolution Focused Ion BeamsFIB and Applications,* Springer, 2002
3. Private communication, Peter Carleson, FEI Company, Hillsboro, OR.
4. SRIM is available from http://www.srim.org/
5. J. Orloff, L.W. Swanson and M. Utlaut, "Fundamental Limits on Imaging Resolution in Focused ion Beam Systems", J. Vac. Sci. Technol. B14, 3759 (1996).
6. R. J. Young, J. R. Cleaver, H. Ahmed, Journal of Vacuum Science Technology, B 11(2), March/April, 1993, pp 234-241
7. D.W. Niles and R.W. Kee, "Success! > 90% Yield for 65nm/40nm Full-thickness Backside Circuit Edit", Proc. ISTFA 2010, 348
8. K.N. Hooghan et al., "Integrated Circuit Device Repair using a FIB system: Tips, Tricks and Strategies", Proceedings of the ISTFA 1999
9. A. N. Campbell and J. M. Soden, "Voltage Contrast in the FIB as a Failure Analysis Tool", Electronic Failure Analysis desk reference, 1998, pp. 129-135
10. J. Benbrick, G. Rolland, P. Perdu, M. Casari, R. Desplats " Focused Ion Beam Irradiation Induced Damages on CMOS and Bipolar Technologies," Proceedings, ISTFA'98, pp. 49-57
11. K. Chen, T. Chatterjee, J. Parker, T. Henderson, R. San Martin, H. Edwards, "Recovery of Shifted MOS parameters Induced by Focused Ion Beam Exposure", Proceedings of the 40th IRPS, pp 194-197
12. S.B. Herchbein et al., "Semi-Automated Full Wafer In-line SRAM Failure Analysis By Dual Beam Focused Ion Beam (FIB)", Proc. ISTFA 2010, 113
13. H.H. Kang et al., "High Volume and Fast Turnaround Automated Inline TEM Sample Preparation for Manufacturing Process Monitoring", Proc. ISTFA 2010, 102

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Circuit Edit at First Silicon

Ted Lundquist and Mark Thompson

Introduction

Recent statistics indicate over 60% of designs miss production at first silicon [1]. For a design to make it to market a mask revision is required. The majority of design misses is due to functional or scan failures typically from timing and voltage marginality [2]. First silicon debug then is of critical importance. The requirement is to determine what design errors or marginality have caused the failures and what needs to be done to change the design. Financial resources forbid going to second silicon without a "proven" level of validation. Internal circuit editing* provides this level of validation.

The 1989, Melngalis review of FIB applications was the first to include circuit edit [3]. Previously FIB edit had only addressed yield issues of IC fabrication masks [4]. The concept of IC circuit edit ("discretionary wiring") however had been proposed earlier to improve financials when IC yields were low [5]. Electron beam technologies were proposed for this discretionary wiring. Lasers, used in production to "wire" or edit ICs by opening fuses, have also been used for editing and recently lasers have been used to remove (or "kill") certain transistors from circuitry [6]. However, the productivity and flexibility provided by FIB circuit edit makes it the choice for first silicon circuit edit and is the focus of this chapter.

Sometimes a design issue is sufficiently identified through test that, straight away, circuit edit can be used to validate the proposed "fix". In other cases internal device measurements are required to identify the fault. Internal measurements with mechanical probes and photon emission microscopy (PEM) have been used for many years in identifying gross failures, which hinder I/O testing. Frequently during initial debug, FIB provides access to internal voltage busses for mechanical probes by bringing a probe point to the surface, even through silicon on flipchip devices. FIB editing can also open a window in top-level metallizations so PEM can be utilized for lower metal localization. Whenever a failure seems too gross (either excessively high or low Iddq) it might be beneficial to use FIB to cross section one corner where via stacks are frequently placed. If the via stack matches the vertical CAD then no gross mask errors occurred within the interconnects.

Most debug is still done on packaged samples and the majority of these are still die-up wire-bonded and editing from the front side is done. For flip chip and wire-bonded die-down packages edits are done through the silicon backside. Wire-bonded die can be packaged and repackaged to enable through silicon editing [7].

Emulation of the Mask/Fabrication Process

The ultimate goal of circuit editing would be to rework the silicon to obtain the performance equivalent deliverable through a mask/fabrication cycle—emulation. In other words reworked silicon should be equivalent to re-fabricated silicon. Today's circuit edit FIB is a "Fab In a Box"—it contains a focused ion beam for direct write lithography; process chemistries to selectively etch various dielectrics and metals and to deposit dielectric and conductor; endpoint detection for process monitoring; and imaging feedback on process quality. As in the fab, each step contributes to the final success (yield).

To accomplish such silicon re-fabrication the universal FIB tool of the 1980's has evolved into today's dedicated circuit edit tool. (Application specific FIB instruments have also been developed for front-end process monitoring, defect review, TEM sample preparation, micro-machining and general failure analysis.) Although almost all FIBs can "cut and paste", it is obvious that complex edits require optimization and thus the need for tool specialization. In addition to the consistency of the specialized edit tool, a great deal of skill and concentration on the part of the FIB operator is needed. FIB operator creativity is amazing.

Science behind Circuit Editing

The major milling function in FIB circuit edit is sputtering (a momentum transfer process whereby secondary atoms and molecules are ejected due to the impact of ions or energetic neutrals). Sputtering depends on several factors such as surface topography, primary mass, primary energy and surface binding energy [8]. Through chemistry the surface binding energy can be altered to increase the etch rate or decrease it to

* ISTFA is the premier venue for circuit edit.

deposit new interconnect materials. However for circuit edit, ion bombardment has some inherent negative artifacts such as grain dependent sputtering (a copper issue), implantation damage and charge accumulation (drift and gate oxide issues). Energetic primaries can also eject SE, which are used to generate an image by scanning the focused ion beam as is done with electrons in an SEM (Scanning Electron Microscopy). The major difference between SIM (scanning ion microscopy) and SEM is that the ion beam is always removing what is imaged/impacted. Another benefit in circuit edit is that the SEs provide a sensitive method for vertically endpointing an edit operation due to SE yield dependences on voltage and materials.

The ion beam chemistry processes are essentially activated by the dissociation of precursor molecules whether chlorine, bromine, iodine, xenon di-fluoride, tungsten hexa-carbonyl or more complicated precursors. These FIB processes, similar to fab related chemical vapor etching and deposition processes, involve three steps:

1. Physi-sorption of precursor molecules;
2. Fragmentation of precursor into etch/deposition and daughter by-products;
3. Desorption of etch and daughter by-products.

Chemistries are delivered by diffusion due to simplified control and delivery. The chemistry is locally injected onto the region of the FIB edit so as to minimize the gas load into sensitive ion column optics.

Chemistry Assisted Etching: SiO_2

Silicon oxides and nitrides are quickly etched using XeF_2. The F-Si bond is very strong and upon activation by the ion beam and damage to the SiO_2 surface the increased etch rate can be > 5x. Sputtering is not an equilibrium process and neither are the enhanced chemistries associated with it; for example, only ~1.5 molecules of XeF_2 were needed to assist the ion etching of SiO_2 [9]. Without XeF_2, upon exposure Al is redeposited over neighboring passivation. With XeF_2, whatever redeposition there might be is in a non-conductive form as determined by voltage contrast. Further when using XeF_2 less Ga^+ current is needed and yet the etching time is shorter.

Chemistry Assisted Etching: Aluminum

With Cl, Br or I chemistries, Al etches quickly and cleanly—when a line is cut, resistance is negligible >1GΩ. Barrier material is etched very rapidly with XeF_2.

Chemistry Assisted Etching: Copper

Transition to copper metallizations has challenged FIB circuit editing because copper is so chemically different than aluminum. The FIB exposure of copper readily reveals crystallite structures through ion channeling contrast. Cu tends to be fine grained, but Cu sputtering is very grain orientation dependent. Some orientations etch very quickly while others etch slowly. Another problem comes about when copper is exposed to halogen chemistries. Exposure to halogens followed by exposure to ambient results in dendrite growths associated with the hygroscopic nature of copper halides, excluding CuF. The conclusion then is that the use of halogens to accelerate copper etching must be eliminated or at least extremely well controlled [10]. XeF_2, does not effect the milling rate of copper. To protect exposed Cu it needs to be sealed either with conductor or insulator. Cu by itself sputters very fast so cuts show a lot of redeposition, which results in poor electrical isolation.

Acceleration is nice but not critical; what is important is selectivity. If a chemistry slows down the milling of Cu and slows down the milling of SiO2 even more it may have value. Lower k dielectrics are not yet standardized, but selective chemistries are ready [11, 12].

One issue with copper technology not generally emphasized is top level planarization as part of damascene processing. This means there is no surface topography to assist navigation, increasing the value of CAD navigation.

Selective etch of copper over dielectrics—SiO_2 and FSG—is illustrated in Fig 1. Voltage contrast imaging is used to verify that a cut has been successful [13]. Lower FIB induced voltage contrast implies cut resistance is $\geq 10^{10}\ \Omega$ [14].

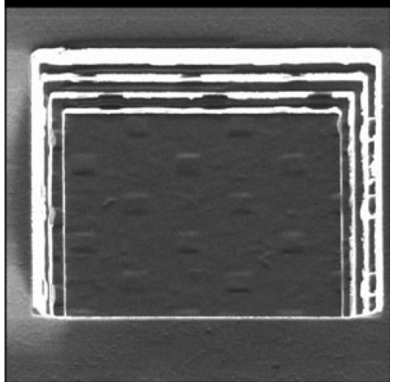


Figure 1. Selective etch of copper through 4 power planes, results in negligible texture from dielectric fill.

Chemistry Assisted Etching: Low-k Dielectrics
The texture of the lowest K materials is porous and "spongy" yet the composite device formed with these materials must be structurally stable under both chemical-mechanical polishing and thermal processing. XeF_2 accelerates the FIB milling of these materials [15]. There is no new issue for FIB enhanced milling with XeF2 of these materials, but for endpoint the nano-scale pores result in extensive micro-topography. This topography enhances SE emission making endpoint to metal more difficult. This materials issue is not expected to be as difficult as copper and etching should be controllable.

The lowest-k dielectrics will be organic with no silicon. With these there sometimes may be a false endpoint signal due to C buildup/Ga implantation but that has negligible effects as it equilibrates very early in the FIB etching process. Depending on the specific organic, a different chemistry might accelerate FIB milling even more than XeF_2 does [15]. FIB milling of the organic low K dielectrics creates voids (SEM beam energies above 2 keV creates voids as well) [16].

Chemistry Assisted Deposition: Conductor
For Pt depositions the precursor is trimethyl-Pt-methyl-cyclopentadienyl [17]. For W depositions the precursor is W(CO)6 tungsten hexa-carbonyl [18]. The balance between precursor flux and ion flux is critical to quality conductor deposition. Fig. 2 is a qualitative graph of this relationship between resistivity ρ and precursor-flux to ion-flux ratio. Note that when chemistry flux is too low compared to ion flux, milling occurs and when chemistry flux, is too high relative to ion flux the precursor is not fragmented sufficiently [19], i.e., carbon content is high and therefore ρ increases.

Increasing beam current density reduces the deposition resistivity, probably due to increased integrated SE density. On the surface where space is relatively unlimited, time is the critical productivity factor in circuit edit. For a via the lowest resistivity, however, should be used.

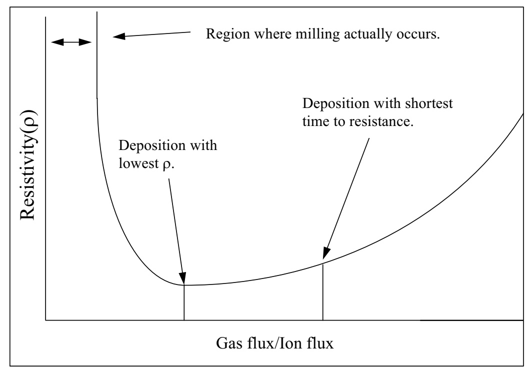


Figure 2. Graph of resistivity of FIB deposited conductor versus chemistry flux to ion flux ratio.

Chemistry Assisted Deposition: Insulator
FIB deposited insulator is illustrated in Fig 3. Voltage contrast readily shows electrical isolation, which occurs both across and through the deposit.

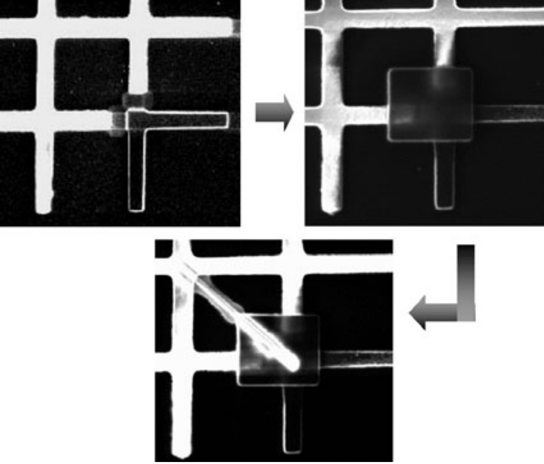


Figure 3. FIB micrographs illustrating insulator deposition. Top left shows FIB isolation of two legs of a test structure. Top right shows FIB insulator deposition covering the isolation cuts. Bottom shows FIB deposited conductor over the insulator deposition.

Insulator deposition is used for redeposition (Fig 4) and to enable edits through power planes and minimize rewiring lengths for balanced circuits. It is especially critical to the capping of copper cuts.

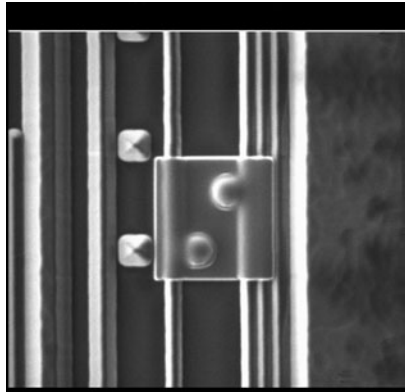


Figure 4. Repassivation over an edit area.

Edit Operations
On unplanarized top metal, a FIB edit is simplest because it is easy to navigate and accurately place the edit. Voltage contrast in the SE image shows when a metal cut has good isolation. Surface coverage can also be seen. Further, there are no issues associated with resistivity of the deposit (the deposit only has to be thicker to decrease its resistance) or milling a HAR hole.

Circuit edit operations are divided into two activities—Cut and Paste. Either a trace must be "cut out" or another trace must be "pasted in". Before an edit is started the correct internal circuit node must be located precisely. To accomplish this, CAD is generally required to locate the operation position. Prerequisites to edit steps:

- Remove polymer coating if present by
  - global removal
  - laser removal
  - chemically assisted FIB etching
- Align stage to CAD database layout
- Move to location of edit operation
- Fine align[*]

Accessing a trace buried under metallizations and microns of dielectric or silicon is challenging. Issues are accurate navigation for edit placement and end-pointing the trace of interest. Placement accuracy illustrated in Fig. 5 is only possible using CAD navigation. (Note: the contact diameter of the M2 via is 200nm.)

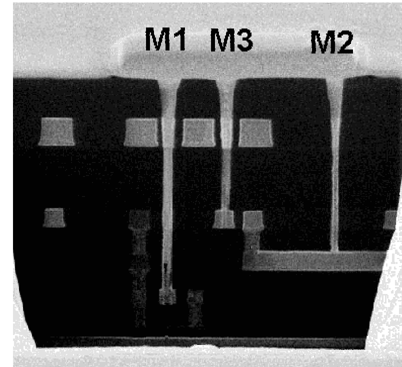


Figure 5. Cross section shows 3 FIB vias "drilled" and filled in a 6M device. Note cross sections are important in developing edit technique in a new process.

CAD is sometimes inaccurate, so operator needs to characterize each new device lot before editing. This exploratory surgery checks both CAD and device before the actual edit (See Fig. 6, showing edit site verification using a second device.)

Often times the operator suggests changes to increase edit success and the dialog between designer and operator can be in person or by FAX, which is illustrated in Fig. 7, which also shows a typical edit.

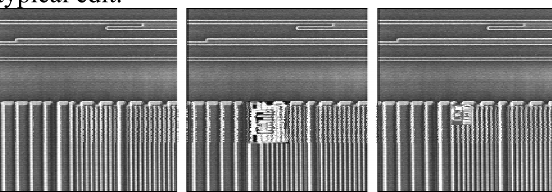


Figure 6. Left micrograph shows edit area. Center shows exploratory operation to verify the edit site neighborhood. Right shows M2-M3 connection, an M1 cut was outside this field of view.

The marked-up print supports identification of the operation on the CAD. Note: if this part had been copper, the vias to the cuts would have been filled with insulator.

[*] Edit placement and fine alignment are assumed in this discussion

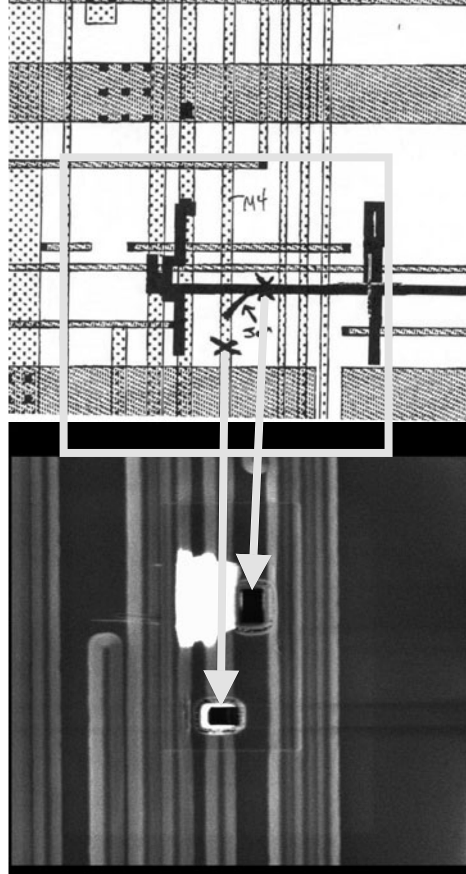


Figure 7. Circuit edit for red-lined ECO testing

To control structural stability for CMP processing, "dummy" fill metal is placed in open areas in the various metallization levels. When this "dummy" metal is part of the fabrication process and not part of design extra effort should be made to obtain its CAD before editing.

For success rate optimization certain steps need to be followed [20]. Table 1 gives these steps as well as the specific concern of the FIB operator.

Table 1: Order of Edit Steps

Step	Task	Goal
1.	Open Vias	End-point accuracy
2.	Fill Vias	Low resistance
3.	Connect Vias	Surface coverage
4.	Clean-up	Reduce leakage
5.	Open Vias	End-point accuracy
6.	Cut traces[1]	No re-deposition
7.	Deposit Insulator	(If Cu) Seal cuts

The first edit step creates an opening in passivation/dielectric to expose the trace. This is conceptually the simplest FIB edit operation—cutting out a volume of dielectric. To edit nodes buried under higher levels of metallization, it is obvious to first remove some dielectric to allow direct contact to a specific trace. This operation through the passivation and dielectric is required for a probe-point, whether e-beam or mechanical. There are two key concerns when exposing a trace: 1) that trace metal is actually exposed and not just thin oxide above the metal and 2) that only the desired trace or traces are exposed.

Exposing multiple traces within a localized area (local area depassivation) is advantageous when needing to probe several neighboring traces. Local area depassivation is only beneficial to device editing when two or more neighboring traces need to be shorted or cut.

Edit Operation: High Aspect Ratio Milling

An optimized HAR milling technique with good endpoint signal can obtain greater than 20-to-1 aspect ratio holes. (Note: Contact to Si substrate or poly-Si is more difficult than to Al or Cu because the SE endpoint signal is less.[*]) HAR milling has two basic requirements:

1. Accurately position a small hole to pass between closely spaced metal traces without exposing any other metal trace;
2. Adequate endpoint to detect the deep buried conductor;

To address 1, the use of CAD overlay techniques is almost mandatory. A SE detector, designed for maximum endpoint sensitivity to SE signals from HAR holes [23], can address 2.

HAR milling requires XeF_2 [9, 24, 25]. Without XeF_2 or some other chemistry assistance the best aspect ratio that can be ion milled is ~6-to-1. This is because some dielectric material sputtered from the bottom of the hole being milled is re-deposited on sidewalls; as the hole becomes deeper the percentage of material striking the sidewalls increases. This material must then be removed by another ion-surface event. So a steady state is reached—as the hole goes deeper, the sidewalls move out. With XeF_2 not only is re-deposition eliminated but also the milling rate increases.

Maximization of milling rate by XeF_2 occurs at a minimal ion-to-chemistry flux ratio. This means

[1] If a trace is cut before reconnected not only does it float but if it is a gate its V_t can shift [21, 22].

[*] Care must be used in contacting Si or poly-Si as Si is etched with XeF2 at a rate 10-100X greater than SiO2.

that when the chemistry flux to ion flux is insufficient ("gas starved") milling performance drops [25].

Edit Operations: Vertical Interconnects
One of the most important functions of the Circuit Edit FIB is laying down conductive material to rewire devices [26, 27, 28]. Total deposition resistance is the sum of contact resistance within each hole plus each deposit resistance plus the surface (lateral) deposition resistance (Fig. 8).

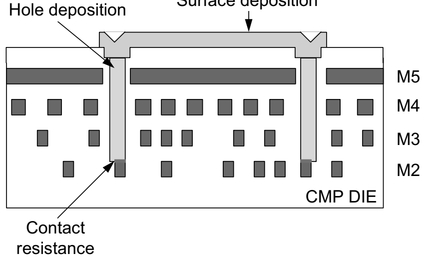


Figure 8. Schematic of an edit interconnect on a 5 metal copper device.

Connections to lower metallizations are made through HAR holes. The resistance $\mathbf{R_h = \rho t/A}$, where t is the dielectric thickness above the trace to be contacted and A the area of the hole to that trace. For minimal resistance A should be as large as possible, however, line spacings restrict A. Using the same process into a hole as for a lateral FIB deposition, resistivity can be unacceptably high. Undoubtedly this is due to processes within the deposition hole, resulting in a decreased metallic percentage. On the surface where space is not limited, a greater than optimized deposition rate gives good step coverage, but within a hole a low resistivity material is needed for good contact to lower metal.

Several factors contribute to a minimal resistance deposition. First minimize contact resistance to Cu, which is dependent on area. Therefore, the milled via to the metal levels should always be as large as possible. Redeposition of trace material onto the sidewalls of the via increases the contact surface thus reducing contact resistance. To achieve redeposition, an area 50% smaller than the via should be used to sputter the exposed trace within the via for about 15 seconds. (Note the resistivity of re-deposited Cu is much lower than FIB deposited conductors [14]). The deposit on the hole walls is ~25nm thick.

A deposition into a HAR hole can easily result in voiding within the deposit. When a deposition rate is too high, voids form. Therefore deposit into a 25% smaller box with 1/4 normal chemistry flux at 5pA/μm^2 and fill the entire via.

Even if there is no voiding the "stuff" does not necessarily have low resistivity. The solid angle for the organic carrier to escape the deposit is much less from a hole than from the surface; therefore, more of the non-conductive organic material is trapped inside the hole, which increases fill resistivity (Fig. 9).

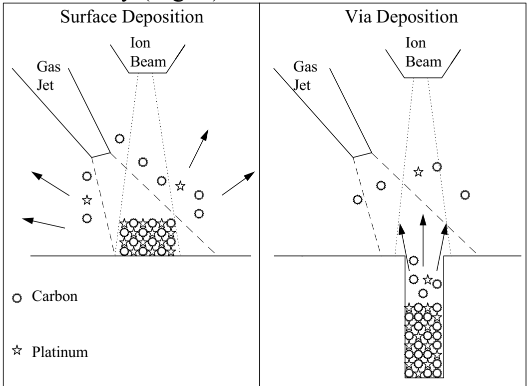


Figure 9. Surface deposition allows greater solid angle of escape for conductor's organic carrier, but within a via this escape angle is less so more carbon material is trapped within the deposition.

The chemistry arrival rate must be reduced as well as the beam current to produce a deposition rate that allows more time for the organic carrier molecules to escape. By varying vapor pressure and ion beam current, the deposition rate is slowed. This gives the organic carrier more time to escape the deposition. By reducing the chemistry and ion flux resistivity within a hole can be reduced 5-20x. To achieve this rate, the current density was 5 pA/μm^2 and the chemistry flux about ~10^{10} molecules/s/μm^2.

Edit Operations: Lateral Interconnects
The purpose of focused beam depositions whether ion, electron or laser is to create a defined, localized deposit--direct-write, maskless lithography. When making a surface deposition the resistivity ρ($\mu\Omega$–cm) of the deposition is not as important as is time-to-resistance, i.e., productivity. For a deposited interconnect, resistance is what is important and not resistivity. Resistance $\mathbf{R = 10^4 \rho l/wh}$, where l($\mu$m) is the length of the deposit, w(μm) is the width of the deposit, and h(μm) is the height or thickness of the deposit. (Note that when ρ increases (Fig. 2) with the precursor-flux-to-ion-flux ratio, it is actually h which is increasing and not necessarily the R of the deposit.) As productivity depends on

minimizing time-to-resistance, then with l and w more or less fixed by device layout, R depends on the variables ρ and h. The result is that the shortest time-to-resistance occurs at a ρ greater than for minimum ρ. This minimum time-to-resistance becomes beneficial for two reasons:

1. In the region of minimum time-to-resistance, the smaller slope of the resistivity curve (Fig. 2) results in greater tolerance of system variations in the precursor-flux-to-ion-flux ratio and
2. In the region of minimum time-to-resistance, the deposition thickness is greater, which overcomes problems with poor topographical coverage.

The resistivity of a surface deposition is important but the minimization of time-to-resistance yields the greater benefit—productivity.

The resistivity of platinum conductor deposited by FIB on a surface is from 70 to 200 μΩ-cm [29]. This is much greater than the typical resistivity (~2μΩ-cm) of a copper trace, the resistivity of platinum (~5 μΩ-cm) and even of Ga (~20 μΩ-cm). The platinum conductor is not pure; Auger analysis shows it to be ~50% platinum, ~34% carbon, ~15% gallium and ~1% Oxygen. Carbon (resistivity ~1400 μΩ-cm) is a remnant from the organo-metallic precursor carrier, the vapor phase of which makes possible the transport of platinum. Background water vapor is most likely the contributor of oxygen to depositions. In fact, the vacuum, which occurs just after venting and re-pumping, generally results in even poorer deposition resistivity. (Background chamber pressure below $1X10^{-5}$ T is important to forming minimal resistivity depositions.)

Deposition lengths are generally limited to the available ion beam defection field. However, this is not really an issue for long depositions as depositions can be "stitched" together. The real issue comes in the time it takes to produce a deposition with acceptable resistance. Typically for a 12nA beam current, a 1mm long deposition will take ~30min and have a resistivity of over 1000Ω. Sometimes when a long deposition must be performed, there has been concern as to the effect of the added capacitance; a typical maximum capacitance for a 1mm long deposit is less than 100fF. If this falls into the area of concern, special care must, therefore, be employed in routing the deposit as well as controlling the trace width and over-spray.

Edit Operations: Over-spray Cleanup

Implied within the primary edit goal of conductor deposition to electrically link two circuit nodes is the requirement to not link other circuit nodes. Edit depositions close to other edit traces or bond/bump pads must remain electrically isolated. This is not always directly achieved due to "over-spray", the deposition of conductor in areas other than the selected area (Fig. 10).

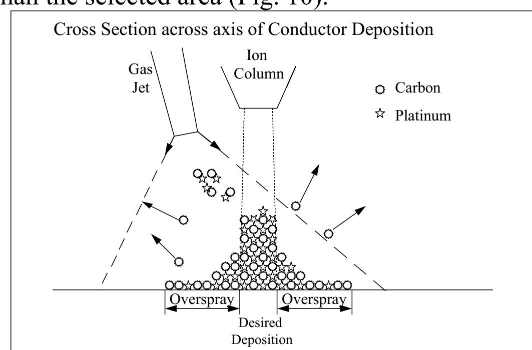


Figure 10. Clarification of "Over-spray"; note the region outside the desired deposition has a higher ratio of organics and, therefore, higher resistivity.

By removing or "cleaning-up" this over-spray, unwanted connections can be unlinked without removing the desired electrical link. Over-spray resistance is on the order of 30kOhm/sq between closely spaced deposits. (Note: Implantation of gallium the passivation results in a surface resistance of 1Gohm/sq; once exposed to air the gallium layer is oxidized in a few days and should not be an issue.) Besides conductance isolation, over-spray may need to be cleaned-up to reduce capacitance of the deposition.

The effect of ion beam tails is the limiting factor in how closely depositions can be placed. This occurs primarily adjacent to the desired deposition area and is due to the ion beam current profile extending beyond the edges of the desired area. As the ion beam current profile is smooth, this ends up rather thick near the desired deposition. (See Fig. 2) Over-spray in the beam tail regions has a resistivity characteristic of the right portion of the graph—where the chemistry flux/ion flux ratio is high. Lower current ion beams generally have smaller beam tails, this is because spherical aberrations decrease with beam current. A narrower yet taller deposition yields less over-spray but also decreased resistivity because of the beam density effect on deposition.

On Si oxide/nitride passivation clean up is best accomplished with XeF_2. Careful clean up can result in incredibly high isolation resistances—>

10^{14} ohms/sq. Monitor every clean up process on the video screen as clean up is very fast and needs to be stopped as soon as the bright over-spray deposit comes dark so as not to degrade the passivation. To minimize passivation degradation when more than one deposition is to be removed clean up of conductor depositions should be done after all depositions in a region are completed.

Edit Operations: Trace Cutting

During circuit validation, an edit is required in which a node must be electrically cut from circuitry. Edit success relies equally on all process steps leading to and including this "cut" operation. The cut should typically be the last operation performed to complete an edit. Cutting a trace, prior to completion of its associated vias and strap operations, can negatively impact edit success. Cutting out the node prematurely can cause end point issues, as the ground reference to the contacting node floats and charges up, and more importantly, this accumulated charge may cause an ESD event possibly damaging the device and effecting functionality.

Access to the trace to be cut is identical to High Aspect Ratio Milling described previously. The difference is in the dimension, as the entire metal line width must be exposed. To aid edit alignment (Fig 11), a good practice is to expose both contact and cut on the same line when practical, although visible in separate node access holes.

Figure 11. FIB image of backside node accesses through silicon and insulator to M1. Left site is to be contacted and right site to be cut.

For cutting aluminum traces, halogen chemistries are very effective in volatilizing sputtered material out of the hole, yielding a clean high resistance cut. In most FIB systems, the cut operation views the metal line going dark as it is cut. Within a HAR hole, imaging the metal trace is difficult, thus reliance is on a high sensitivity graphical endpoint to monitor electron yield over time. The electron yield drops as the metal is etched away, exposing underlying oxide. Where end pointing is difficult, it is possible to perform a cut based solely on time for a given metal thickness, beam current, and chemistry. This technique requires process characterization and system repeatability to get consistent results.

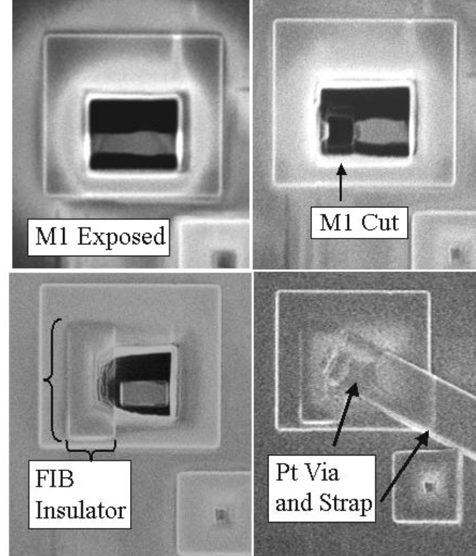


Figure 12. Backside edit images of a M1 cut before (UL) and after (UR), cut region protected with insulator (LL) and via and strap created to exposed M1-M2 via pad (LR).

Circuit layout constraints for a proposed edit may require that a cut and contact be attempted at the same node opening. Although not the preferred application, figure 12 shows such an edit. In this case, charging was not a problem due to both sides of the cut being tied back to ground at a prior edit step. If a cut and contact within the same access hole can not be avoided, minimizing beam current and depositing conductor material can assist in minimizing damage due to accumulated charge.

In the case of cutting copper power planes and traces, challenges with leakage due to redeposition of sputtered copper in the immediate vicinity is an issue. Because the sputter yield of copper is very high, re-deposition can induce leakage to near by circuitry. In applications where power copper planes must be removed to gain access to lower circuitry, both unassisted sputtering and gas assisted milling techniques are used [52,11]. As seen in figure 1, when cutting through multiple layers of copper, decreasing the opening of each subsequent lower layer minimizes connectivity between layers due to redeposition. This results in an inverted pyramid structure, and where redeposition occurs, can generally be minimized with an XeF_2 etch. Once the structure layers are electrically isolated from each other, a FIB

insulator can be deposited to passivate the exposed edges from subsequent etching steps.

When exposing smaller, single copper lines for cutting, it is important that ILD on both sides is sufficiently removed so that the full thickness of metal can be etched completely (Fig. 13). Without doing this, copper stringers or runners can remain, prohibiting effective isolation.

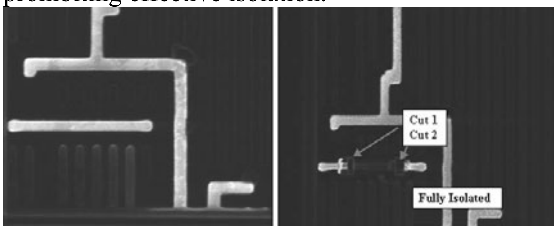


Figure 13. FIB depassivated copper traces and floating dummy metal. Right image shows isolated line following 2 cuts.

For cutting minimum geometry copper lines, XeF_2 can provide adequate isolation of the cut. Using XeF2 to cut copper over Low k layers can quickly expose underlying metal. FIB chemistries that protect Low k interfaces can be used [11]. Copper cuts, after cleaning, should be passivated with insulator deposition for protection and minimizing the possibility of corrosion.

Edit Operations: Reconstructive Surgery
Although the shrinkage of metallization line widths is a challenge for circuit edit, the greater challenge is multiple layers. In fact, because of the increasing number of metallization layers, it has become necessary to edit through the backside silicon. However, another method not used very extensively is to reconstruct the edit neighborhood to edit lower metallizations. The optical path from the passivation to the silicon surface is often completely blocked by traces, not even counting power-planes making up top metallizations at 9 levels of metallization. The only access to the underlying trace is to cut out and higher level trace rerouting it if only temporarily to make an edit. Reconstruction is possible with a FIB, but performance can be degraded as FIB processes are not as good as those of the fab.

Edit Operations: Backside editing:
To edit an IC through its silicon requires some new understanding. In regards to editing through the interconnection side, interconnections are essentially passive elements of the integrated active components and thus do not require more than high isolation resistance cuts and low resistance contacts. However, there can be cases due to ion beam interactions with the dielectric surface where charge is dumped into a floating gate and damages the gate as evidenced by changes in transistor parameters [30].

Editing through silicon results in new issues. Those most critical are those arising because the sensitive active areas must be crossed before the interconnections are reached [31]. Intel has reported that ~50% of the through silicon edits done there involve diffusions [32]. This is for 2 reasons, first the availability of the diffusions and secondly the diffusion density precludes some FIB operations without milling through diffusions. The reliability of through-silicon repairs has been tested through accelerated temperatures and voltages and found the edits to be robust for design validation [32].

Devices need to be thinned before beginning an edit. The thinner the remaining silicon the better the navigational resolution can be because transparency of shorter wavelength light increases as Si thickness decreases. Further, by reducing remaining silicon thickness detailed FIB operations become less time consuming. Experiments at several frequencies and voltages found no detectable performance changes [32]. Die thinning has evolved from the 100 μm barrier to below 30 μm [33].

The interconnection side of an IC is generally covered with a dielectric, which enables the direct rerouting of interconnections. The silicon side is not covered with dielectric. As the doped substrate is a conductor and part of the circuit, rerouting of interconnections across its surface requires a dielectric coating. This is not bad as the coating if properly tuned to a wavelength of light for forming an anti-reflective (AR) coating [34] actually improves the ease of navigation while it enables trace rerouting. However, a global ARC would need to be removed in the edit regions.

Because silicon is a conductor, vias through it must be lined with dielectric before filling with conductor. This is a complication that front-side editing only dealt with when editing through power planes. FIB assisted insulator deposition is therefore required for every through silicon [35].

Active regions must remain unchanged (unless that is the purpose). Exposure of p-channels and n-channels has been reported to render cells non-functional [36]. To terminate a signal the diffusion has been removed from either the output or the

power [32]. As spacing between active regions shrink, this goal of necessity becomes more challenging as already editing to M1, M2 and M3 is being done to avoid active regions. Note: it may sometimes be advantageous to actually change a transistor parameter by trimming a channel [32].

N-well size reduction may be required. As the n-well is a passive component to a transistor structure, changes to it ought not to effect transistor operation. As transistor density increases there will be increasing need to reduce the well diffusions. Unpublished work [37] has found that small reductions in an n-well dimension to enable an interconnect to M1 does not alter transistor characteristics. Livengood et al. has also reported success in this area [32].

Edit Operations: Probe-Point Creation
The main features of the probing method determine requirements on a probe-point. E-beam probe-point creation [38] is simpler than that for mechanical probing. For mechanical probing, the large physical probe size, relative to a trace, means that the probe point window in the passivation must be large.[2] or that a large pad must be created on the surface for touchdown (a 3x3 µm pad is recommended to guarantee probability). When contact to a lower level metallization is required, the signal must almost always be extended to a "surface". As the mechanical probe has finite impedance, care must be made to minimize the resistance of any connecting deposited conductor, especially if the probe is for current injection. In e-beam probing, resistance of a connecting deposit is not as important because e-beam's inherent impedance is effectively infinite. Further, for e-beam, the probe-point window does not even need to actually expose the trace as waveforms can be obtained capacitively through thin dielectric [39], but for mechanical probing the probe-point window must expose the trace to direct probe contact. Signal measurements with the e-beam tester rely on detection of electrons ejected. Accurate measurement of signal levels depends on the smoothness of trace exposure or deposition smoothness because electron yield depends on this [40]. On the other hand, always minimizing deposition resistance simplifies the debug process as e-beam probing sometimes gives way to mechanical probing for device debug/characterization.

For density restrictions on the die surface, the pad may need to be created to one side of the exposure window in this case it may be beneficial to shape the deposition topography to form a divot to trap the microprobe tip. Because the use of a divot in the probe-point pad minimizes the skating or slipping a probe tip experiences, a minimal radius probe tip with long conical taper can be employed. [Note: Once all probe-point pads are laid down in an area, careful attention to further aspects of the debug process needs to be made. Over-spray needs to be removed not only to eliminate leakage currents but also to assist in future debug steps. FIB conductor depositions are very resistant to the chemistries used to remove passivation globally; remnant over-spray debris potentially can ruin further debug steps on depassivated devices.]

Because of turn-around-time between mechanical prober and FIB, sets of probe-points and possibly several edits are executed on the FIB and then delivered to the prober. After set-up on the prober, measurements associated with each probe-point set and edit are obtained. From these measurements, feedback from the simulations, and as testing through the design proceeds, new lists of measurements are generated and the DUT is moved back and forth between the FIB and the prober. Obviously this process of going back and forth between the FIB and the mechanical prober is time consuming; estimates are that it takes several hours at best suggesting a value to having micro-probes in the FIB chamber [41].

Success Statistics

The following are from circuit edit service work and are typical for a good operator using good equipment (Fig. 14 & 15). It is important to note that even for 130nm process technology the success rate per job is >80% and that even at deeply buried metallizations the success rate is high. Success was determined only by feedback from customers, as no test programs were available in-house for these devices. When a job failed, most often it was repeated and then successful the second time, but this information is not expressed in the data.

[2] For both probe types it is depth from the surface, which drives the probe-point window requirement.

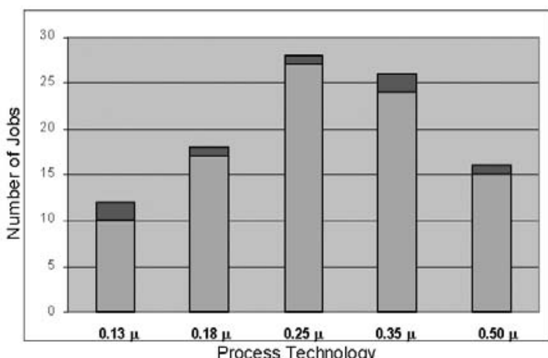


Figure 14. This graph shows, as a function of process technology, the total number of jobs that were completed in several months—100, those that had successful edits (light shading) were 85 and those that had no successful edits (dark shading) were 7.

FIB edit service tries to minimize turn-around-time with the goal being to provide at least one working device as only one is needed to provide a valid emulation of a proposed design change. So in each of the 100 "jobs" graphed in Fig 13 and 14 between 1 and 11 parts were generally received from the customer (3.3 on average).

For 130nm technology devices the individual edit success rate average was 90%. Of the 334 devices edited, 293 were successfully edited and 41 failed.

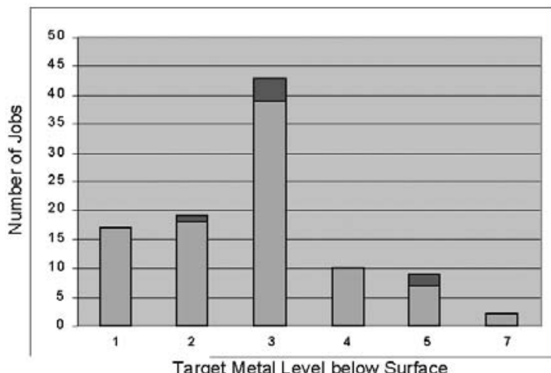


Figure 15. The same data as Fig. 13 only as a function of edited metallization below the surface.

Why Edits Fail

The number one reason for edit failure is in correct edit specification—the FIB operator was told to edit the wrong thing. Beyond this why do edits physically fail? Several things cause a FIB edit to fail: improper edit order, shortening to upper metallization when filling a HAR hole, excessive resistance of the FIB interconnect, cut resistance may be too low, Capacitance changes from long traces and/or probe-pad construction, etc.

Resistivity of FIB deposits is greater than that from the fab often by ~2 orders of magnitude. Across the surface deposit resistivity is not critical as traces can be made thicker, but within narrow vias this is an issue. When gates float, either intrinsic or FIB created, charge injection into gate oxides shifts V_t. (Shifted transistor parameters can be90% recovered and V_t reduced to <10mV by heating to ~425C [42].)

Design Rules for Diagnostics

To take full advantage of the internal measurements design rules are developed. These are ready for post silicon installation of FIB probe points needed to assist microprobe and E-beam measurements, as well as circuit editing [43]. Preparation at design not only can speed debug but in many cases enable it. For example, to aid navigation the incorporation of a "dummy" aluminum top layer on CMP dielectric speeds the editing and therefore debug of copper parts [44]. When providing instructions to the operator, in spite of a high success rate, it is still recommended to edit higher metal rather than lower metal.

Circuit editing enables a proposed "design fix" to be tested before going to fabrication. The value of circuit editing has been recognized for sometime as critical to getting ICs to market [45]. Circuit editing is not able to provide a complete emulation of the fabrication process, as some fabrication-inherent and edit-process issues exist. Following certain design rules however can maximize the emulation capability.

Moving forward, the fear is evident that edit success may decrease unless Design for FIB rules are implemented. These are a subset of what Design for Diagnostics includes. Rules for enabling access to probe points also enables access for editing. The number one rule already referred to is to bring up potential edit points to higher level metallizations or for flipchips and devices with more than 5 levels of metallization to bring potential edit points to M1. Methods to design for through silicon circuit probing [46, 47] can also be utilized for circuit editing [48].

With the increasing complexity of ICs, the actual device design layout becomes important to the success and turn-around-time of edits. Design for Test DFT has become the plan of record at most IC manufacturers. The incorporation of DFT into the design process not only saves Test time but allows greater flexibility in Tester usage and provides value to first silicon debug. In the same

manner, Design for Diagnostics [49], which includes Design for FIB saves development time by making edits easier. Implementing edit design rules not only enables confirmation before a mask change but can facilitate prototype production from first silicon [50].

The goal of the design team is one Mask/Fab cycle per design; however, this is not often feasible due design complexity and design/process interactions. Implementation of DFF rules reduces the number of Mask/Fab cycles per design. The realistic goal is 2 Mask/Fab cycles; by properly utilizing circuit edit through employment of DFD, this goal can be achieved. When the designer better understands tradeoffs in design rules, optimization of both silicon utilization and circuit edit occurs. Then products get to market faster and with less cost.

Conclusions

Circuit edit has become the standard first silicon practice among IC manufacturers [7, 8, 9, 20, 21 26, 31, 35, 38, 40, 44, 45,46] to prove a proposed design change before going to second silicon. Essentially circuit edit addresses the design yield issue—the design must be modified before usable devices can be fabricated. Any proposed edit must be proven before going to fabrication. First silicon is not generally edited to produce product, although there are exceptions especially to deliver prototypes. FIB editing remains a valuable part of the IC product development process because it can have a high success rate. In spite of increased complexity edit success remains high. Circuit edit processes are not foolproof. The FIB instrument has become a miniaturized fabrication facility—a "Fab In a Box". The same types of expertise which are applied in the fab needs to be utilized in the FIB, yet all the techniques are not as readily executed so the tool can not be used carelessly—its limits must be known.

For most of IC history, Al and SiO_2 dominated device interconnects. Now that copper has been standardized, which makes circuit edit more difficult, low K dielectrics are being introduced. The driving force for migrating to new materials is mainly performance—faster and less power consumption but also process cost reductions. However, these are only the beginning of many revolutionary process changes that are expected to require changes in circuit edit chemistries. In spite of high success rates designers must Design for FIB if the value of circuit edit is to continue.

Circuit edits can actually provide value to the next design. For example, the present design can be edited to improve its performance. Simulations make suggestions (informed guesses), proposed changes are implemented and then these are tested. The one that is best is then incorporated into the next design. As an example, pin drivers needed to work faster in the next design—performance needed to improve. Testing of the old design at higher frequency showed there were issues with the starting of certain ramps so the device was probed to find out where improved performance ramp start error was originating. Modeling of the probing results was inaccurate so several edits were performed to understand the source of the error and provide feedback to the model. To develop an accurate model, capacitance was added and subtracted, lines were re-routed to reduce coupling capacitance and components were removed. This rework enabled the next design's first silicon to work as planned.

Manufacturers monitor the cost of mask changes and require a "guaranteed" success with the next mask set. This cannot be achieved without proper emulation of the proposed mask/fabrication change. Mask costs [51] will continue to drive the need for IC circuit editing with FIB.

Just as for fabrication as the circuit edit challenge increases, the importance of understanding the issue the edit seeks to address will increase.

References

1 Electronic Business (June 2003)
2 Collett International Research (April 2002)
3 J Melngailis, JVST (1989) 469.
4 A Wagner, Proc SPIE 393 (1983) 167
5 CK Crawford, PE Kudirka, IEEE-Sym Electron, Ion and Laser Beam Technology (1971)
6 K Dickson, A Blakely, T Garyet, ASM-ISTFA 27 (2001) 465.
7 S Kolachina et al, ASM-ISTFA 27 (2001) 51.
8 K Nikawa, JVST B9 (1991) 2566.
9 HF Winters, JVST B1 (1983) 927.
10 T Ohuchi, TR Lundquist, VV Makarov, LSI Testing (2003) 107.
11 VV Makarov, WB Thompson, TR Lundquist, TT Miau, AVS-Inter Conf Miroelectronics & Interfaces 3 (2002) 115.
12 J.C. Gonzalez, D.P. Griffis, T.T. Miau, P.E. Russell, JVSTB 19 (2001) 2539.
13 H. Ahmed J Microscopy 139 (1985) 167.
14 J Melngailis, CR Musil, EH Stevens, M Utlaut, EM Kellogg, RT Post, MW Geis, RW Mountain, JVSTB4 (1986) 176.
15 H Bender, RA Donaton, ASM-ISTFA 26 (2000) 397.

16 JC Gonzalez, MIN da Silva, DP Griffis, PE Russell, JVSTB20 (2002) 2700.
17 T Tao, J Melngailis, US Patent 5,104,684 (1992).
18 T Kaito, T Adachi, US Patent #5,086,230 (1988).
19 RA Lee, TR Lundquist, ASM-ISTFA 22 (1996) 85
20 S Czapski, TR Lundquist, LSI Testing Sym (1995) 96.
21 KN Hooghan, KS Willis, PA Rodriguez, S O'Connell, ASM-ISTFA 25 (1999) 247.
22 J Benbrik, P Perdu, B Benteo, R Desplats, N Labat, A Toubol, Y Danto, Proc Int Sym Plasma-Induced Damage 3 (1998) 128.
23 L. Wang, D. Keckley, D. Bui, Rev. Sci. Instrum 69 (1998) 91.
24 H Ximen, CG Talbot, ASM-ISTFA 20 (1994) 141.
25 LR Harriot, JVST B11 (1993) 2012.
26 J. Mohr and R. Oviedo, ASM-ISTFA 19 (1993) 391.
27 T Tao, J-S Ro, J Melngailis, JVST B8 (1990) .
28 T Tao, W Wilkinson, J Melngailis, JVST B9 (1991) 162.
29 J Puretz, LW Swanson, JVST B10 (1992) .
30 A Campbell, ASM-ISTFA 25 (1999) 273.
31 PF Ullmann, CG Talbot, RA Lee, C Orjuela, R Nicholson, ASM-ISTFA 22 (1996) 381.
32 R Livengood, P Winer, J Giacobbe, J Stinson, J Finnegan, ASM-ISTFA 25 (1999) 477.
33 Jim Colvin, private communication, 2004
34 B Davis, W Chi, ASM-ISTFA 26 (2000) 155.
35 R Lee, N Antoniou, ASM-ISTFA 24 (1998) 459.
36 R Ring, R Goruganthu, L Stevenson, EDFAS News (Feb 2001) 32.
37 M A Thompson, private communication (2000).
38 CG Talbot, in FA: Tools and Techniques, L Wagner editor (1998) 113.
39 W Baerg, VRM Rao, RH Livengood, IEEE-IRPS (1992) 320.
40 X Larduinat, D Kratzer, CG Talbot, Test & Measurement Europe (Summer 1994) 25.
41 J Brown, M DeSiltes, D Masnaghetti, CG Talbot, Microelectronic Engineering 31 (1996) 149.
42 K Chen et al, IEEE-IRPS 40 (2002) 194.
43 N Kuji T Ishihara, IEEE-Asian Test Symposium (2001) 179.
44 T Malik, private communication 2004]
45 K Nakamae, H Sakamoto, H Fujioka, IEICE Trans. Fundamentals E77-A (1994) 698.
46 RH Livengood, IEEE-ITC. (1999) 877.
47 RH Livengood, US Patent 6,020,746 (2000).
48 RA Lee, P Wolpert, A Pedneker, ASM-ISTFA 27 (2001) 285.
49 W Lee, IEEE Design &Test of Computers (June 1989), 36.
50 R Flores, EDFA 6 (2004) 6.
51 D. Hutcheson,The Chip Insider™ (August 15, 2002).
52 JD Casey et al, JVSTB 20 (2002) 2682

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

The Process of Editing Circuits Through the Bulk Silicon

Nicholas Antoniou
FEI Company, Peabody, MA, USA

Introduction

For about 20 years the semiconductor industry has enjoyed a short cut to product introduction through the use of focused ion beam technology (FIB). With FIB, an integrated circuit can be modified to fix design flaws or to fine-tune its performance. In a matter of hours a design change can be implemented and tested in silicon instead of waiting for the fabrication of silicon with this change.

Technological advances in semiconductor manufacturing and assembly have presented difficulties in performing FIB based circuit editing. Flip chip packaging was the first serious obstacle to present itself and it was around 1995 that the first flip chip editing FIB was introduced. This system was capable of editing circuitry through the bulk silicon. The proliferation of flip chip technology has not advanced as fast as expected but other advancements have made access to critical nodes of the circuit from the front side very difficult. A design engineer can use as many as 9 or more layers of interconnect, most if not all, in copper. The interlayer dielectric can be made of a low-k material and filler pattern is used across the chip and at most layers to even out (planarize) each layer. Finally, the top few layers of interconnect are often used for power, ground and clock distribution and are therefore wide and thick buses. To the failure analyst, the front side is now almost inaccessible. Emission microscopy and circuit editing are all moving to the backside where access to the circuit is becoming easier than the front side. Once through most of the bulk silicon one has an unobstructed view of the active circuit where all the switching takes place.

However, now more than ever, cooperation between the design, package technology and test groups is critical to the successful and rapid debug of circuits. IC's are very complex and the materials can be difficult to deal with but a small amount of planning at the design phase can drastically improve the success of a product.

Background

The key technologies needed to edit a chip from the backside are:

1. Die decapsulation technology
2. Die level silicon thinning and polishing equipment
3. Localized silicon thinning capability (Laser Micro-Chemichal and/or FIB)
4. Infra-Red microscopy for navigation through silicon
5. Endpoint technology for the control of bulk silicon removal
6. High quality direct write dielectric and metal technology
7. Deep submicron via creation

Items 3, 5, 6 and 7 are typically found in FIB systems. Items 1 and 2 are typically stand-alone systems and finally infra-red microscopy needs to be integrated into as much of the process as possible. Today, IR microscopy is available inside an FIB system and this is a key differentiation between a front side and backside FIB system. Other differentiations are bulk silicon removal capability with endpoint and blind navigation through CAD with the proper rotation and mirroring of the layout.

The process flow

Editing a circuit from the bulk silicon side requires some preparation and steps that are not needed in more traditional front side editing (Figure 1). The bulk silicon has to be exposed and preferably thinned and coated with dielectric.

1. Thinning the bulk silicon

The bulk silicon is exposed and mechanically thinned to about 100 microns of remaining silicon. The surface has to be polished so that IR imaging can be achieved through the bulk. A polishing wheel works well for flip chip devices or when the device is not in a cavity. For all other cases a specially made mechanical drill system is used. There are several companies that offer such systems today.

2. Polishing and cleaning the bulk silicon surface

The surface of the bulk silicon has to be smooth and free of debris. Imaging through the silicon for navigation is improved if the surface of the bulk silicon is polished to a mirror finish. The polishing and cleaning are important not only for imaging purposes but also for the subsequent local thinning. If a laser micro-chemical system is not available, residue from the clean will result in a trench as that shown in Figure 2.

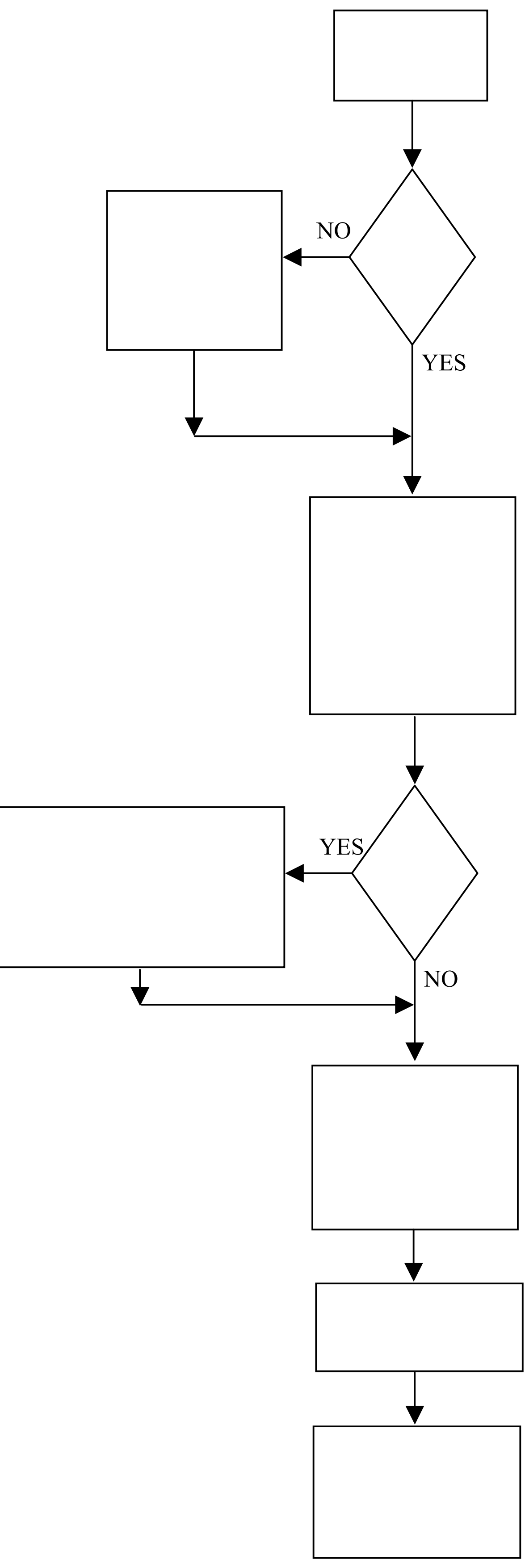


Figure 2. The process flow for backside editing

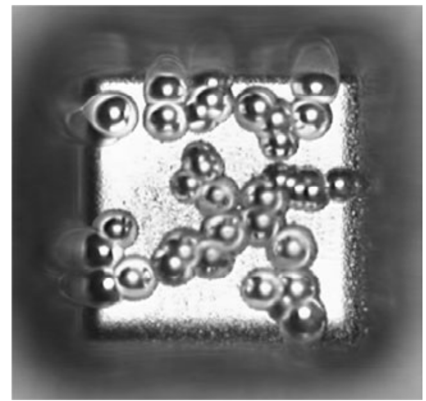


Figure 2. Surface contamination obstructing the local removal of silicon.

3. Navigation and Trenching

In backside editing, navigation is mostly blind to the FIB imaging system. The scanned ion beam imaging cannot obtain images through the bulk silicon. CAD navigation is critical as is IR microscopy. An IR microscope is used to image through the silicon so that alignment marks can be found (Figure 3). Once these alignment mark are found one can directly located and a trench created as outlined below.

To improve imaging through silicon, anti reflective coatings are sometimes used. These coatings are easily applied but add time to the local thinning of the silicon.

4. Trenching with Laser Micro Chemical Technology

Using Laser Micro-Chemical (LMC) processing, trenches of 1 mm x 1 mm or larger can be opened to about 10 microns of remaining silicon [1].

Alternatively, the FIB can be used to open trenches of a few hundred microns on a side and down to just above the active circuitry. Optical beam induced current (OBIC) endpoint is used to reach the wells of the active circuitry. OBIC is used in both the LMC and the FIB system. In the OBIC endpoint system, the chip power and ground are monitored and a laser is used to induce a photo current in the chip. This photocurrent is measured externally and is the OBIC signal. As the silicon is locally thinned, the OBIC signal rises exponentially (Figure 5). The value of the current can be calibrated to a thickness or one can wait for a transition in this signal.

This transition (drop in current) occurs when the space charge region between the well and substrate is reached. A layer of silicon just above the well is all that remains at this point [2]. Typically the total amount of remaining silicon at the transition is about 3 microns for advanced IC technologies.

Other techniques used for endpoint include passive voltage contrast [3] and optical thin film measuring techniques. Passive voltage contrast (Figure 4) can be used under certain well to substrate conditions.

5. Dielctric Deposition

Once the area of interest is locally thinned, dielectric deposition is used to passivate the locally thinned silicon. Most FIB systems produced for circuit editing provide adequate dielectric deposition capability.

At this point in the flow, the process of circuit editing is very similar to that used on the front side. There is a thin layer of material to go through to access metal lines for re-wiring or cutting. The only significant difference is that part of the material to go through is a conductor (the thin layer of silicon remaining). This thin layer has to be passivated with FIB deposited dielectric.

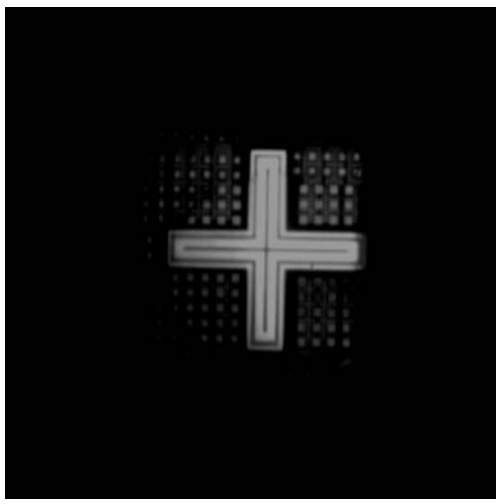


Figure 3: Alignment mark located in the IR microscope through bulk silicon

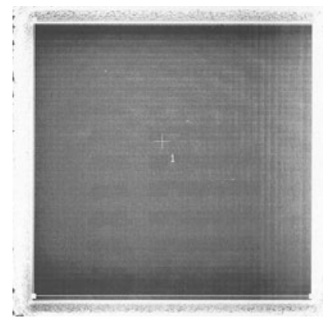


Figure 4: A locally thinned area with the well areas appearing as voltage contrast.

6. Fine Navigation

Once the trench is passivated, high accuracy navigation is needed. Either the optical microscope or the FIB can be used for this fine navigation. Two scenarios are proposed to achieve fine navigation. Fiducials are found in the optical microscope (Figure 3) and used to map the die to the CAD. Three fiducials are sufficient for CAD navigation.

a. Use of a local alignment mark

If there is a feature near the edit area that can be used to improve navigation then the overall process of navigation can be accelerated. The edge alignment marks are found via the IR microscope and the CAD aligned to the chip. A local alignment feature is then exposed and used as an additional locking point. This can dramatically improve the accuracy or navigation and is described in Figure 6.

Another advantage of the backside approach is that some ways to modify the circuit become available that were not available on the front side approach. These techniques involve removal of part of the active area of a transistor to adjust its current drive [4]. Another technique involves making contact directly to the active area of a transistor for re-wiring [5].

b. Exposure of edge alignment marks

Another approach is to expose edge fiducials so that they may be used by the FIB system for precise deep sub-micron navigation. In this approach fiducials are exposed and then the CAD aligned in the FIB imaging system using the CAD for reference (Figure 7). The area of the edit is located using the CAD navigation and trenched to a few microns of remaining silicon.

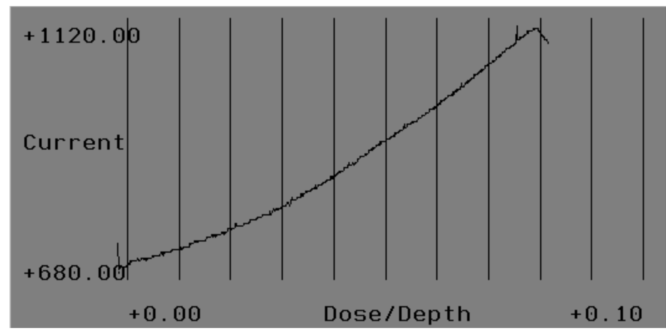


Figure 5: OBIC endpoint plot.

7. Editing the circuit

Once the CAD is matched to the chip, the node of interest is located in the CAD and its coordinates sent to the FIB. A mark is created in the FIB for the location of this node.
At this time, a circuit feature is now in a known and discernable location and can be exposed, cut or contacted as is typically done in FIB circuit editing (Figure 8). The only additional step needed is to passivate vias in silicon if they will be contacted. FIB deposited dielectric performs this quite adequately.

Areas of concern

Several issues have arisen with the advent of backside editing. Solutions are either available or being worked on as

this is written. These issues are listed below so that those who perpare to do backside editing can consider these.

1. Die distortion

The assembly process can cause die to distort. This can then present problems in precise navigation. Solutions have been already implemented but are mostly procedural. A local alignment can help avoid this problem as can the use of the voltage contrast image to locally align the CAD image to the voltage contrast image.

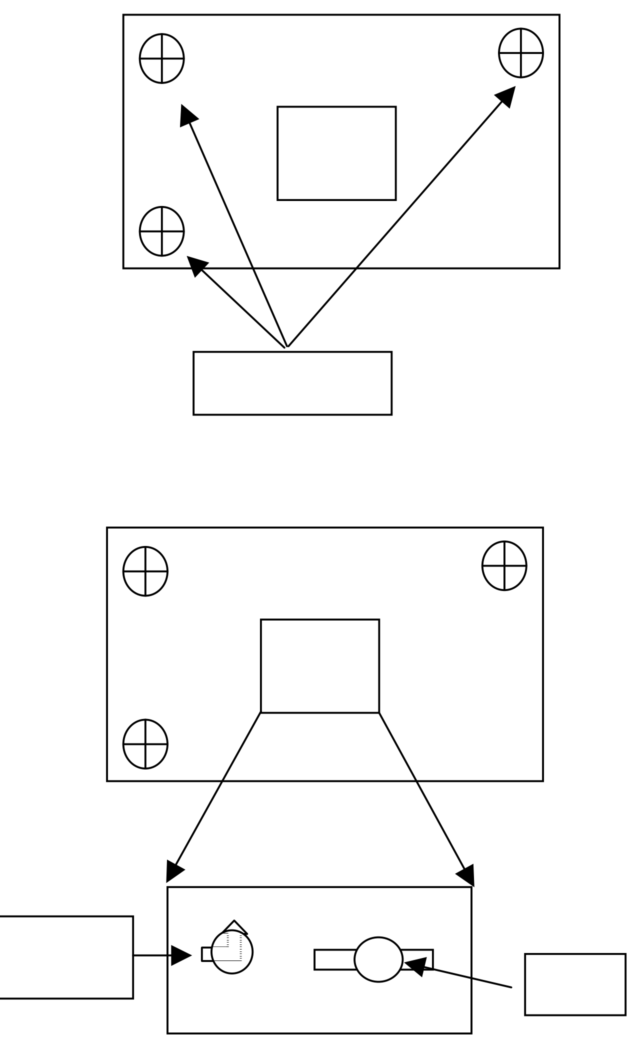


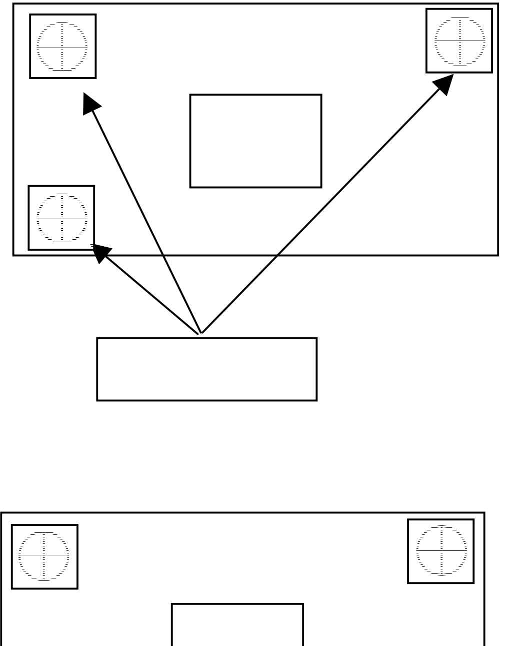
Figure 6: Process of aligning using an IR microscope and a local alignment mark.

2. Enpoint in silicon

Even though techniques are available and reliable (OBIC for example) they require some preparation. In the past, interface cards were used for E-Beam probing. These interface cards are still needed for signal acquisition and emission microscopy but thought has to be given to make these useful on several tools.

3. Silicon on insulator

This technology is expected to make endpoint in silicon easier and in some cases it has. However, the bulk silicon is electrically isolated in this case and care must be given to the grounding of the part. Also, some of the techniques of making SOI material introduce crystalographic defects that might interfere with backside editing.



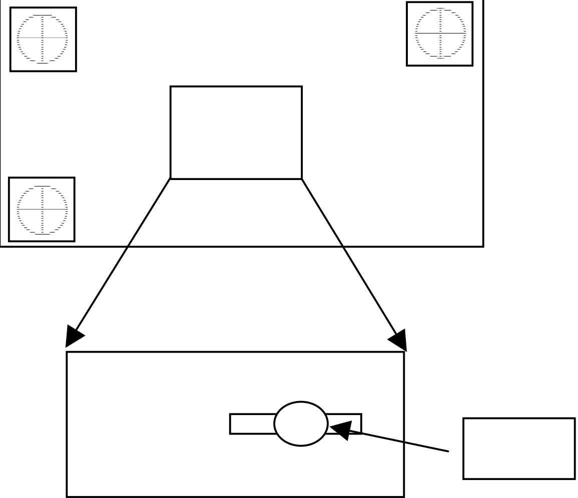


Figure 7: Navigation using exposed fiducials

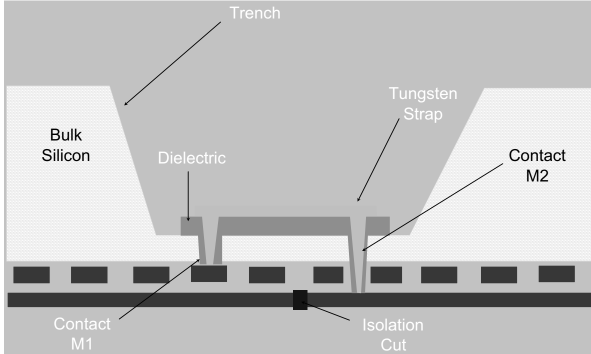


Figure8. Schematic of edit through bulk silicon

4. Charge control

The following is offered to reduce charge effects.

a. Use lower beam current
b. Coat the sample with a conductive layer such as carbon
c. Use charge neutralization
d. Ground the sample
e. Make isolation cuts last

5. Desing Rules

The idea of implementing design rules for FIB has been discussed at meetings and is beginning to gain momentum. If a designer can allow access to a node that is critical to the operation of the chip then the debug can be facilitated. Design rules can be easily implemented in the design phase and broken where necessary. The additional space required to allow access tocritical nodes, especially from the backside, is not terribly demanding. One can think of the access to a node as being an upside-down wedding cake where the access area has to increas as the depth increases.

References

1. S. Silverman et al., *OBIC Endpointing Method for Laser Thinning of Flip_Chip Circuits, Proceedings of the 24th ISTFA.,* 461-464, (1998)
2. Rama R. Goruganthu et al. , *"Controlled Silicon Thinning for Design Debug of C4 packaged ICs", IRPS Proceedings, 37th Annual IRPS*, pp. 327-332, March 1999
3. C. Boit, K. R. Wirth, E Le Roy, *Voltage Contrast Like Imaging of N-Wells,* Proceedings of the 29th ISTFA, 331-337, Santa Clara CA, November 2003.
4. R. H. Livengood, P. Winer, J. A. Giacobbe, J. Stinson, J. D. Finnegan, *Advanced Micro-Surgery Techniques and Material Parasitics for Debug of Flip-Chip Microprocessor*, *Proceedings of the 25th ISTFA*, p 477-483, November 1999, Santa Clara, CA.
5. R. Lee and N. Antoniou, *FIB Micro-Surgery on Flip-Chips From the Backside, Proceedings of the 24th ISTFA*, p 455-459, November 1998, DFW, TX.

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Education and Training for the Analyst

Christopher L. Henderson
Semitracks, Inc.
Albuquerque, NM USA

Introduction

Analysis of semiconductor components has become an increasingly complex task. Today's analyst is called on to find "a needle in a haystack." Every year, the needles get smaller, the haystacks get bigger, and the customer wants it found faster. With this in mind, how should we train our analysts to perform this daunting task? There are several major thrusts our education/training efforts need to take if we are to be successful analyzing modern semiconductor components. These thrusts include: process, technology, cross-training, and techniques. I will discuss the history of training activities in the failure/product analysis discipline, and describe where this area is heading.

Historical Perspective

In 1966, Fairchild Semiconductor introduced the quad, two-input NAND gate (see Fig. 1). Soon afterwards, Texas Instruments, Motorola, National Semiconductor and others introduced their own lines of standard logic parts. In the late 1960's much of the integrated circuit development was performed for the U.S. military. At this time, the military also began a push to increase the reliability of its systems. This brought about the need for analysis capabilities to understand failure mechanisms associated with these integrated circuits. As these parts grew in popularity, the requirement for failure analysis grew. During the late 1960's many companies began taking the first steps toward establishing what we now know as a failure analysis laboratory. The failure analyst was born.

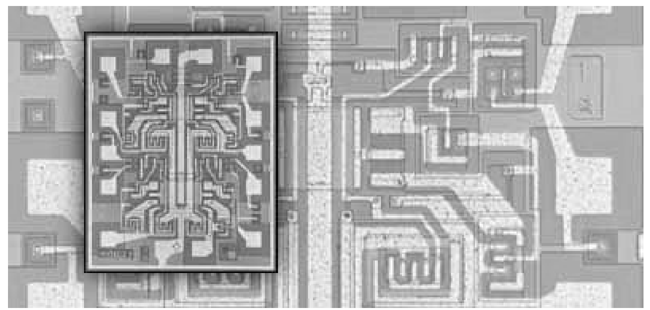


Fig. 1. Photograph showing the Fairchild Semiconductor quad, two-input NAND integrated circuit (photo courtesy Fairchild Semiconductor).

In the late 1960's, much of the failure analysis work was done on field returns. Systems companies returned devices that failed to the manufacturer for analysis. This meant that the analyst served in a reactive mode to problems the customer might experience. The semiconductor manufacturers produced components in relatively low volumes; they also manufactured the same components over a long period of time. This meant that the analyst could feed back information from field returns to the manufacturing line in time to make an impact. The process engineers could then make changes to the manufacturing line that would correct the problems seen by the customer.

In the late 1960's, product engineers performed failure analysis for the most part. The product engineer knew everything there was to know about their product. For example, a competent product engineer knew the schematic, recognized the layout, could map between the schematic and layout, knew the processing steps, understood the mask levels, knew how to test the part, knew the packaging process, and understood the customer's requirements (from regular visits and/or telephone conversations).

Quite often, failure analysis could be accomplished with a curve tracer, some package decapsulation tools and techniques, and an optical microscope. More subtle problems required a scanning electron microscope (SEM) and possibly energy dispersive spectroscopy (EDS). The investment in failure analysis tools was modest. $250,000 bought all of the tools that were necessary to do most any

analysis job. More importantly, the investment in skills and training was modest. A competent electronics technician could perform and manage all aspects of the analysis.

In the late 1960's there was no training program for analysts. Public training courses, like those offered by Bud Trapp, John Devaney, and Howard Dicken, were still a decade away. Training occurred "on the job." The analyst talked to the process engineers to understand the process, the designers to understand the layout, schematic, and test procedures, and the packaging engineer to understand the packaging process. The rest of his or her training came from experience. Good analysts were typically individuals who had years of experience and kept good records that allowed them to look for trends and recurring problems.

Analysis – The Current Situation

Today's analyst faces a much different situation than the analyst of the late 1960's. While we still have standard digital logic circuits like the quad, two-input NAND gate, we also have integrated circuits that exceed 10 million gates. In addition the bipolar silicon circuits of the late 1960's, we have a bewildering array of technologies, such as complementary metal oxide semiconductor (CMOS) circuits, BiCMOS circuits, gallium arsenide circuits, indium phosphide circuits, silicon carbide transistors, gallium nitride diodes, complex heterojunction structures, and microelectromechanical systems (MEMS). We also have a variety of design complexities, including discrete components, digital circuits, analog circuits, memory circuits, MEMS devices, radio frequency and optoelectronic components. An example of the circuits analysts no face is shown in Fig. 2.

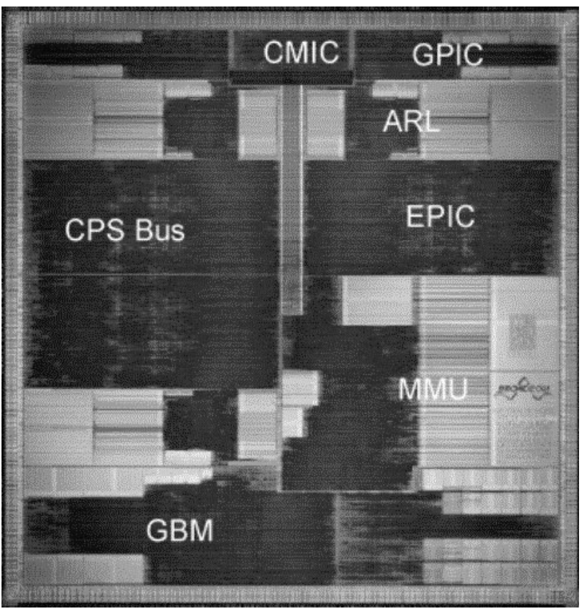


Fig. 2. Photograph showing a wire speed 3-7 switch system on a chip containing in excess of 60 million transistors (photo courtesy Broadcom).

Today's analyst also faces more complex equipment sets. In addition to the curve tracer, optical microscope and decapsulation tools, the analyst must be familiar with a variety of electrical testing hardware, endless electrical fixture configurations, x-ray and acoustic microscopy, electron beam tools, optical beam tools, thermal detection techniques, the focused ion beam (FIB), the scanning probe/atomic force microscope, and a bevy of surface science tools. Many times, the analyst must make do with a limited set of tools, as the cost of purchasing these tools exceeds the budget of all but the most lavishly funded operations.

Today's analyst also faces enormous pressures from the customer. The customer wants an answer to the problem immediately. Even so, there may be little, if any, ability to impact the manufacturing process. Because product cycle times are so short, many components are manufactured with a few wafer runs and are not manufactured again. The analyst's company may not even manufacture the device; it may be sent to a foundry for fabrication. The analyst may be doing activities other than analyzing field return failures. He or she may be performing design debug, yield improvement, or qualification analysis.

Today's analyst must know a staggering amount of information and be able to lucidly process that information to successfully guide an analysis through to completion. While the analyst of the late 1960's could get by with a basic understanding of design, test, packaging, and some analysis tools and techniques, today's analyst must know a whole lot more. We are expected to understand the correlation between the design and layout of multi-million transistor

circuits, even though the design engineers can't perform that task (all they know anymore is VHSIC Hardware Description Language (VHDL), Register Transfer Language (RTL), and other abstraction languages). We are expected to understand how to test complex integrated circuits, even though the test engineering community cannot figure out how to adequately test these devices. We are expected to have non-destructive depackaging techniques available for an ever-increasing number of package configurations, even though the packaging engineers cannot successfully rework these configurations. We have to understand obscure properties like the Franz-Keldyish effect, the Seebeck effect, and escape peaks.

The task of performing failure analysis is daunting. Over a career, an analyst is likely to experience many of these issues listed above. Education and training play an important role if the analyst is to be successful in his or her work. How do we prepare analysts for this environment? Below, I give an approach that can help the analyst succeed in this seemingly impossible environment.

Philosophy of Training

An effective training program requires that we start with the end in mind. Therefore, we should ask ourselves what the characteristics of an effective analyst are. Once we understand their characteristics, we can develop an education/training program to develop these characteristics. I have listed some of the characteristics that make a good failure analyst.

- Clear, logical thinking
- Does the job right the first time
- Organized (both in actions and in documentaton)
- Has a good understanding of semiconductor technology (design, process, test, packaging, reliability, etc.)
- Understands the information a tool or technique can yield
- Can operate a variety of analysis tools

I have ordered the list in the order of importance for training. The top items on this list are things that are unchanging as technology advances. The bottom items tend to change with respect to time. Techniques change and evolve over time, while the actual tools tend to change most quickly. Ideally, we would like to ensure analysts understand and practice the upper items in the list. Once they understand those upper items, the lower items on the list will become easier to teach. Many analysts (and their managers) become fixated on particular tools or techniques. In a sense, this is putting the cart before the horse. Analysts who spend all their training time learning and re-learning the latest tools and techniques do not gain an appreciation of how their contributions fit into the big picture. Many times, they don't even make the transition to the ability to guide an analysis through from beginning to end.

Education and training for the analyst can be broken down into three areas: analysis process, technology, and technique training. Because failure analysis requires a broad set of skills and knowledge, companies need to be willing to invest more time in education and training than might be required for other areas, such as design, process, and packaging.

Process Training

Clear, logical thinking is essential to the analysis process. We need to understand up-front what the customer wants as an outcome. This will help guide the overall process. One approach to this clear, logical thinking process is outlined by Ferrier in this past year's International Symposium for Testing and Failure Analysis proceedings[1]. There are also a number of philosophical principles that are outlined by Henderson, Cole, Barton and Strizich in the 21st Century Product Analysis Manual [2]. Clear, logical thinking, coupled with an understanding of the technology, allows the analyst to develop a process flow for successfully analyzing the device of interest. Every analyst should be proficient in this area.

The analyst also needs to adopt the attitude that he or she will do it right the first time. Doing the job right the first time may seem to require more time, but the opposite is true. The president of your firm would much rather see 10 yield problems thoroughly addressed and solved, than 50 yield problems partially solved. In the first case, these problems are unlikely to surface again, while in the second case, the problems are likely to resurface in the future and cause more grief. This requires that the analyst, and especially his or her management, put back-pressure on over-demanding customers who want results yesterday. It is hard to apply clear logical thinking to an analysis problem when the customer is breathing down your neck

The analyst must be organized in their work. Each analyst should maintain a laboratory notebook and document the analysis as it progresses. Most analysts perform more than one analysis at a time. While many analysts can keep the details of a single analysis straight in their minds, six or eight analyses is a different story. Good documentation allows the analyst to quickly generate a report at the end of the analysis. The analyst should generate a written report for every significant analysis finding. These should then be archived in the corporate knowledge based for future referece. Every analyst should be proficient in writing up reports. If an analyst is weak in this area, he or she should receive training.

There is considerable debate in the analysis community as to whether good analysts are born with their abilities (genetic), or successfully learn their skills. While genetics may

certainly play a role, we need to be training people in this area.

Technology Training

All analysts should have a basic understanding of the circuit technology they analyze. While many engineers and technicians graduating from college should have a basic knowledge of circuit design and processing, one cannot assume this is so. Analysts have a variety of different educational backgrounds including: electrical engineering, physics, chemistry, and materials science. Because of reorganizations, layoffs, and other events, there may even be analysts with other backgrounds. Even within electrical engineering, some specialized in design, while others specialized in processing. Furthermore, many analysts may have not understood their classes the first time (or paid attention, for that matter). Analysts should be encouraged to take basic semiconductor technology courses to introduce them to the subject or refresh their memories.

By their nature, analysts must be generalists. They need a moderate depth of knowledge in a variety of subjects. To accomplish this, analysts should be encouraged to learn design, processing, test, packaging, and reliability. The more an analyst can understand these fields, the better they will be able to perform their jobs.

Technique/Tool Training

Finally, analysts need to understand the techniques and tools required to successfully analyze the devices manufactured by their companies. First, analysts need to understand the techniques. This would include an understanding of electrical test techniques and methods, package level techniques, delidding/decapping techniques, backside sample preparation techniques, optical and scanning electron microscopy, electron beam techniques, optical beam techniques, thermal detection techniques, scanned probe techniques, the focused ion beam, and analytical characterization techniques. Important, but often overlooked, is the knowledge to choose substitute techniques, implement techniques cost-effectively, and how/when to contract for techniques.

There is a misconception that “hands on” training will solve all of our problems. For example, I can train an analyst to be proficient at running a light emission microscope system, but without further knowledge, the analyst will not know how properly set up the electrical stimulus to the part, what to do if he or she is not getting results, how to interpret the results, and what to do next. The analyst needs training to understand all of these issues. He or she therefore needs education in design and test, device physics, and failure analysis process. Furthermore, the analyst will need to be retrained when his or her company purchases a new light emission microscope system. Finally, the analyst should know how to perform light emission on a variety of platforms (turn-key systems, charge-coupled device cameras mounted to probe stations, standard optical microscope in a dark room, etc.). This is not adequately accomplished with “hands on” training.

The Future of Analysis Training

As our world moves ever faster, analysts are in need of “just in time” information. Although it would be nice to provide six weeks of training each year for our analysts, we do not have the time or the budgets to do so. This is one area where the internet can help. Companies like Semitracks are developing internet-based course and reference material that can be viewed at any time. The material is broken into small “chunks,” allowing analysts to learn during short breaks (equipment pumping down, data acquisition, lunch, first thing in the morning, end of the day, etc.). The material contains video, graphics and textual data (see Fig. 3). Moreover, the material covers a wide range of subjects, including design, process, packaging, test, reliability and technology, in addition to the subject of analysis.

Conclusion

Analysis of semiconductor devices has grown increasingly complex. Analysts now require extensive, continuous education to successfully perform their jobs. Analysts should concentrate on clear logical thinking, for this underpins the discipline of FA. Analysts also require technology-based training for understanding semiconductor fields such as design, process, technology, packaging, reliability and test. Finally, analysts need to understand the techniques and tools used for analysis, with the caveat that an overemphasis on tools can actually reduce the effectiveness of the analyst. Internet-based training may help alleviate some of the time constraints placed on the analyst. Training and education will play an increasingly critical role in the success of the analyst.

References

1. S. Ferrier, “A Standardized Scientific Method Formulation for Failure Analysis Application,” *Proc. Int. Symp. Test. & Failure Analysis (ISTFA)*, Nov. 2002, pp. 341-348.
2. C. Henderson, E. Cole, D. Barton, M. Strizich, *21st Century Product Analysis Manual*, published by Semitracks Inc. 2001-2004.

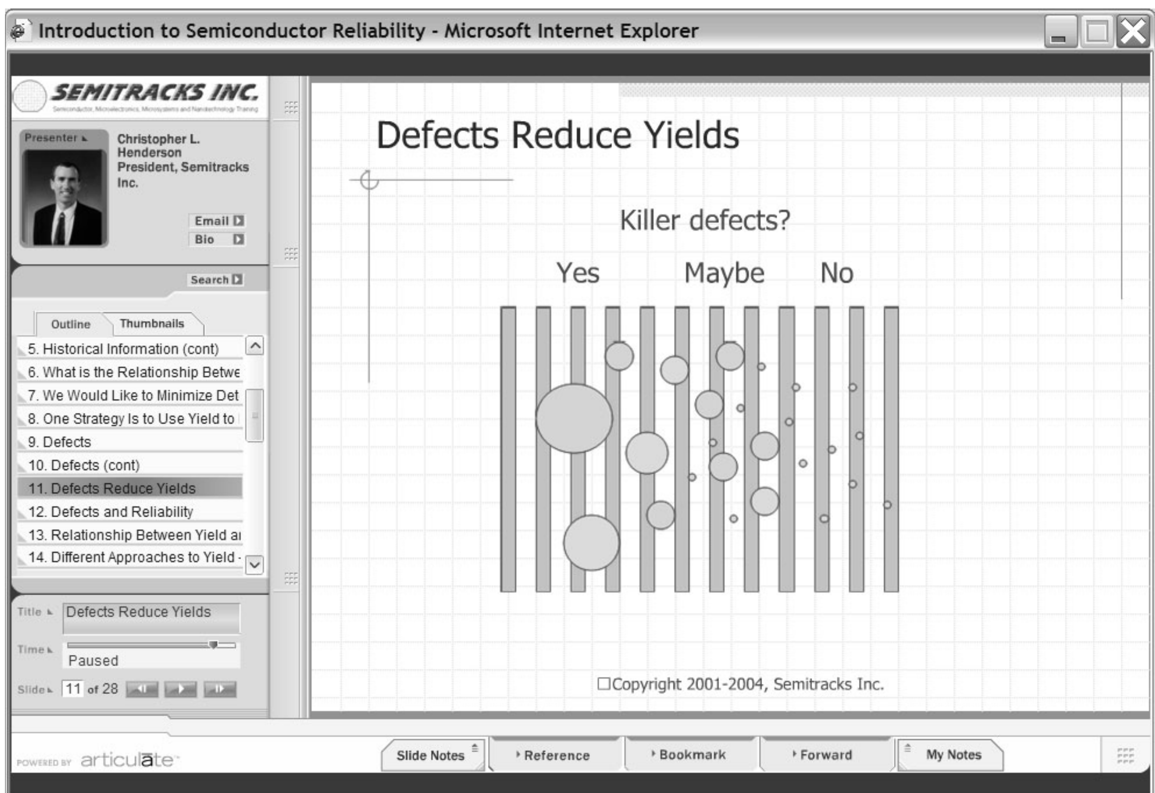


Fig. 3. Screen shot showing an example of internet-based training (image courtesy Semitracks Inc.).

Microelectronics Failure Analysis
Desk Reference, Sixth Edition
R.J. Ross, editor

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Management Principles and Practices for the Failure Analysis Laboratory

Richard J. Ross
IBM Systems & Technology Group
Essex Junction, Vermont

Introduction:

Failure Analysis (FA) of electronic components is a highly technical activity requiring highly skilled personnel, increasingly complex (and increasingly costly) equipment, and the development of increasingly sophisticated techniques and methods to discern the location, nature, and root-cause of defects causing non-conforming device operation. The devices to be analyzed may range from discrete, passive components to ultra-high density integrated circuits (ULSI). In all cases, the operation and development of a Failure Analysis laboratory need to be managed with an eye toward cost-effectiveness, customer satisfaction and future challenges. The purpose of this article is to stimulate die reader to consider the various aspects of Failure Analysis laboratory operations and their respective business management requirements. While the general focus is primarily based on semiconductor integrated circuit (IC) operations — particularly those of integrated device manufacturers (IDM) — the general ideas and operations can be applied to a wide variety of other laboratory configurations sod settings. References for further reading and examples of resource materials are also included.

Outline

FA operations and management requirements can be divided into the following areas:

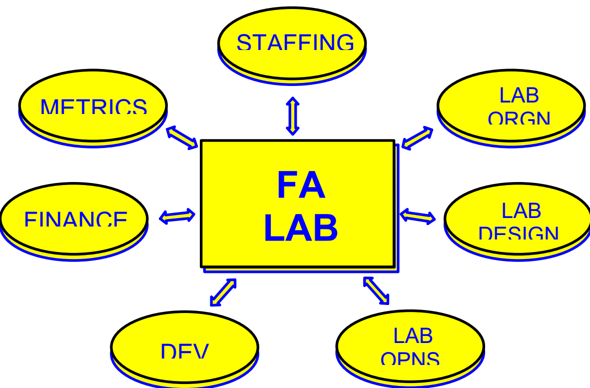


Figure 1: Outline of FA Lab Management

Staffing:
The staffing function is a key, if not the key management function No amount of complex and clever equipment cost efficiencies or fancy graphics-heavy reports can replace dedicated motivated skilled engineers and technicians The staffing function can be divided into the areas of Recruitment, Retention, Training, and Skills Mix.

Recruitment:
Recruitment of FA personnel is often a tricky business. A good analyst should have a working knowledge of device physics, circuit analysis, manufacturing processes, electrical characterization techniques, and chemistry. Further, (s)he must hove strong oral and written communications skills- be a team player and be able to manage the stress which conies from multiple tasks and multiple customers Above all, analysts require a strong technical curiosity.

Job Role: FAILURE ANALYSIS ELECTROPHYSICAL ENGINEER
Principal Academic Field: ELECTRICAL/ELECTRONICS ENG/MAT.SCI./PHYSICS

JOB ROLE DESCRIPTION

Engineer responsible for providing technical direction to and/or participation in an analytical team which verifies, diagnoses, isolates, identifies, and characterizes to root cause defects in semiconductor VLSI processes and/or products. Interacts with customer (inside and/or outside the division/company), manufacturing, development, and qualitypeers and management to insure timely and accurate response to concerns and influence appropriate corrective actions. In-depth knowledge of semiconductor device physics, integrated circuits, as well as strong communication skills required. Employee communicates results to customers.

Significant technical guidance to technicians is required

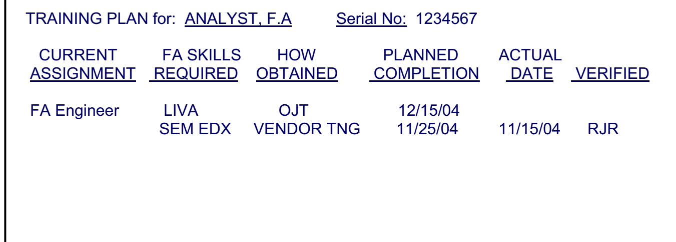
Figure 2: FA Engineer Job Description

Where, then, does one go to recruit such people? The simple answer is that one *can't*. Engineers and technicians from good schools with degrees in Electronics, Chemistry, Physics, or Materials Science can be found, but there are very few of these new graduates who will have the full set of requisite skills. Another source is to find experienced design, process, or characterization engineers/technicians from within the industry or within the company, but

this leads to a fairly rapidly diminishing list of the same key prospects. The ultimate answer is to start with the newly minted engineer/technician as “raw "material" and, by carefully applying time, patience, and mentoring, grow them into successful and competent analysts. This then, leads us to...

Training:

How do we take these new, eager people and turn them into failure analysts? There are few academic programs teaching the art of FA. Several workshops and seminars such as those held in conjunction with The International Symposium for Test & Failure Analysis (ISTFA) and/or those given by a number of consultant firms exist and these are very valuable, however, even these can only cover a broad overview of the skills necessary to become a full-fledged analyst. There are also self-paced reading materials — books (such as this one) and multimedia reference works, which are excellent references, but again, can oily provide broad guidance The most prevalent and successful training methodology on-the-job training (OJT) under the guidance of an experienced, highly skilled FA mentor. The model is similar to the "see one, do one, teach one" technique used in medical training. This requires to addition to the mentor noted above, a clear training plan and well-documented procedures. OJT cannot be unplanned or haphazard – such an approach can easily lead to chaos, disillusionment of the trainee, and frustration for the trainer. Goals must be established and documented, training processes established, timelines set, and, above all, management oversight must be on-going and supportive. A specific job description needs to be in place and the mentor (experienced analyst) as well as the trainee (new analyst) needs to be fully aware of the expectations and goals. Workplace distractions will also need to be minimized. Several independent estimates of the time required to transform an engineer/technician into a qualified failure analyst indicate that the process takes 18-24 months and may cost up to $250,000 including salary, benefits, materials, etc. Further, this need for training will continue (albeit at less intensive levels) for the duration of the analyst's career. The electronics industry especially the IC industry, is moving at a rapid rate —so much so that each year much or what we know and the tools we use become at best outmoded and inefficient, and at worst, completely obsolete. Tools and techniques exist today that were essentially non-existent in FA ten years ago (e.g. Backside Focused Ion Beam, xIVA, Scanning Probe Microscopy, Time-Resolved Photon Emission Imaging) Simultaneously, materials and process complexities are being introduced which make current methods ineffective (e.g. Cu metallurgy, Flip-chip packaging, Silicon-Germanium, Low-k dielectrics, sub-white-light optical pattern dimensions). The need to grow and *learn* is endless. One can think of this as Moore’s Law for Failure Analysis – the techniques you use today will be obsolete in 24-36 months. With this need and the cost and time involved, a major issue becomes...



TRAINING PLAN for: ANALYST, F.A Serial No: 1234567

CURRENT ASSIGNMENT	FA SKILLS REQUIRED	HOW OBTAINED	PLANNED COMPLETION	ACTUAL DATE	VERIFIED
FA Engineer	LIVA	OJT	12/15/04		
	SEM EDX	VENDOR TNG	11/25/04	11/15/04	RJR

Figure 3: Sample Training Plan

Retention:

It is an acknowledged fact that retaining skilled FA personnel continues to be a challenge across the industry. The average "lifetime" of a failure analyst is less than five years. Many young analysts get the impression that FA is not a leading-edge career path like design and development. Some feel that FA has low corporate visibility and/or that it is not good that the nature of the mission is dealing with "failure". On the contrary, I would make the argument that FA is a leading-edge career path. It is continuously varying, a challenge to the mind, an opportunity to master multiple disciplines, a way to keep up with the newest technology advances, and, we get to play with some really *cool* toys! It all comes down to the fact that people need to be recognized for their contributions. The negative impact of turnover is not limited to lost expertise to the group. Indeed, stress on the remaining analysis personnel associated with transition, wasted efforts and incomplete projects and loss of social networks all impact the work force. The fully-loaded cost of continuous turnover includes the constant need to train new people (see above), the churn in customer interaction, the reduced efficiency of a constantly changing workforce, and the time spent by managers in location, recruiting, and hiring new people. The team needs to be built and management needs to insure that the FA team shares in the successes of the overall organization. Rotational assignments, certainly within the FA organization, if it is a large one, are useful – even more so if rotation between FA and design,

process engineering, and even technical marketing can be accommodated. Attendance at outside conferences and workshops are also key ways to, reward and provide incentive to analysts — and to accomplish some training as well. Also, never forget the positive effects of just saying "thanks"...

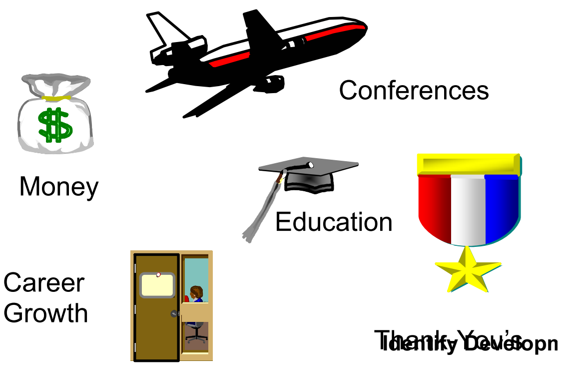


Figure 4: Recognition/Retention Techniques

Skills Mix:
In a given setting, the "mix" of professional to technical personnel (engineers to technicians) is dependent on several factors. Product mix and complexity, task(s) to be performed, tool set automation, education/experience level, and proximity to design support all factor into the equation. In the IC world, typical engineer/technician ratios run from 1 : 2 -- 1 : 4 in SRAM/DRAM to 3 : 1 — 1 : 1 for custom logic or microprocessors. The use of computer-aided diagnostics techniques such as Level Sensitive Scan Design (LSSD) can reduce the Logic ratios by up to a factor of three. Characterization FA operations in a manufacturing environment can typically run with several skilled operators working under the supervision of an FA engineer or technician to evaluate test structures or defect-density monitors. The key factor here is that one size does not fit all -- the operation needs to be tailored to the task and setting.

Laboratory Organization

The organization aspect of FA Lab Management includes the reporting structure, the guiding management philosophy, the decision to operate as "generalists" or "specialists", and, especially significant in today's global economy, how to handle support of a world-wide business operating environment.

Reporting Structure:
Who "owns" the FA operation depends in large part on the type and size of the overall business unit. In a large corporation, especially an IDM, the FA organization may report to corporate or division quality/reliability. An individual fabricator (foundry) or design laboratory (fabless) may have its FA operation linked to its local quality/reliability operation. Manufacturing and/or development functions may claim the rights to FA on their specific product sets. Or, the FA operation may be an independent entity -- clearly for independent, contract FA labs this is the norm. Again, there is no one, right model, however, in actual practice, among IDMs, the most common reporting structure is for FA to be contained in the quality/reliability function.

Management Philosophy:
Many texts in management deal with the possible structures and philosophies of management of technical areas. The three most common FPA management philosophies are the traditional functional organization, the matrix organization, and, a new and empowering option, the self-directed team. The traditional *functional* organization works best with less experienced personnel in a single, task-oriented environment with fixed tools and established procedures. The *matrix* organization gives a mix of task and customer orientation, which allows the shifting of resources quickly to meet tactical needs, but can give rise to loyalty conflicts between the “owning” and the “using” organizations. The newest organizational paradigm, the *self-directed team*, gives the employees more control and empowerment over their assignments and procedures. It requires high levels of maturity, expertise, and trust – both among the employees and between management and employee. There is an up-front team-training requirement to develop die team-building process and all factions must buy into the process from the start. When this is working, it provides the maximum in flexibility and, often in productivity.

On the technical side, the organization of the lab can be segmented according to the various product types serviced (memory, logic, processor, analog, discrete, passive…), by technology (bipolar, CMOS, mixed..), or by customer set (automotive, data processing, consumer, communications...). It could also be divided by the source of the work (qualification, reliability, returns, yield support,…), or by FA process, (test, electrical characterization, fault localization, deprocessing, inspection,…). Each of these methods has advantages and disadvantages, and each may work in a given setting. The key challenge is the optimization of (usually very limited) analytical resources to

maintain operational flexibility and customer satisfaction.

Specialization versus Generalization:
Analysts may be organized by task or by program supported. That is, they may be assigned one specific task or task set linked to key tools or they may be assigned to the complete analysis task from start to finish. *Specialists* may be used to maximize critical expertise, to optimize key techniques, or optimize tool performance. They may become well-known experts is their specific field and perform their specific task in a very efficient manner. The main disadvantages to specialization include the risk of boredom or "burnout", the potential that the job may have little variation or flexibility, and the risk of a tool or technique becoming obsolete and thereby exposing the individual connected with it. *Generalists* have the advantage of very great variation in job tasks on a day-to-day basis. Flexibility is optimized as resources are more easily shifted from one job or product to another. The analyst feels a higher sense of ownership of the job from start to finish and gets a greater sense of accomplishment when it is completed. The mains disadvantages of generalization include the risk that there will be less than optimal results from a specific tool or technique, the potential lack of "experts" in one specific field, and an increase in training time as more operations need to be learned.

As is usual in FA, there is not a hard and fast rule as to what is the best mode of operation. Lab resources may dictate one model, as there may not be enough people to have single procedure experts for all procedures. A good compromise is to use specialists for tools/processes where optimum performance is critical (e.g. Scanning Probe Microscopy, FIB Device Modification, Materials Analysis tools, etc.) and generalists for broader applications (e.g. deprocessing, electrical characterization, inspection, etc.). As always, the key is maintaining optimal flexibility and customer satisfaction.

World-Wide Operations:
Today, all markets are global. Industry-wide, revenue and profit growth will increasingly come from non-US sources. The demand for wireless communications is worldwide as the cost for developing economies to build huge landline infrastructure will be exorbitant as compared *to* the cost for wireless relay stations. Silicon foundries and build, assembly, and test (BAT) operations increasingly are spread across geographies. End products are shipped around Ito world. Alliances and partnerships are proliferating will, many of these now multi-national in composition as the cost of development and fabrication facilities soars. All of these worldwide customers want FA answers and want them *now*. Both IDMs and silicon foundries often have multiple fabricator locations around the world making the same products or exercising the same technologies -- the ability to find and share information saves time and time is still money. Alternatively, some suppliers are defining specific fabricators as worldwide sources for a given product or technology with all customers dependent on that single facility for key hardware. How then, can FA labs respond to this global challenge?

Again, several possible operational models exist, each with its own advantages and disadvantages. For a fairly large IDM, a *geographic* or *regional* model can optimize the turnaround-time (TAT) from fail to root-cause analysis. The premise is that the failing device is transmitted from the customer to the nearest laboratory to provide the shortest analysis time. This model typically requires multiple full capability labs — one in each general geography where the company operates or sells (e.g. North America, Europe, Pacific Rim) — with, if possible common tools and procedures. Close communication and a worldwide, commonly accessible database among FA labs and design and fabrication facilities are essential to build the trust and "buy-in" needed to accept FA results regardless of the location of analysis. This motel works well when TAT is the overriding issue. A second option is the *site of manufacture* model. In this model, the failing part is always returned to the location of manufacture for analysis. The premise is "the place that made it, owns it" and the fabrication site has the expertise required to link the specific product, process, and tooling, which may have caused the defect to the fail mechanism. There is a TAT impact due to ship time from point of fail to fabrication location (Customs, transit rime, etc.), but controversies caused by distance or differing toots sets are eliminated. The people responsible for the analysis and the corrective action are also now co-located. This model works well for worldwide single-sourced products as it requires only one set of process or product specific tooling or expertise.

Lab Design & Operations

Lab Design:

The design of a Failure Analysis lab revolves around tie expected workload. Space requirements are driven by tools (number and footprint) and personnel (numbers). "Wet" space used in chemical deprocessing is typically the most expensive space (ventilation, drains, and environmental concerns) as well as being the most heavily utilized space. Safety concerns which, while primarily governed by national or local government regulations and laws, are also highly affected by specific corporate policies. The population of the lab is driven by the organizational structure, the workload, cost constraints, TAT experience and requirements, and the tool set. The cost of operation can be divided into the costs of people and equipment to provide basic tool coverage and analytical support (*fixed* costs) and the costs driven by the quantity of work to he done and/or the TAT (*variable* cost). Fixed costs are essentially the depreciation, utilities, maintenance, materials, and labor costs associated with the ability to do one sample. If for example, the average TAT for a particular product or product type is 2.0 days, than one analyst can, on average, complete 2.5 parts/week or about 125 parts/year. Therefore, for every 125 parts of a given type expected, one analyst would be required – not counting administrative or support tasks.

Lab Operations:
Several models for day-by-day FA operations exist. They include:

- The Corporate Asset
- Cost (or Profit) Center
- Open Campus
- Dedicated Resource
- Tiered Support

Many of these models presume the lab exists in a corporate setting, usually an IDM, however several of them are applicable to independent laboratories and even university environments. The *Corporate or Division Asset* model implies that the FA lab is independent of any single/local manufacturing, development, or quality organization or product line. The costs are spread across all elements of the corporation or division as a "tax" as part of the cost of doing business and are not normally causally allocated. If the cost of the tooling is a major factor, and the product set is reasonably sized, this is a good model for medium-sized suppliers with multiple fabricators.

The *Cost or Profit Center* model implies again that the FA tab is an autonomous entity, but recovers its costs (or makes a profit) by causally charging its services to the requesting customers on an "as-used" basis. The lab is expected to be either self-funding (within a corporate environment) or to be a profit-making operation (corporate/division or independent lab) providing a return on investment over and above cost recovery. This model serves the purpose of accruing the cost of the FA operation to the customers in a causal fashion and allows the customers, in turn, to manage their costs by managing the submission of parts.

An interesting paradigm used in some companies is the "*Open Campus*" model. In this model, the FA organization consists of a small cadre of experts who provide the facilities, tools, technique development/documentation/training to a large number of people from a variety of organizations who use the laboratory and perform their own analyses. The FA personnel may perform a small number of very complex analyses, but have the main function of maintaining the lab, providing training and certification, and developing or procuring new techniques/tools as needed. The lab costs, including personnel, are allocated across the using organizations as a percentage of lab time used. The effect of this is to minimize the size of the FA organization per se. Concerns include the potential erosion of skills when an individual is an infrequent user, and the exposure for lack of careful tool use and resultant increase in maintenance costs.

The *Dedicated Resource* model is normally employed by large companies, although it also he usable by independent contact laboratories or research organizations/universities working on specific contract items. The premise of this model is that resource requirements are negotiated with and allocated to customers during the planning cycle, usually annually. Past history, planned volume (wafer starts. shipments, reliability monitoring strategy, qualifications, etc), process/product mix, and technology complexity, are normally used to project resource needs. Some flexibility is built in to allow for tactical changes, but the resource allocation is essentially like a service contact or insurance policy.

Tiered Support is not so much an operating model as it is a process to minimize TAT and manage workflow. The premise is that an analysis be performed at a facility as close in time and distance to the occurrence of failure. A hierarchy of laboratories is established, ranging from small, quick inspection facilities which may be at design centers or even sales offices,

through moderate-sized and equipped regional facilities, up to a full-function centralized total support facility. The paradigm is decentralized support and quick TAT so that relatively straightforward analyses (e.g. package mechanical damage) do not load down the main laboratory. This expected savings in time come at some capital and stalling cost.

Strategic Development

As has been mentioned in many articles and publications, the pace of technology continues to accelerate. Whether or not Moore's Law continues to hold true, new materials, product designs, processes, and customer quality requirements will drive the FA community to require more and more advanced techniques and tools, The role of strategic development in FA is to insure the preparedness of the FA laboratory to analyze these advanced products and/or technologies. This process must be pro-active as design schedules are shrinking and time-to-market becomes a key competitive advantage. Strategies for identifying gaps in needs versus current capabilities, the prioritization of development dollars and resources, and project planning, scheduling, and tracking are all key ingredients for strategic development teams. Another factor in the process is the increasing influence of external standards and expectations such as EIA Standard 671, CIQ-C, and ISO-900x/1400x. Continuous process improvement in terms of TAT, accuracy, and financial efficiency also drive the need for strategic development activities. Unfortunately, given the too—frequent situation where an PA organization is understaffed and overloaded, the drive to get today's work out often conflicts with the need to address tomorrow's needs. The also too-frequent actuality is that tomorrow's needs are not addressed until a crisis situation develops.

A strategy for staffing and managing strategic development projects for FA can be established using standard project management techniques which hove been published end taught for some time. The key techniques include Project Definition, Scheduling, Budgeting, Staffing, Supervising, Reporting, and Inspecting. In addition to many books and management courses, several computer programs are available to aid in the process. In medium to large size FA organizations with mature populations, the matrix organization structure can be an excellent management method. Each analyst is expected, as a part his/her job performance, to spend some period of time (e.g. 20%) working on strategic development projects. Each project is assigned an owner and staffing is pulled from across 'he organization. Obviously, this requires agreement and championship from the line management team and a realization that there may be some TAT impact on non-urgent analyses—customer involvement is also a key component.

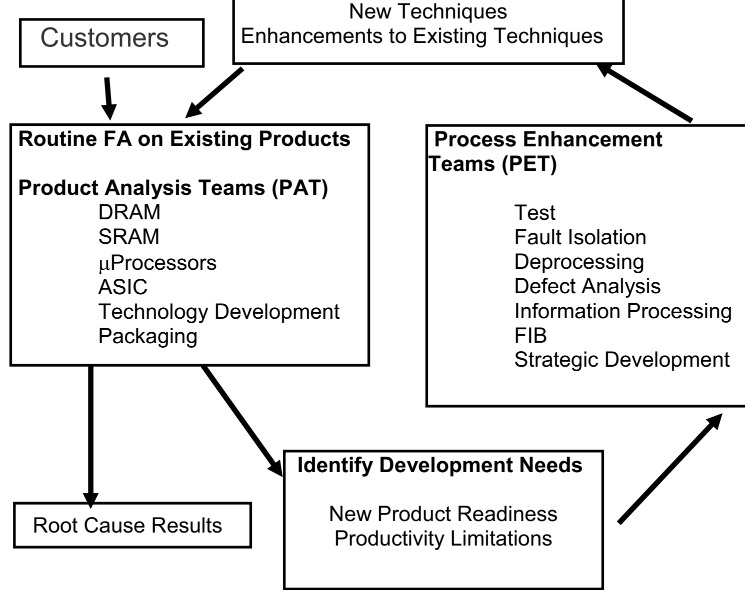


Figure 5: Continuous Improvement Flow

In today's environment, complex problems — some of which push the fundamental laws of Physics (e.g., root-cause optical inspection and/or mechanical probe placement of sub-0.25 micrometer features) -- often surface. The cost and time of strategic development and the knowledge base/expertise of an individual FA organization or even an individual company may be insufficient. In addition to the "go it alone' approach which is traditional in organizations which have bent typically highly competitive with one another, several alternative paradigms have emerged. The use, in larger corporations- of corporate resources such as Research & Development tabs, or, for organizations of any size, of independent tabs, universities, and, increasingly, partnerships with equipment vendors can reduce the time and money required to resolve strategic issues. Consortia of companies, which have traditionally been bitter competitors (e.g. SEMATECH and JESSI), are also an increasing means of providing funding and impetus for the development of tools and techniques to meet common. On or all of these strategies may be applied simultaneously to maximize the effective resource brought to tear

on difficult problems.

Financial Management

As the complexity of the problems continues to increase, the cost of analysis continues to increase. Personnel costs still account for almost 60% of total FA lab costs and are driven by the number and skill mix of human resources, the cost of labor, and the associated overhead costs. Tooling (depreciation) accounts for the next largest percentage — about 20% (and this percentage may well rise as the cost of new, complex tools continues to escalate). Once upon a time, FA tools were fairly inexpensive — the Scanning Electron Microscope (SEM) was probably the most expensive piece of equipment in a high-end FA laboratory and it cost around $100K. Today, new-generation FE-SEMs with associated EDX detection units cost upwards of $500K and some specialized tools (e.g. FIBs, SIMS, TEM, PICA, high-speed electrical characterization testers, etc.) can cost in excess of $1.0M. Clearly, to paraphrase Sen. Everett McKinley Dirksen's comments on the U.S. budget in the 1960's, a million here and a million there and pretty soon we're talking about real money. Financial management and cost-efficient operation are here to stay, whether, as technical folks, we like it or not. Since FA is usually looked at as a service operation (and a necessary evil) as opposed to directly revenue-producing activities, cost constraints are often more severe in FA operations than in product development or manufacturing – especially in "down" years for the industry or company. This is ironic, given that prompt, accurate FA can often result in improved yield (lower manufacturing costs), improved reliability (improved customer satisfaction) and indeed, often save The business for the company by providing quick and effective information for corrective action to resolve customer concerns. In any event prudence in financial matters is highly important and appropriate – and it just might save your job.

Given the high cost and high visibility of capital tooling the need for care in toot evaluation and selection becomes ever more important. Decision points for tool purchases will vary with the size and scope of the organization. Key factors such as whether or not an existing tool cast be upgraded to support the need, the cost of maintenance for older tools. The frequency at which the proposed solution is/will be needed, and personnel training costs will have significant influence on the decision process. If a technique/tool/process will be used infrequently, the best solution may be to, purchase it from an outside source. The drive to maximize the use of older tooling may seem sensible (tool is fully depreciated), but serviceability and increased maintenance may actually wind up costing more than the effect of depreciation of a new piece of equipment.

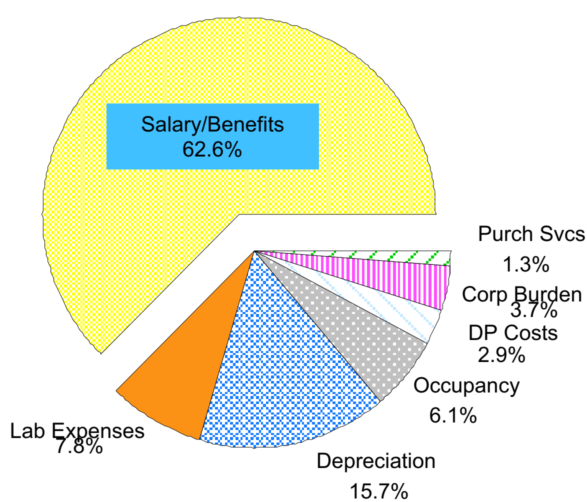


Figure 6: Typical FA Lab Cost Breakout

Once the decision is made that a new piece of equipment is justified, how shall we decide which tool to buy? In some cases where the market is essentially served by a single or dominant vendor, the choice is easy; however, this is not normally the case. Given variables like tool capability, cost (including parts. Supplies, and service), vendor expertise, user-friendliness, extendibility to new technologies, compatibility with existing equipment, delivery schedule, and service TAT, is there a cogent and methodic way to make an objective decision? There are, of course, several methodologies. One that provides a straightforward and logical approach is taught in Management Science courses as Multi-Criteria Decision Making (MCDM). The key' elements of MCDM are: identification of critical decision elements; assignment of weights to each element; pair-wise comparison of each alternative solution to each decision element; and the aggregation of this data to a decision point. The decision elements may be subdivided hierarchically as needed to establish the appropriate granularity of the elements. The process may be done by hand for the case of' three or fewer alternatives, but is best done using a spreadsheet or one of several commonly available software packages for larger numbers of alternatives and/or "what-if" analyses.

For techniques, the cost point comes down to

the equivalent of a “make or buy” decision. Frequency cost per use, service support, training, and proximity of service are factors in this type of decision. If a university, vendor, or independent laboratory offering the service in question is sufficiently nearby, the frequency of need is low, and the cost to purchase favorably compares to the cost to develop or own, the service should be purchased. If’ TAT would significantly suffer, or the workload imposed upon the outside provider would overwhelm their capacity, the service should be developed in-house. A traditional payback analysis for short-term needs (18 months or less) is adequate, but for longer-term needs, a financial analysis using an established internal rate of return (IRR) is more appropriate.

Financial concerns can and will affect the general management philosophy and operation of the FA organization. Multi-shift operation may become cost-effective to improve utilization of high-priced assets. A vigorous program of preventive maintenance can reduce costly tool downtime. Cross training of personnel and work schedule adjustments can optimize the efficient use of both human and capital resources. Reducing/eliminating duplication of effort among laboratories, control of expense items like chemicals, DP costs, travel, and supplies can also reduce the costs of operations. One approach toward the development of an understanding and methodology for operational cost control and efficiency, which has been widely heralded in recent years, is the concept of *Activity-Based Costing (ABC).* ABC involves the braking down of’ all aspects of the operation into basic units or “activities”. These activities can be hierarchically divided into Facility sustaining, Product sustaining, Batch, and Unit costs. Facility sustaing costs are those needed to keep the laboratory in an operating condition – maintenance, cleaning services, etc. Product costs would include tooling, techniques, interface boards, etc. nneded to analyze specific products. Batch costs include set-up time, inspection, etc. Unit costs include materials, energy, and labor. All the costs of the operation are then assigned, using historical data for a specified reference period, to one or more of these activities. Each activity and associated cost is then inspected in detail to assess its relevance to the mission of the organization and its efficiency. The entire organization is invited to “brainstorm”’ ideas for better ways to perform the activity at reduced cost or, even if the activity can be eliminated. As always, follow-up and implementation schedules are critical. Organizations employing ABC have reported overall cost efficiencies in the range of 15-30%, the reduction of “non-value-add time”, improved use of capital and human assets, improved TAT, and improved communications, a better understanding of the business processes across the organization, and more thoughtful decision making. Ideas can be as simple as the order quantity of supplies. For example, one organization saved 44% on chemical costs by purchasing high-use chemicals in larger unit sizes. Key ingredients for success of an ABC program include the commitment of senior management and the education of the personnel to obtain “buy-in”. The program is not intended to be a reduction-in-force program and agreements and commitments must be fulfilled — it cannot become the “program du jour”. Many books and articles exist on the process and concept of ABC and a growing number of independent consultants can be found to help tailor the process to an organization‘s specific culture and needs.

With all this, it cannot be left unsaid that FA needs to be “sold” to upper management and customers as a critical, even *essential*, enabler for the company. The cost of quality is a much-discussed issue in economics of corporations. Customers lost for technological or price reasons may be recouped with enhanced technical features or lower price points in the next generation of products. Customers lost for poor quality field performance rarely come back. FA provides the business the information needed to swiftly engage and resolve quality problems – you can’t fix a problem if you don’t know where it’s broken. Further, especially in the early phases of new technology or new product development and/or manufacturing ramp-up, yield is a crucial element of product cost. One point of yield in the first several months of manufacturing ramp-up may be worth more than the entire annual cost of the FA operation. Further, studies have shown that the major fraction of profit generated by a new or upgraded product occurs in the first 6-12 months after introduction. Time-to-market is increasingly critical and ever-shortening development/qualification schedules need to be met the first time – again, FA is a key enabler. Close working relationships with manufacturing, product development, and technical marketing can pave the way for a well-developed Return on Investment (ROI) case for FA personnel and equipment.

Metrics & Measurements

Since FA is so important and since we are all scientists and/or technical people, why should

we take the bother to internally measure anything? Just as the customer of an analysis can't resolve a problem until the root cause is established, the operational efficiency of the FA organization can't improve until we know where the root cause defects are. We must, in effect, failure analyze the FA process. External to the FA organization, measurements of customer satisfaction can help us to improve service quality. Internal to the FA organization, we want to optimize the use of resources, control costs, justify needed capital or personnel additions, and assess the readiness for new technologies. Standards, such as customer requirements and/or EIA Standard 671 on TAT expectations for analysis of customer returns, or JEDEC/EIA standards/publications on analysis report content and format may govern the operation's metrics. Corporate guidelines for expenses or other measures may be another source of standards. What is important is the decision of when to measure and when *not* to measure. A measurement which cannot or is not used is worse than no measurement at all, as it becomes a morale dissatisfier. What is it, then, that we really need to know? Key items for control and assessment of an FA operation include total TAT (queue time as well as process time), the rate (and trend) of unsuccessful or unresolved analyses, workload volume and source, and, most important, customer satisfaction. The measure of the components of TAT and unsuccessful analyses can serve to identity tooling pinch-points, obsolete tools, training deficiencies, overall staffing shortfalls, and customer communication issues.
Measurements are least effective when used as a club and most effective when used as a lever. Often new tooling, increased resources, and/or enhanced training opportunities can he justified to higher technical or financial management with a cogent set of metrics — especially if the technical metrics are combined with financial data which show' efficient use of existing equipment.

Job #:_____ Analyzing Dept.: ____ Eng./ Analyst: ___

Job Description: __________________________

1. Were the RESULTS received in a TIMELY and EFFECTIVE manner considering the assigned priority?

Yes ___ No ____

2. What is your assessment of the QUALITY of this work?

Excellent ___ Very Good ___ Average ___ Fair ___ Poor ___

3. What is your overall LEVEL OF SATISFACTION with our work on this particular job/project?

Excellent ___ Very Good ___ Average ___ Fair ___ Poor ___

If any of the above questions were answered Fair or Poor please explain below.

Figure 7: Customer Satisfaction Questionnaire

EXCEPTION: TAT CSI NDF/UNSUCCESFUL (Please circle as appropriate)

DATE **PROGRAM;**
JOB NUMBER : **TEAM LEADER:**

NO. OF PARTS AFFECTED: _______________
ANALYSIS CLOSED DATE: _______________

PROBLEM: __________________________________

ROOT CAUSE (Check one that best applies):
Tool/technique insufficient:
Tool/technique inappropriate:
Insufficient training:
Miscommunication:
Inadequate diagnostic information:
Inadequate customer data:
Inadequate verification/retest:
Inadequate outside lab support:
Unusually high complexity:
Experimental work:
Priority conflict:
Other (explain below):

Figure 8: FA Corrective Action Report

Summary

The management of a Failure Analysis laboratory requires a broad range of activities to optimize the efficiency of the operation. These have been described and some suggestions for approaches to the various activities have been outlined. The key item for consideration is that the pace of technology development continues to accelerate, and, with it, the need for timely, cost-efficient Failure Analysis. "When all else fails- ask yourself: 'How would I do this if I were running the business out of my garage?' If it

makes sense to the “garage test”, it’s probably a good idea.

References

Byham, William C., Zapp! The Lightning of Empowerment, Harmony Books, New York, 1991

Frank, R. and Lee, T., “Business Aspects of Failure Analysis”, Proceedings of the 22nd International Symposium for Test & Failure Analysis, ASM International, Materials Park, OH, 1996, pp 255ff

Hussein, Mohammed, Tracking and Controlling Costs – 25 Keys to Cost Management, New York Times Pocked MBA Series, Lebhar-Friedman Books, New York, 1999

Kerzner, Harold, Project Management – A Systems Approach to Planning, Scheduling, and Controlling, 2nd Edition, van Nostrand, New York, 1984

Krumwiede, Kip R., “ABC—Why It’s Tried and Why Its Needed”, Management Accounting V ol. 79, No. 10, April I 998 pp. 32-38

Kudva, S. et al, “The SEMATECH Failure Analysis Roadmap”, Proceedings of The 21st International Symposium for Test & Failure Analysis, ASM International, Materials Park, OH, 1995, pp 1ff

O’Guin Michael C., The Complete Guide to ActivityBased Costing, Prentice-Hall, Englewood Cliffs, NJ, 1991

Phillips, Jack J. and O’Connell, Adele O., Managing Employee Retention: A Strategic Accountability Approach, Elsevier, Burlington, MA, 2003

Ranfil, Robert M., R&D Productivity, 2nd Edition, Hughes Aircraft Company Study Report, Los Angeles, CA, 1978

Ross, Richard J., “The Development of a Total Quality Management System for a Semiconductor Failure Analysis Operation”, Master’s Thesis, National Technological University, 1991

Rothwell, William J. and Kazanas, H. C., Improving On-the-Job Training: How to Establish and Operate a Comprehensive OJT Program, 2nd Edition, Pfeiffer (an Imprint of John Wiley & Sons), San Francisco, CA, 2004

Saatv, Thomas L., Decision Making for leaders, 2nd Edition, RWS Publications, Pittsburgh, PA, 1990

Semiconductor Industry Association, “The National Technology Roadmap for Semiconductors”, SIA, San Jose, CA, 2003

Silverman, Melvin, The Technical Manager’s Survival Book. McGraw-Hill, New York, 1985

Materials Sources (Examples)

MCDM:

Expert Choice Software, Inc.
4922 Ellsworth Avenue
Pittsburgh, PA
www.expertchoice.com

Project Management:

Microsoft PROJECT 2003
Microsoft Corporation
Redmond, WA
www.microsoft.com

CA-SuperProject
Computer Associates
International Inc.
One Computer Associates
Plaza
Islandia, NJ
www.cai.com

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Managing the Unpredictable – A Business Model for Failure Analysis Service

C. Boit, K. Scholtens*, R. Weiland*, S. Görlich*, D. Schlenker**
*TUB Berlin University of Technology, Berlin, Germany / *Infineon Technologies AG, Munich, Germany*
***Robert Bosch, Reutlingen, Germany*

This article presents a complete business model for Failure Analysis (FA) as a high tech provider in the world of microelectronics. It starts with the definitions of a business process, and analysis flows as foundation are discussed. Then, a Key Performance Indicator (KPI) based operation is developed. The implementation of such a model in a FA lab is discussed with the interdependencies of workload and cycle times – the pipeline management. This opens the path to quantitative target setting agreements with the customers. A complete system builds a database that can calculate all the parameters for a FA lab tailored exactly to the demand of the customer. Such a database acts as a reference lab and represents best FA practice.

I. Introduction

Historically, Failure Analysis has been very much on its own, and not really integrated into the business processes of the semiconductor industry. Rather, FA is a parallel world prepared for the unpredictable, and often a stepchild in terms of roadmaps and investment. FA engineers are proud of their independence and knowledgeable work that only a few can deliver, but they are also torn between making the impossible happen and missing feedback or recognition.

This changed in many companies as FA became integrated into manufacturing process control, and product oriented FA had to face a paradigm shift to analysis techniques from the backside of the chip. All this improved the FA invest situation. However, the performance expectation of the customers remained paired with their unwillingness to define fixed rules. The nuts and bolds of FA lab operation are frequent topics for panel discussions, but little literature exists. While ref.[1] discusses a variety of operational models, there is no focus on quantitative aspects, and ref.[2] only looks at the cost of FA innovation. Nowhere do we find a detailed business model.

We present a complete business model for a high tech FA service. The model demystifies some of the business romanticism's of FA, and offers the chance to come to terms with the customer and help identify weak areas of their own performance. The core of this business model is a concise system of analysis flows. It helps to deliver a professional analysis that the customer can rely on. We discuss how the size of the operational unit, the variety of techniques, and the structure of the customers can influence key performance indicators.

Several examples show how the model can identify weak spots, and improve the performance of an FA lab to quantify targets that can be defined between the labs and their customers. The result is a lab that operates by transparent objectives, and with it come much higher confidence and satisfaction levels for the customers, and last but not least the analysts.

II. Elements of a Business Model

The Business Process

Every business model starts with a business process. This first step touches several common grounds of failure analysis, such as every job is different, failures are unpredictable, and the urgent problems may require such an immediate action that no model fits those jobs. So, the first rule of a business model is to keep it as simple as possible, and to see how much you can learn even on this basic level. In the end, you may find that about 80% of the jobs have a repetitive and a serial character. Failure analysts tend to be more sensitive to differences than to similarities, an attitude necessary for their analytical success, but an obstacle for introduction of a business model.

An FA business process that can be used as a foundation for a model is shown in ***Fig 1***. It starts with a job entry procedure followed by the analysis work. This is tightly connected with Step 3, the evaluation and conclusion portion. A conclusion may be that Steps 2 & 3 must be cycled several times with different technical processes until the failure mechanisms is identified and isolated. Usually the customer was already involved in Steps 2 & 3, but the final feedback of the result to the customer is an output that builds with the final report in Steps 4 & 5. This business process can identify weak and strong parts of the operation, and it may be important to qualify all the required information in the entry period until it makes sense to start the job.

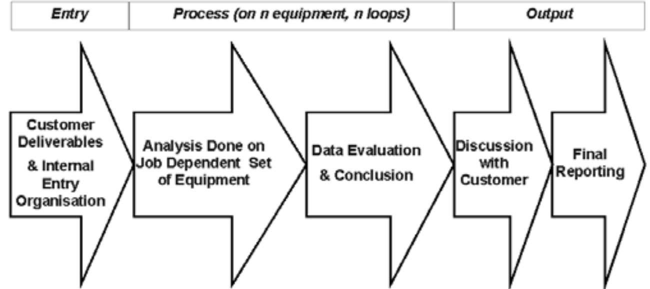


Fig 1: Business process for failure analysis

Process Flows as Analysis Guideline

When the business process is defined, then the next important structuring element is to categorize types of FA jobs and specify the process flows. Analysis process flows should be developed for all relevant fail types, and be specified for all respective technologies. An example for SRAM analysis based on a process flow is given in ref. [3] and ref. [4]. It does not include every possible failure mode, but presents a selection of frequently occurring fails. It is just a general demonstration how such a process flow leads almost automatically to the isolation of the failure. The physical analysis to nail down the exact root cause is not described there. It gives a sample, not the complete detailed case history.

The creation of an analysis process flow for fail categories is only justified if there is a gain in productivity - in other words a considerable job volume behind it. It moves the technical analysis to a professional level as a trademark of the lab. Even less experienced employees can reach best possible results with some additional supervision by the gurus. This adds a flavor of predictability to FA, and is the key for FA becoming a professional partner of microelectronics development and fabrication. When fully developed, it delivers a catalogue of process steps and modules that can estimate the complexity, expenses, and cycle time. These processes may vary from lab to lab and are not a topic of this article.

Key Performance Indicators (KPI) and Targets as Metrics Guideline

A third ingredient details the performance of the lab, and these are the Key Performance Indicators (KPI) and the targets that provide metric guidelines. KPIs on cost and logistics create a transparent correlation of workload and cycle time. If the lab operates on rules targeting job quantities per priority class, then speed and delivery reliability for the majority of the jobs improves. Sometimes customers and analysts may fear losing flexibility for the few jobs that demand personal processing. Again, customers and FA analysts usually overestimate the one extraordinary urgent job per month that makes them oppose such models. An agreement with key customers of the lab on targets and exceptions to the rules is an excellent way to gain experience and continuous improvement of the goals and targets of the lab. Application of such a business model creates transparency in operation. It has high impact on reduction of cycle time and increase of delivery reliability. This may astonish customers and analysts, but it more than justifies the additional discipline and steering force that is required.

FA reference lab

One last step remains in the model – the FA reference lab. An elaborate business process and an operating model allow concise calculation of a FA lab operation with the necessary personnel and set of equipment. This is based upon information about the technologies and product mix that the lab will serve, the job volume, as well as the integration of FA in all the development programs of the company.

III. Metrics and Target Setting: Aspects of an Operational Business Model

KPIs (Key Performance Indicators)

Each operational model needs a list of measurables that classify the performance of the lab, such as time to result, quality of the product, and cost of operation. We selected time to result to demonstrate the steps of KPI generation and implementation because time is critical in most analysis jobs, and the financial leverage of quick response exceeds by far the cost of FA operation. Criteria for the quality of the FA product differ from company to company, and in many cases are not easy to quantify.

The most important prerequisite of a KPI based operation is process orientation of the lab. Many FA lab organizations are segmented into technical units such as electrical probing, mechanical preparation and so on. This organization gives the best know how per technique, since the range of techniques one person has to deal with is limited. This may optimize throughput when most of the jobs follow a similar scheme. The disadvantage is that an optimization of the lab for job logistics such as cycle time and feedback to the customer is more complex.

The application of analysis flows also supports the philosophy that one job is generally performed by one analyst. The lab organization and technical responsibilities can be built up among the group as a project matrix. Another requirement is a database that allows access to job logistics following the business process. The first KPI step partitions the complete cycle time (CT) into distinguishable and accessible segments in the database (***Fig 2***).

The business process (BP) in ***Fig 1*** shows that the productive working time is typically the equipment working time plus the time for evaluation and administration. The equipment working time is identical to the analyst working time as long as the tools do not run unattended during operation. This is summed up as Analysis Time (AT). After subtraction of AT from CT, the remaining time is Hold Time. Hold Time is divided into Queue Hold Time (QHT) that belongs to the Entry phase in the BP, Job Execution Hold Time, and Hold

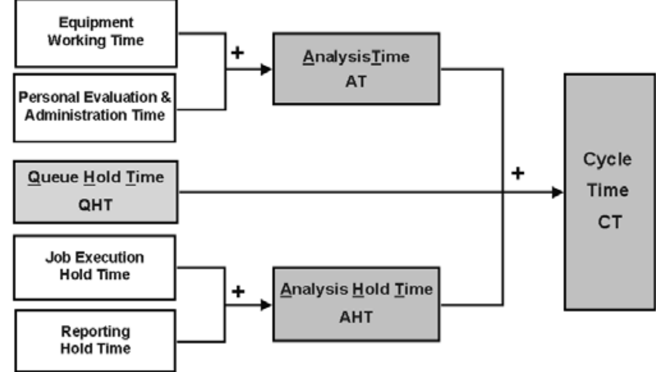


Fig 2: KPI tree logistics

Time to Report. The latter two combine to the Analysis Hold Time AHT.
With this segmentation, we can try a first assessment of possible hold time sources in the three basic phases of the Business Process (BP)

- Entry should ideally contain information about the verified failure, all required data from the customer such as layout, netlist, converted test program, etc. The assignment of priority class, resources and cost items is also part of the entry process segment. Hold times may increase if
  - too many jobs are in the lab (pipeline full)
  - priority changes on request of customer, or
  - job start requirements are not provided by the customer

- The Analysis Process starts with the question: "Failure verified?" and contains all the fail investigations on the system, the package, and the chip. This includes mechanical-chemical preparation and imaging. The process may repeat when the electrical signature of the defective needs to be re-characterized after layer-by-layer deprocessing (usually with feedback loops to the customer) and ends with the fault isolation. Hold times may occur through
  - Parallel execution of jobs (if more than one job per analyst at a time)
  - Utilization of the tools
  - Additional requirements from customers that arise during job execution
  - Delay due to outsourced parts of the process
  - Priority change on request of the customer
  - Lab maintenance
- The reporting phase begins with a quick feedback to the customer about the results, but it is not finished until the paperwork is done. This includes a written report, an invoice, and closing the job in the database. This phase is very sensitive to hold times. The most probable reasons are
  - Priority rules (report may have low priority after 1st feedback to customer)
  - Parallel execution of jobs (plus the above)
  - Lack of discipline

IV. Classification of Analysis Jobs

After segmenting the process flow, we can identify improvement targets tailored to each sector. From the business process viewpoint, all jobs look similar in their structure. The performance measure of an FA lab must respect the enormous range of complexity of the analysis jobs, from a quick cross section preparation to a rocket-science analysis of a soft timing problem in a 10-metal layer flip chip technology. Job categories solve the problem. The art is the selection of the job classes. There must be enough to group the jobs into classes that allow class correlated performance investigations, and they should be as few as possible to keep the database simple and the evaluation handy. A rule of thumb is not more than ten classes.

If flow-oriented job classes are hard to create, then it makes sense to start with a grouping by the level of Analysis Time (AT) per job. The more complex jobs usually are AT-intensive, and most of the less complex are quick. There is of course an error margin as the categories are not overly specific. Even a simple job with a cluster of 25 samples using only AT as a criterion would in the end also be counted in the high complexity class with an error of about 10-15%. But improvement potentials of 50% or more can be easily identified in the introduction phase of an operational business model.

An example of job classes is presented in ***Table 1***.

Table 1: Examples for Job Classes

Class	Example of Analysis	Explanation/Example
1	Qualification, simple	Qualification analyses after ESC-test fail; e.g., EOS, melted metal: I-V characteristics, local decapsulation, optical insprection
2	Qualification, medium	Qualification analyses after different kind of stress; e.g., electrical measurement, hotspot (liquid crystal or emission microscope), lift-off, SEM
3	Qualification, complex	A design/technology related problem or functional fail: tester integrated/SW-localization, (E-beam) probing, backside preparation, emission microscope, internal measurement by microprobing, delayering, cross sections, SEM
4	Qualification, very complex	A design/technology related problem of functional fail: tester integrated/SW-localization, (E-beam) probing, backside preparation, emission microscope, internal measurement by microprobing, delayering, cross sections, SEM with several refinement loops

This model uses four classes of analysis jobs from simple to very complex, which will be in focus from now on. In fact there are three more – Internal Physical Inspection, Circuit Edit, and Partial Decapsulation of packages that are not the subject of this article. Explanations are listed in ***Table 2*** for the analysis job classes.

Table 2:List of Analysis Jobs as Assigned to Job Classes

Complexity Classes for Analysis Jobs		**Number of working hours**
1 Phys/chem..	simple	≤ 4
2 Electr./Phys.	medium	4-16
3 Failure	complex	16-40
4 Analysis	very complex	≥ 40

The highest level of job classification is the analysis process modules that are identified as operational units. Every FA job can then be described as a specific composition of the modules. Such a system has many advantages

- Each module has its own performance program
- Easy job classification has by number of process modules or application of previously defined key modules per category (i.e., analysis from chip backside = a job class automatically very complex)
- Good overview of expected job complexity based on a fail description already at the job start helps to define the scope of the work from the beginning
- Estimation of job price and time of delivery with the customer at the job start with better approximation of complexity level of the investigated fail and better basis to decide on priority level

A model for the definition of analysis process modules and a description of a typical analysis job as a serial application of some of these modules is shown in ***Table 3***.

Table 3: Modules of Analysis Jobs

Analysis Modules	**Metal Layers of Technology**		
	2	4	8
Electrical Verification / DC or Board	x	(x)	
Electrical Verification / Tester	(x)	x	x
ATPG Analysis		(x)	x
DC-Localization 2/Photo Emission or LC	x	x	x
Backside Analysis Preparation		(x)	x
Backside Localization		(x)	x
Delayering 1 Metal Layer – Parallel Polishing		xx	xxx
Delayering 1 Metal Layer-Wet/dry Etch	x		
Pad depo by FIB		xx	xxx
Voltage Contrast by FIB	x	xx	xxx
Voltage Contrast by SEM	(x)		
Probing of Failing Microcircuitry	x	xxx	xxxx
Cross Section Preparation – Mechanical	(x)	(x)	(x)
Cross Section Preparation – FIB	x	x	x
Cross Section Imaging – SEM	x	x	x
Cross Section Imaging – TEM		(x)	(x)
Doping Profile CSM	(x)	(x)	(x)

x – module does not apply
(x) – module may apply
multiple x – number of times to apply module

V. Pipeline Management

Analysis today is much less predictable than wafer manufacturing, and we find that hold times are the major share of FA cycle time. To manage and reduce the hold times, we used the following manufacturing model to describe the FA lab logistics.

Pipeline management means handling of the dynamics between CT (Cycle Time) and throughput. This tradeoff is optimized when the hold times are as low as possible with respect to the job priority. So, the most important instrument is a unit that expresses the factor by which the hold times exceed the Actual Process Time (AT). This Factor is the Flow Factor (FF)

Eq 1

FF = CT / AT

FF also allows a clear overview of the hold times independent of job classes (which in turn are valuable to identify improvement potentials). The numbers from one sample lab of our department in Munich (out of five) were chosen. *Fig 3* shows how a time scale transforms into a FF scale.

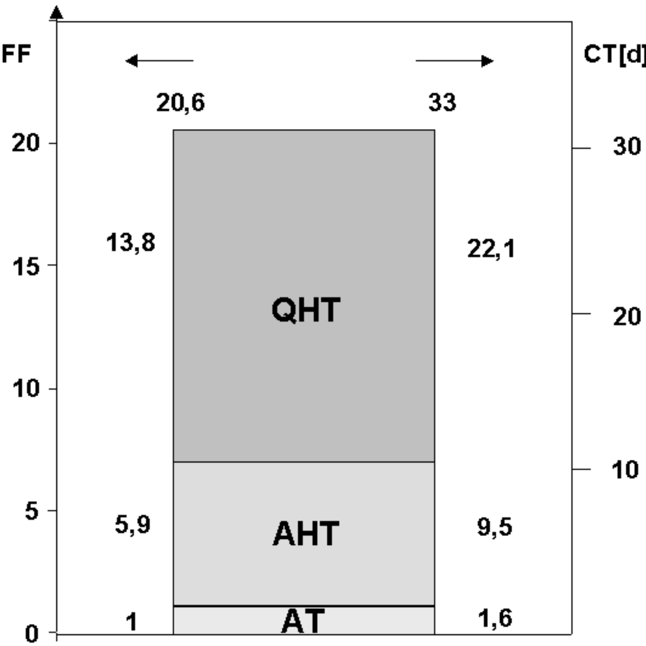


Fig 3: Example of an average FA cycle time of a lab expressed in time and FF with shares of AT, AHT and QHT

FF = 1 means CT = AT following the definition of ***Eq 1*** as long as parallel work of several persons for one job is not a significant factor. The calculations become more complex for the parallel work per job and are not included here. In our sample lab we see that CT exceeds the physical cycle time by an order of magnitude. This is common in FA labs, and we have seen similar numbers in several companies.

The implementation of a KPI system, the tracking of the parameters, and the mere discussions about improvement potentials usually yield a considerable decrease of the cycle time as demonstrated in ***Fig 4***, 1st step. The improvement potentials can be identified with a survey about the origins of hold times in the lab operation. This information should be arranged in a Pareto chart, and for each of the topics can be identified as an improvement potential (***Fig 5***). Typically, there are two major areas of improvement: internal lab logistics and the interface of the lab with the customer.

Further targets must be set to fully develop the full CT reduction potential, and this must include customer commitments to keep the job input to the agreed workload including priority classes. This helps understand lab logistics and their dynamics in the application of a factory model. There may be some resistance to accept a factory model as a base for FA operation. Analysis engineers as well as customers tend to see themselves and the problems they deal with as unique that do not compare with each other, and no factory model ever seems to meet their needs. But, after implementation of such a model, cycle times become much more predictable, delivery more reliable, and the throughput of top priority jobs increases. As a result, the operation is more transparent, the customer is more satisfied, and the FA engineers can plan their work.

Factory as Operational Model & Operating Curve

Now we describe a calculation model that shows the lab logistics in an operational manner. It will be done in four steps. First, we take the list of job complexity classes (C1, C2, etc.), and specify the average physical cycle time AT per job class ***Table 4***. Second, we assign specific average flow factors FF to each priority level (P1, P2, etc.). The result is an average CT per class of the category complexity and priority ***Table 5***.

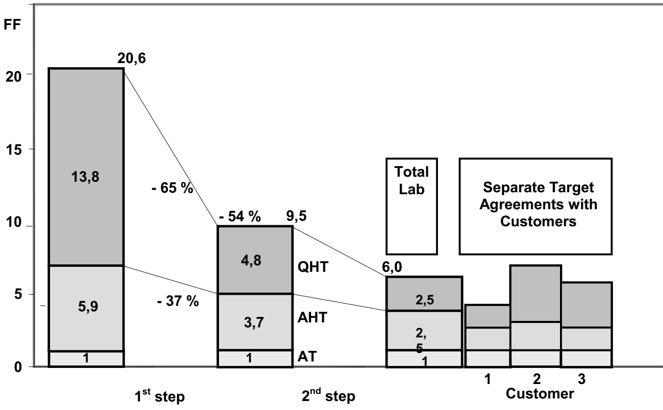


Fig 4: CT Reduction steps: 1st step, implementation of KPI Model; 2nd step, quantitative agreements with customer

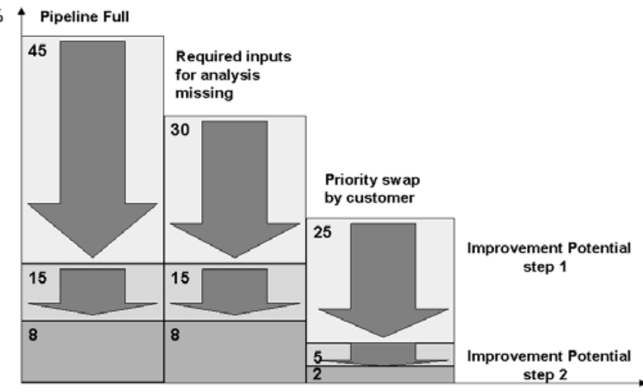


Fig 5: Hold Time Pareto for QHT with improvement potential assigned

Table 4: AT (Analysis Time) of Jobs per Complexity Class (C1, C2)

Complexity Class		
C1	C2	Etc.
AT_1	AT_2	

Table 5: Cycle Times (CT) and Flow Factors (FF) of Jobs per Complexity (C) and Priority (P) Class

	C1	C2	Etc.	FF
P1	CT_1	CT_2		FF_1
P2				
Etc.				

The third step makes an assumption about the maximum workload in the lab in terms of the job volume distributed over the mix of complexity and priority in the lab. The result is a table that contains Work in Progress (WIP) per priority and complexity class ***Table 6***. This looks simple, but contains all the knowledge about the composition of the lab performance. How many Priority-1 jobs can be operated? And depending on the distribution between the priority and complexity classes, what FF can be assigned to them? Experience shows that if the Priority-1 share exceeds 15%, then the FF must increase to over 1.5. All hold time contributors must be taken in account when the CT and WIP tables are built.

Now we add another equation, known as ‘Little’s Law’, to correlate WIP and throughput (called the ‘Going Rate’ (GR))

Eq 2

$$GR = WIP / CT$$

The going rate is usually counted daily (DGR), weekly (WGR), or monthly (MGR). In our case, typical throughputs of FA labs relate best to MGR that shows a good statistical

Table 6: Work In Progress (WIP) in Lab (jobs in pipeline)

	C1	C2	Etc.	All
P1	WIP			
P2				
All				WIP total

Unit: analysis jobs per complexity (C1, C2) and priority class (P1, P2) under assumption of job mix as defined in Table 5

Table 7: Monthly Going Rate (MGR) of lab (throughput per month)

	C1	C2	Etc.	All
P1	MGR			
P2				
All			MGR total	

Unit: analysis jobs per complexity and priority class, under assumption of job mix as defined in Table 5. Note: WIP higher than lab capacity increases CT. In a constant condition of overload, MGR is independent of WIP, but the finished jobs have a long time. MGR or Throughput for Above Assumed Job Mix.

correlation. ***Table 7*** shows the respective MGR table. We recommend working with the model to get used to the dynamics. For example, hold times (or FFs) do not affect throughput because they increase both WIP and CT. ***Eq 2*** shows that GR is hold time invariant. GR is basically a direct function of AT. This calculation can encourage customers to cooperate on hold time reduction, because the rumor is widely spread that shorter CT means loss in throughput, and the model clearly defeats this opinion.

With this model, it is easy to predict CT and throughput for given job mixes in terms of priority and complexity classes. It tells how the WIP increases with hold times. The FA manager can detail how a job input that markedly exceeds the agreed WIP will pile up the QHT, and how the timely supply of information such as netlists, test program conversion, etc., will improve the performance. The FA manager can discuss how many parallel jobs per analyst make sense (WIP). It can easily be shown that a limited number of Priority-1 jobs with ultra-short FF not only delivers excellent CTs, but also considerably increases the MGR of Priority-1 jobs. So with a little more discipline in job steering, the output shifts to more Priority-1 jobs that are finished much quicker than before.

Now it is possible to identify logistics improvement potentials, define measures to lift the potentials, and set quantitative targets based on the action plan. This is a complete process for KPI and target setting of FA lab logistics. This transparent model should yield a more exact target agreement with the customer, implement a job steering process, allow better control with a KPI report, and introduce high tech service providers to target tracking and balanced score cards.

The CT performance improves immediately after implementation of an exact agreed upon job input with the customer according to the WIP. This is especially important for high priority jobs with a maximum limit of about 15% share of the job mix. A performance check after 2 months, with no additional measure taken at that time shows (***Fig 6***) the improvement for a limit of 15% Priority-1 share of the job mix.

The heart of a manufacturing site model is the operating curve. This curve shows the average FF of the fab as a function of utilization U (utilization = actual workload / maximum workload). The basic formula for an operational curve is

Eq 3

$$FF = \alpha\,(U/(1-U)) + 1 \quad \text{with} \quad U = U_{act} / U_{max}$$

where α is a parameter indicating how rapidly hold times increase when the workload increases. In ideal operation, FF is constant at 1 until workload reaches its maximum value, where it has a step function into infinity. This happens when α is close to 0. With higher values of α, the curve flattens out and FF is already increasing at lower utilization rates. A wafer fab operates at an α well below 1. When we fit our sample lab data with ***Eq 3*** (***Fig 7***), we can see how α improves from about 10 before step-1 of the KPI program to about 5 after the first step. So, an operating curve also works as a performance indicator for a FA lab.

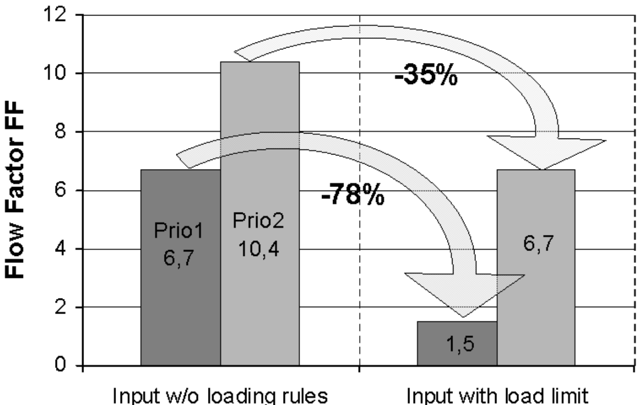


Fig 6: Improvement of CT

Since wafer fabs operate at better parameters, it is not surprising that the approximation of FA operation with a factory model is very rough. An optimized FA lab can reach α values around 2, and that is a real surprise. A FA lab has no reason to hide its operational parameters that account for the low degree of repetition or automation of the process and the flexibility requirements.

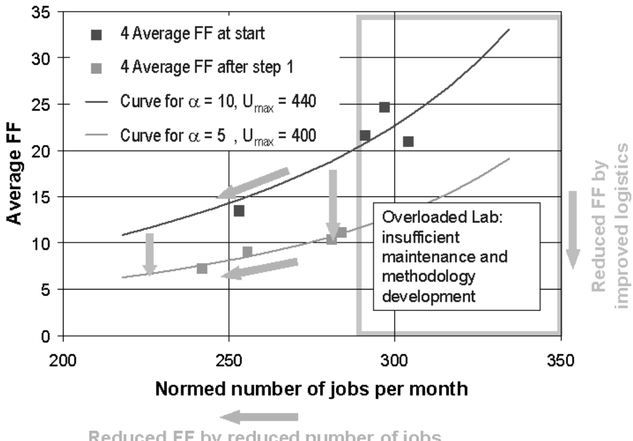


Fig 7: Operating Curve for FA lab

But the other fit parameter, U_{max}, requires some discussion. In the model, FF asymptotically approaches large values at U_{max}. No lab or fab wants to operate close to such a FF. Manufacturing utilization is typically kept to 75-80% of U_{max} to achieve reasonable FF. We see, the FA lab again can compete.

The degree that a FA lab can reach such high utilization values depends on the flexibility of additional duties such as lab maintenance, readiness-for-the-future development projects, and training. There are also many more duties that require workforce assignments of up to 30-50% of the total working hours available in the lab, and this can be organized around the job execution. This is the answer to the still existing mysteries of job input in the FA business, and allows the lab to breathe with the customer's demand, but only for a limited period of time. So, a FA lab can operate at workload peaks closer to U_{max} (the marked area in ***Fig 7***) as an additional performance offer to the customer.

VI Integration of FA into Technology Development Roadmap of the Company: Innovation Aspects of the Business Model

In addition to technical flows and operational metrics, a professional business model can be developed that secures the readiness for the challenges of the future. This business model has two parts. One is FA-inherent and compiles a technical roadmap of the FA capabilities and requirements. These correlate to the SIA-roadmap plus development of analytical equipment. The other one is the explicit participation of FA in the default business process of all innovation and development projects in the company with their own objectives. This adjusts the kick-offs and milestones of FA development projects to the general innovation strategy of the company.

Capabilities and Requirements for Technical Roadmaps for FA

The technical roadmap of FA techniques can be obtained by watching the SIA roadmap, the FA roadmap as compiled by the SEMATEC Product Analysis Forum[2] and by gathering more detailed information at equipment supplier fairs, conferences, and by developing in-house inventions. But all those technical capabilities can only secure readiness for the future when technologies or products make best use of the advanced analysis techniques. The prerequisite for this is *Design for FA*, something that is kept very secret by the companies. More public discussion would at least help define a minimum standard. That would also help the equipment vendors to standardize their products.

FA in Innovation Business Process of the Company

The roadmap of FA capabilities and requirements of the future indicate that FA is ready for integration into the business process of all the development projects in the company. For every innovation in the front end or back end technology, or product design or test that gets implemented in a development project, the FA roadmap contributes its part. A working integration relies on an introduction of the required FA project milestones in the milestone plan of the major development projects of the company. Some progress on this topic is seen in most of the companies in recent years, although a full integration is yet ahead.

FA Reference Lab - The Ultimate Database for Design of FA to Demand

A reference lab is a virtual lab. It is a database that calculates all the necessary parameters to build a lab from scratch for any service specification fed into the database. With such a tool, a FA lab can be designed to the anticipated specific mix of technologies and products for the demands of the customer. To calculate the parameters, the model must be based on measurables introduced in the previous chapters. The more real labs have been evaluated to create the data base, the closer the virtual lab can define best practice.

This built-in best practice experience makes the database a reference lab. The business model with all its classifications and standardizations is a mandatory integral part.

Many further differentiations are necessary to identify the required equipment and personnel resources of the lab such as

- Front End (wafer level) Manufacturing
- Back End (package level) Manufacturing
- Product Definition/Quality.
- FA for development projects
- Specific technologies (such as deep trenches, security cards, etc.)
- Specific packages and many more.

Analysis process modules must express all of these specifics in module quantities. By multiplying the expected number of analysis jobs, we obtain the workload, and after several detailled calculations we can define the final equipment and personnel requirements of the lab. This is design to demand.

VII Conclusion

The introduction of a business model for failure analysis service labs was discussed in seven steps

- Definition of the foundation of the model – a business process.
- Introduction of a process flow to assure a high quality level to the product of a FA lab, the analysis result, or root cause identification. With the introduction of process flows comes a standardization that we can use for the next step, the implementation of
- Performance metrics and target setting with key performance indicators, presented here for cycle time logistics. To get more resolution on critical areas in the lab, it is important to
- Classify the analysis jobs by task and complexity. With these specifications we can
- Apply a factory logistics model with an operational curve to measure the performance of the lab and start target setting with the customer.
- A firm FA roadmap with lab readiness for the future is assured.
- The final phase of the business model is a FA reference lab, a database as virtual lab. This is fed all information developed in the previous steps, plus a hierarchical decision flow on success versus effort. This can be used for designing FA labs and for best practice learning.

This business model can

- Improve the performance of the lab
- Achieve clear target agreements with the customer on throughput, cycle time, maximum number of jobs in the lab per priority, and complexity class
- Build a solid ground of trademark result quality
- Assure readiness of the future
- Integrate FA in the full business process of the company from idea to product
- Reduce the burden of the FA personnel because of a better predictable workload and more realistic claims of the customers

For the first time a transparent business model was successfully introduced to a high tech service provider such as FA. It can as well be transferred to similar groups.

Acknowledgement

We want to thank the Infineon FA crew in Munich for making this project happen. This occurred from the phase when it seemed to be only extra overhead work that kept us from doing the analysis job, through all the contradictory and pointless results that may only have told trivial results on a higher level, to the final success with an improved performance and customer reputation. Thanks for your patience, everybody. It paid off with a high level of FA integration at Infineon and appreciation from the executive board.

References

1. R. Ross, **"Failure Analysis Laboratory Management,"** *Microelectronic Failure Analysis Desk Reference* , Editors R. Ross, C. Boit, and D. Staab, 4th edition, *ASM International*, Materials Park OH 1999, p 569-574.
2. C. Boit, et al., **"Can Failure Analysis Keep Pace with IC Technology Development,"** *Proc.7th IEEE IPFA* 1999, Singapore, p 9-14.
3. C. Brillert, et al., **"SRAM Failure Analysis Flow,"** *Proceedings of 27th ISTFA 2001*, ASM International, Materials Park OH, p 323-329.
4. P. Egger, C. Burmer, "**SRAM Failure Analysis Strategy"**, Proc. 29th *ISTFA, 2003*, p. 177

Copyright © 2011 ASM International®.
All rights reserved.
Product code: 09110Z

Failure Analysis Terms and Definitions

Ryan Ong
National Semiconductor Corporation, Santa Clara, California

Acknowledgements:

This terms and definitions section, first started by Chris Henderson, is a compilation of terms that have been addressed in the previous proceedings of the International Symposium for Testing and Failure Analysis (ISTFA).

ASTM definitions have been reprinted, with permission, from the ASTM Standards, copyright American Society for Testing and Materials, 1916 Race Street, Philadelphia, PA, 19103-1187

ASME definitions have been reprinted with permission from the American society of Mechanical Engineers, 345 East 47th Street, New York, NY 10017-2392

EIA definitions have been reprinted with permission from the Electronic Industries Association, 2500 Wilson Boulevard, Arlington, VA, 222001-2392

IEEE definitions have been reprinted from IEEE Std 100-1992 IEEE Standard Dictionary of Electrical and Electronics Terms, copyright 1993 by the Institute of Electrical and Electronics Engineers, Inc. The IEEE disclaims any responsibility or liability resulting from the placement and use in this publication. Information is reprinted with the permission of IEEE.

National Technology Roadmap for Semiconductors definite ions are reprinted with permission from Semiconductor Industry Association (SIA), 4300 Stevens Creek Boulevard, Suite 271, San Jose, CA, 95129

SEMI definitions are reprinted by SEMATECH with permission from Semicondcutor Equipment and Materials International, 805 E. Middlefield Road, Mountain View, CA, 94043

SEMATECH definitions are reprinted from the SEMATECH Official Dictionary Version 5.0, copyright, 1995 by SEMATECH, 2706 Monotopolis Drive, Austin TX, 78741

ISTFA definitions are taken from the symposiums from 2000-2003.

Acceptor n: in a semiconductor, an impurity in a semiconductor that accepts electrons excited from the valence band, leading to hole conduction. [SEMI M1-94 and ASTM F1241] Also see hole

Access time n : a time interval that is characteristic of a storage device and is essentially a measure of time required to communicate with that device. [IEEE]

Acoustic Micro Imaging (AMI) : A method of evaluating materials and bonding for various microelectronic application by using high frequency ultrasound from 5-300 MHz to image the internal features of the samples. [ISTFA]

Accumulation condition n: the region of the capacitance-voltage (C-V) curve for which a 5-V increment toward a more negative voltage for p-type material, or toward a more positive voltage for n-type material, results in a change of less than 1% in the maximum capacitance, Cmax. [ASTM F1241]

Active area n : the region of thin oxide on a die or wafer in which transistors and other circuits reside. [SEMATECH]

Active devices n: semiconductor devices that have active function, such as integrated circuits and transistors. [SEMI G35-87] Contrast passive devices.

Adhesion, resist edge n : the ability of the edge of an image in a development resist coating to adhere to its substrate under applied physical or chemical stress. [ASTM F127-84]

Adhesive stringer n : on a photolithographic pellicle, any detectable protrusion from the edge of the adhesive. [SEMI p5-94]

Aeolotropic : see anisotropic

AES : see Auger electron spectroscopy

AFM : see Atomic Force Microscopy

Alignment n 1: the accuracy of the relative position of an image on a reticle with reference to an existing image on a substrate [SEMATECH] 2 : a procedure in which a wafer is correctly positioned relative to a reticle. [SEMATECH] 3 : the mechanical positioning of reference points on a wafer or flat panel display substrate (also called alignment marks or

alignment targets) to the corresponding points on the reticle or reticles. The measure of alignment is the overlay at the positions on the wafer or substrate where the alignment marks are placed. [Adapted from SEMI P18-92 and D8-94] Also see direct alignment and indirect alignment.

Alloy n 1: a composite of two or more elements, of which at least one is metal. [SEMATECH] 2 : a thermal cycle in which two or more discrete layers (of which at least one is metal) react to allow good electrical contacts. [SEMATECH}

Aluminized area n : in a cerdip or cerpack semiconductor package, the lead frame area coated with aluminum to provide a surface suitable for wire bonding. The maximum area is defined by the inside dimension of the cap or ceramic ring. In some cases, the die attach area is also coated if a full lead frame is used. The coating may be vacuum deposited or bonded. [SEMATECH]

Aluminized width n : in a semiconductor package, the width of the area coated with a protective layer of aluminum. This area covers most of the top formed width. [SEMATECH] Also see package, bond finger, top formed width, and aluminized area.

Aluminum (Al) n : a metal used to interconnect the devices on a wafer and to interconnect external devices or components. [SEMATECH]

Ambient temperature (TA) 1: the temperature of the surrounding medium, such as air or liquid, that comes into contact with the device or apparatus. [SEMATECH} 2: the temperature of the specified , surrounding medium (such as air, nitrogen, or a liquid) that comes into contact with a semiconductor device being tested for thermal resistance. [SEMI G38-87]

AMI : See Acoustic Micro Imaging

Ammonium fluoride (NH_4F) : a white crystalline salt used to buffer hydrofluoric acid etches that dissolve silicon dioxide by not silicon. An example of such an etch is the buffered oxide etch. [SEMATECH] Also see pinhole.

Ammonium hydroxide (NH_4OH) : a weak base formed when ammonia is dissolved in water. [SEMATECH]

Amorphous silicon : silicon with no discernible crystalline structure. [SEMATECH] Contrast polycrystalline silicon.

Analog adj : A signal in an electronic circuit that takes on a continuous range of values rather than only a few discrete values; a circuit or system that processes analog signals. [1994 National Technology Roadmap for Semiconductors] Contrast discrete

Angle-resolved scattering (ARS) n : technique that measures light scattered from particles as a function of angle; used to characterize particles. [SEMATECH]

Angstrom (Å) n : unit of linear measure equal to one ten billionths of a meter (10^{-10} m). The diameter of a human hair is approximately 750,000 Å.) The preferred SI unit is nanometers. 10 Å = 1 nm. [SEMATECH]

Anion n: an ion that is negatively charged. [SEMATECH]

Anisotropic adj. : exhibiting different physical properties in different directions. NOTE – In semiconductor technology, the different directions are defined by the crystallographic planes. [SEMI M1-94 and ASTM F1241] Also called non-isotropic and aeolotropic. Also see anisotropic etch

Anisotropic etch n: a selective etch that exhibits an accelerated etch rate along specific crystallographic planes. NOTE – Anisotropic etches are used to determine crystal orientation, to expose crystal defects, and to facilitate dielectric component isolation. [SEMI M1-94 and ASTM F1241] Also called preferential etch. Also see anisotropic.

Anneal n : a high-temperature operation that relieves stress in silicon, activates ion-implanted dopants, reduces structure defects and stress, and reduces interface charge at the silicon-silicon dioxide interface [SEMATECH]

Anomaly: see defect

Antireflective coating (ARC) n : a layer of dielectric material deposited on a wafer before resist to minimize reflections during resist exposure. [SEMATECH]

Antimony (Sb) n: a brittle, tin-white, metallic chemical element of crystalline structure. Antimony is used as an n-type dopant in silicon, often for the buried layer. [SEMATECH]

Avalanche photo-diode (APD): Solid-state detector with high quantum efficiency in the near infrared region. [ISTFA]

APD : See Avalanche photo-diode.

ARC : see antireflective coating.

Architecture n : of a computer system, a defined structure based on a set of design principles . The definition of the structure includes its components, their functions, and their relationships and interactions. [SEMATECH]

Area contamination n : foreign matter on localized portions of a wafer or substrate surface. [SEMI M3-88]

Arsenic (As) n: a highly poisonous chemical element, which is brittle and steel gray in color. Arsenic is often used as a n-type dopant for buried layer predisposition. [SEMATECH] Also see n-type.

Artifact n 1: a physical standard against which a parameter is measured; for example, a test wafer used for testing parametric drift in a machine. [SEMATECH] Also called

standard reference material. 2 : a superficial or unessential attribute of a process or characteristic under examination; for example, a piece of lint on a lens that appears through a microscope to be a defect on a die. [SEMATECH] 3 : in surface characterization, any contribution to an image from other than true surface morphology. Examples include contamination, vibration, electronic noise, and instrument imperfections. [SEMATECH]

Ash v : to apply heat to a material until the material has been reduced to a mineral residue [SEMATECH]

Asher n : a machine used to remove resist from substrates. [SEMATECH]

Ashing n : the operation of removing resist from a substrate by oxidation; a reaction of resist with oxygen to remove the resist from the substrate. [SEMATECH]

Aspect ratio n 1: in etch, the depth-to-width ratio of an opening of a wafer. 2 : in feature profile, the ratio of height to width of a feature. [SEMATECH]

Atomic force microscopy (AFM) n: a microscopy technique based on profilometry using an atomically sharp probe that provides three-dimensional highly magnified images. During AFM, the probe scans across a sample surface. The changes in force between the sample and the probe tip cause a deflection of the probe tip that is monitored and used to form the magnified image. [SEMATECH]

Atomic percent n: in electron spectroscopy for chemical analysis (ESCA) of plastic surface composition, the number of atoms of a particular element present in every hundred atoms within the ESCA detection volume. [SEMATECH]

ATPG : see automatic test pattern generation.

At-speed test n: any test performed on an integrated circuit that tests the device at its normal operating clock frequency. [1994 National Technology Roadmap for Semiconductors]

Auger electron spectroscopy (AES) n : the energy analysis of auger electrons produced when an excited atom relaxes by a radiationless process after ionization by a high-energy electron, ion, or X-ray beam [SEMATECH] Also called Atomic-emission spectroscopy, SAM, Scanning Auger microprobe

Auger process n: the radiationless relaxation of an atom involving a vacancy in an inner electron shell. An electron is emitted, which is referred as an Auger electron [ASTM E673-90]

Autodoping n : in the manufacture of silicon epitaxial wafers, the incorporation of dopant originating from the substrate into the expitaxial layer [SEMI M1-94 and ASTM F1241] Also called self-doping. Also see doping and substrate.

Automatic test pattern generation (ATPG) n : the automatic development of vectors which, when applied to an integrated circuit, permit faults to be detected in the performance of the integrated circuit. [1994 National Technology Roadmap for Semiconductors]

Back-end of line (BEOL) n : process steps from contact through completion of the wafer prior to electrical test. Also called back end. [SEMATECH]

Backgrind n : an operation using an abrasive on the back side of a substrate to achieve the necessary thinness for scribing, cutting, and packaging of die. [SEMATEC]

Back oxide n : a layer of silicon dioxide formed on the back of a wafer during oxidation. [SEMATECH]

Back surface n : of a semiconductor wafer, the exposed surface opposite to that on which active semiconductor devices have been or will be fabricated. [ASTM F1241] Also called backside.

Bake n: in wafer manufacturing, a process step in which a wafer is heated in order to harden resist, remove moisture, or cure a film deposited on the wafer. [SEMATECH]

Ball-grid array (BGA) n : an integrated circuit surface mount package with an area array of solder balls that are attached to the bottom side of a substrate with routing layers. The die is attached to the substrate using die and wire bonding or flip-chip interconnection. [SEMATECH] Also called land-grid array, pad-grid array, or pad-array carrier

Bar : see die, crossbar, and bar end.

Bare die n : individual, unpackaged silicon integrated circuits. [1994 National technology Roadmap for Semiconductors]

Barrier n : a physical layer designed to prevent intermixing of the layers above and below the barrier layer; for example, titanium tungsten and titanium-nitride layers. [SEMATECH]

Barrier layer : see depletion layer.

Base n 1 : in semiconductor manufacturing chemicals, a substance that dissociates in water to liberate hydroxyl ions, accepts a proton, has an unshared pair of electron, or reacts with acid to forma salt. A base ha a pH greater than 7 and turns litmus paper blue. [SEMATECH] 2 : in facilities and safety, a corrosive material with the chemical reaction characteristic of an electron donor. [SEMI S4-92] 3 : in quartz and high temperature carriers, the material at the bottom of a wafer carrier on which the wafer carrier rests when placed on a flat surface [SEMI E2-93] 4 : of a frame, a window frame, and the cap are attached to the base – generally with devitrifying solder glass – during package/device manufacture. [SEMI G1-85] Also see cap and window frame.

Behavioral n :a level of logic designed that involves describing a system at a level of abstraction that does not involve detailed circuit elements, but instead expresses the circuit functionality linguistically or as equations. [1994 National Technology Roadmap for Semiconductors]

BEOL : see back-end of line

BFM : see bit map fail.

BGA : see ball grid array.

BiCMOS design n : the combination of bipolar and complementary metal oxide semiconductor design and processing principles on a single wafer or substrate. [SEMATECH]

Bimetal mask : see mask, bi-metal.

Binding energy n : the value obtained by subtracting the instrumentally measured kinetic energy of an electron from the energy of the incident photon, corrected for an instrument work function. [SEMATECH]

Bipolar adj : a semiconductor device fabrication that produces transistors which use both holes and electrons as charge carriers. [SEMI M1-94 and ASTM F1241]

Bird's beak n : a structural feature produced as a result of the lifting of the edges of the nitride layer during subsequent oxidation. [SEMATECH]

BIST : see built-in self test.

Bit Fail Map : A matrix of 0s and 1s describing passing and failing SRAM cells at test. [ISTFA]

Blister ceramic n : an enclosed, localized separation within or between the layers of a ceramic package that does not expose an underlying layer of ceramic or metallization. [SEMI G61-94] Also called bubble ceramic.

Blister metal n : in packaging, an enclosed, localized separation of a metallization layer from its base material (such as ceramic or another metal layer) that does not expose the underlying layer. [SEMI G8-94] Also called bubble metal, blister metallization, and bubble metallization. Also see package.

BOAC : See copper bonding over active circuit.

Bonding pads n : relatively large metal areas on a die used for electrical contact with a package or prove pins. [SEMATECH]

Boundary scan n : a scan path that allows the input/output pads of an integrated circuit to be both controlled and observed. [1994 National Technology Roadmap for Semiconductors]

Bridge n 1: a defect in which two adjacent areas connect because of mis-processing such as poor lithography, particle contamination, underdevelop, or etch problems. [SEMATECH] Also called short. 2 : software that allows access to, and combination of, data from incompatible databases. [SEMATECH]

Bridging fault n : a fault modeled as a short-circuit between two nets on a die. [1994 National Technology Roadmap for Semiconductors]

Brightfield illumination n : (transmission electron microscopy) : the illumination of an object so that it appears on a bright background.

Buffered hydrofluoric acid n : an extremely hazardous corrosive used to etch silicon dioxide from a wafer. This aid has a 20- to 30-minute reaction delay after contact with skin or eyes. [SEMATECH]

Built-in self test (BIST) : any of the methods of testing an integrated circuit (IC) that uses special circuits designed into the IC. This circuitry then performs test functions on the IC and signals whether the parts of the IC covered by the BIST circuits are working properly. [1994 National Technology Roadmap for Semiconductors]

Buried contact n : a conductive region between two less conductive regions. [SEMATECH]

Buried layer n 1 : a conductive layer between two less conductive films; for example, a localized n+ region in a p-type wafer that reduces the npn collector series resistance for integrated circuit transistors fabricated in an n-type epitaxial layer deposited on the p-type wafer. [SEMATECH] 2: in epitaxial silicon wafers, a diffused region in a substrate that is, or is intended to be, covered with an epitaxial layer. [SEMI M18-94 and ASTM F1241] Also called subdiffused layer and diffusion under film.

Burn-in n : the process of exercising an integrated circuit at elevated voltage and temperature. This process accelerates failure normally seen as "infant mortality" in a chip. The resultant test product is of high quality. [1994 National Technology Roadmap for Semiconductors] Also see infant mortality.

C4 (controlled collapse chip connect) : see flip chip.

C-AFM : see conductive atomic force microscopy.

Cap deposition : see passivation.

Carrier n 1 : an entity capable of carrying electric charge through a solid; for example, mobile holes and condition electrons in semiconductors. [SEMI M1-94 and ASTM F1241] Also called charge carrier. Also see majority carrier and minority carrier. 2 : slang for wafer carrier. [SEMATECH]

Cavity-down packages n : in cofired ceramic packages, packages on which the die surface faces the mounting board. [SEMI G61-94]

Cavity-up packages n : in cofired ceramic packages, packages on which the die surface faces away from the mounting board. [SEMI G61-94]

Cerdip : abbreviation for ceramic dual-in-line package.

Channel : the portion of a MOS integrated circuit that allows current to flow when biased by the overlying gate region [Sandia Labs]

Chemical-mechanical polish (CMP) n : a process for the removal of surface material from a wafer. The process uses chemical and mechanical actions to achieve a mirror-like surface for subsequent processing. [SEMI M1-94 and ASTM F1241] Contrast physical vapor deposition.

Chem-mech polish : See chemical-mechanical polish.

Chip n 1 : in semiconductor wafers, a region where material has been unintentionally removed from the surface or edge of the wafer. [ASTM F1241] Contrast indent. 2 : see die. 3 : in packaging, a region of material missing from a component; for example, ceramic from a package or solder from a preform. The region does not process completely through the component and is formed after the component is manufactured. The chip size is given by its length, width, and depth from a projection of the design plan-form. [SEMI G61-94] Also called chip-out. Contrast pit. 4 : in flat panel display substrates, a region of material missing from the edge of the glass substrate, which is sometimes caused by breakage or handling. [SEMI D9-94]

Chip carrier (CC) n : a small footprint semiconductor package generally with terminals on all four sides. The package may be manufactured by cofired ceramic or multilayer printed circuit board technologies. [SEMATECH] Also see castellation and ceramic chip carrier

Chip-out : see chip.

Circuit n : the combination of a number of connected electrical elements or parts to accomplish a desired function. [SEMATECH]

Circuit design n : techniques used to connect active (transistors) and passive (resistors, capacitors, and inductors) elements in a manner to perform a function (that is, logic, analog). [1994 National Technology Roadmap for Semiconductors]

Circuit geometries n : the relative shapes and sizes of features on a die. [SEMATECH]

CMOS : see complementary metal oxide semiconductor.

CMP : see chemical-mechanical polish.

COA : see copper over anything

Comet n : on a substrate, a buildup of resist shaped like a comet and generated by a defect. [SEMI P3-90] Also called motorboat.

Complementary metal oxide semiconductor (CMOS) n: a fabrication process that incorporates p-channel and n-channel MOS transistors within the same silicon substrate. [SEMATECH]

Component n 1 : an individual electronic part, such as a device, diode, or capacitor that is fabricated in a metal oxide semiconductor or bipolar process. [SEMATECH] 2 : an individual piece or a complete assembly of individual pieces, including industrial products that are manufactured as independent units, capable of being joined with other pieces or components. The typical components referred to by the specification are valves, fittings, regulators, gauges, instrument sensors, a single length of tubing, several pieces of tubing welded together, tubing welded to fittings, and the like. [SEMI F1-90] 3 : the fundamental parts of an object, its entities, or relationships. [SEMATECH] 4 : the hardware and software that work in sets (functional entities) to perform the operation (s). [SEMATECH]

Conchoidal fracture n : a fracture having smooth convexities and concavities like a clamshell. [SEMATECH] Also see chip.

Conductive atomic force microscopy (C-AFM): An SPM technique that has been widely used for electrical characterization of dielectric film and gate oxide integrity. [ISTFA]

Conductor n : refers to a microscope design with superior abilities to image submicron features on a wafer. [1994 national Technology Roadmap for Semiconductors]

Contact n : in an oxide layer, an opening that allows electrical connection between metal and silicon layers. [SEMATECH] Also see window and via.

Contamination n 1 : the presence of particles, chemicals and other undesirable substances, such as on or in a process tool, in a process liquid or in a clean room environment. [SEMATECH] Also see area contamination and particulate contamination. 2 : three-dimensional foreign material adhering to a package)plastic or ceramic) or leadframe, or parent material displaced from its normal location and similarly adhered. Adherence means that the particle cannot be removed by an air or nitrogen blast at 20 psi. [SEMATECH] Also see foreign material and stain.

Controlled collapse chip connect (C4) : see flip chip

Copper bonding over active circuit: growing copper over any underlying aluminum metallization. Bonding is done on the top most copper layer. [ISTFA]

Copper over anything : having copper over anything but does not require copper on the top layer.

Correlation n 1 : a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone. [Webster's Dictionary]

Crack n 1 : on semiconductor wafers, a cleavage or fracture that extends to the surface and may or may not pass through the entire thickness of the wafer. [ASTM F1241] 2 : of a semiconductor package or solder preform, a cleavage or fracture that extends to the surface. The crack may or may not pass through the entire thickness of the package or preform. [SEMI G61-94] 3 : in flat panel display substrates, a fissure located at the sheet edge or central area. [SEMI D9-94]

Crater n : on the surface of a slice or wafer, an individually distinguishable bowl-shaped cavity. A crater is visible when viewed under diffused illumination. [SEMATECH]

Cratering n : on a slice or wafer, a surface texture of irregular closed ridges with smooth central regions. [ASTM F1241]

Crescents n : structures with parallel major axes, attributed to substrate defects either above or below the surface plane of silicon substrates after epitaxial deposition. [ASTM F1241] Also see fishtails.

Critical area n : the area in which the center of a defect must occur to cause a failure or fault. [SEMATECH] Also see fault and fault probability.

Critical dimension (CD) n : the width of a patterned line or the distance between two lines, monitored to maintain device performance consistency; that dimension of a specified geometry that must be within design tolerances. [ASTM F127-84] Also see linewidth.

Crosstalk n: the undesirable addition of one signal to another in a circuit usually caused by coupling through parasitic elements. An example would be inductive or capacitive coupling between adjacent conductors. [1994 National Technology Roadmap for Semiconductors]

Crossunder n : on a die, the point at which a conductor crosses under a second conductor without making electrical contact. [SEMATECH]

Crow's foot n : on a semiconductor wafer, intersecting cracks in a pattern resembling a "crow's foot" Y on {111} surfaces and a cross "+" on {100} surfaces. [ASTM F1241]

Crystal n : a solid composed of atoms, ions, or molecules arranged in a pattern that is periodic in three dimensions. [ASTM F1241]

Crystal defect n : departure from the regular arrangement of atoms in the ideal crystal lattice. [ASTM F1241] Also see crystal lattice and damage.

Crystal indices : see Miller indices. Also see crystallographic notation.

Crystal lattice n : in a crystal, the three-dimensional and repeating pattern of atoms. [SEMATECH]

Crystallographic notation n : a symbolism based on Miller indices used to label planes and directions in a crystal as follows: (111) plan [111] direction {111} family of planes <111 family of directions. [SEMI M1-94 and ASTM F1241]

Crystal originated particle (COP) n : a surface depression that is formed during soft alkaline chemical treatment of silicon wafer surfaces that contain crystal defects at or close to the wafer surface and that scatters light similarly to a very small particle. [ASTM F1241] Also called surface micro defect.

CTE : see coefficient of thermal expansion.

CVD : see chemical vapor deposition.

Cycle time n : (1) the length of time required for a wafer to complete a specified process or set of processes. [SEMATECH] (2) the length of time required to complete a failure analysis job from receipt in the failure analysis lab to the time results (written or verbal) are communicated back to the immediate requestor. [Sandia Labs] Also see equipment cycle, minimum theoretical cycle time, and theoretical cycle time.

Damascene n : an integrated circuit process by which a metal conductor pattern is embedded in a dielectric film on the silicon substrate. The result is a planar interconnection layer. The creation of a damascene structure most often involves chemical mechanical polishing of a nonplanar surface resulting from multiple process steps. A damascene trench is a filled trench. [1994 National Technology Roadmap for Semiconductors]

Damage n 1 : of a single-crystal silicon specimen, a defect of the crystal lattice in the form of irreversible deformation that results from mechanical surface treatments such as sawing, lapping, grinding, sandblasting, and shot peening at room temperature without subsequent heat treatments. [ASTM F1241] Also see crystal lattice. 2 : any yield or reliability detractors other than those related to design, process specification violations, or particles. [SEMATECH]

DC test : A sequence of direct current (DC) measurements performed on integrated circuit pads to determine probe

contact, leakage currents, voltage levels on input and output, power supply currents, etc. [1994 national Technology Roadmap or Semiconductors]

Deep level impurity n : a chemical element that, when introduced into a semiconductor , has an energy level (or levels) that lies on the midrange of the forbidden energy gap, between the energy levels of the dopant impurity species. [ASTM F1241]

Defect n : for silicon crystals, a chemical or structural irregularity that degrades the ideal silicon crystal structure or the thin films built over the silicon wafer. 2 : a pit, tear, groove, inclusion, grain boundary, or other surface feature that is either characteristic of the material or a result of its processing and that is not a result of the sample preparation. [SEMATECH] Also called anomaly.

Defect density n : the number of imperfections per unit area, where imperfections are specified by type and dimension. [ASTM F127-84] Also see defect.

Defect level n : the number of die in parts-per-million that are shipped to customers and that are defective en though the test program declares them to be good. [1994 National technology Roadmap for Semiconductors]

Defect, photomask n : any flaw or imperfection in the opaque coating or functional pattern that will reproduce itself in a resist film to such a degree that it is pernicious to the proper functioning of the microelectronic device being fabricated. [SEMI P2-86]

Deformation n : a difference between the IC structure of the fabricated device and the desired structure of the nominal device. [ISTFA]

Deforming event : The deviation from nominal manufacturing conditions resulting in an IC deformation. [ISTFA]

Delamination n : in a cofired ceramic package, chip carrier, dual inline, pin grid array, etc., the separation of one ceramic layer from another . [SEMI G61-94] Also see package.

Delay fault n : a fault that has the effect of causing a signal to appear late in arriving at a destination. [1994 National Technology Roadmap for Semiconductors]

Design for test (DFT) n : design of logic circuits to facilitate electrical testing. [SEMATECH]

Destructive physical analysis n 1: the examination and testing of components to ensure proper operation and behavior. [Sandia labs]

Device n : a specific kind of electronic component (such as a MOS transistor, resistor, diode or capacitor) on a die. The diode and transistor are referred to as active devices the capacitor and resistor, as passive devices. [SEMATECH]

Dew point n : the temperature at which liquid first condenses when vapor is cooled. [SEMI C3-94]

DFT : see design for test.

Die n (sing or pl) : a small piece of silicon wafer, bounded by adjacent scribe lines in the horizontal and vertical directions, that contains the complete device being manufactured. [SEMATECH] Also called chip and microchip. Obsolete: bar, slice.

Die attach area n : the nominal area designated for die attaching to the package or leadframe. [SEMI G22-86] Contrast effective die attach area and die attach pad.

Die attach pad n : the nominal area designated for die attaching to the package or leadframe. Die attach pad is usually applied to leadframes. The term die attach area is usually applied to ceramic packages. [SEMATECH] Also see package and die.

Die attach surface n : in a ceramic semiconductor package, a dimensional outline designated for die attach. [SEMI G33-90] Also see package and die.

Die bonding (D/B) : an assembly technique that bonds the back side of an integrated circuit die to a substrate, header or leadframe. [SEMATECH]

Dielectric n 1 : a nonconductive material; an insulator. Examples are silicon dioxide and silicon nitride. [SEMATECH] 2 : a material applied to the surface of a ceramic or preformed plastic package to provide functions such as electrical insulation, passivation of underlying metallization, and limitations to solder flow. [SEMI G33-90]

Dielectric isolation (DI) n : a nonconductive barrier layer grown or deposited between two adjacent regions on a die to prevent electrical contact between the regions. [SEMATECH] Also see isolation.

Diffusion n : a high-temperature process in which desired chemicals (dopants) on a wafer are redistributed within the silicon to form a device component. [SEMATECH]

Dimple n : on a semiconductor wafer, a shallow depression in a wafer surface with a concave, spherical shape and gently sloping sides. NOTE – Dimples are macroscopic features that are visible to the unaided eye under proper lighting conditions. [ASTM F1241]

DIP : see dual inline package.

Dislocation n : a line imperfection in a crystal that either forms the boundary between slipped and nonslipped areas of a crystal or that is characterized by a closure failure of the

Burger's circuit. [ASTM F1241] Also called line defect. Also see slip.

Dopant n : in silicon technology, a chemical element incorporated in trace amounts in a semiconductor crystal or epitaxial layer to establish its conductivity type and resistivity. [Adapted from SEMI M9-90] Also see conductivity type, n-type, and p-type.

Dopant density n : in an uncompensated extrinsic semiconductor, the number of dopant impurity atoms per unit volume, usually given in atoms/cm^3, although the SI unit is atoms/m^3. Symbols: ND for donor impurities and NA for acceptor impurities. [ASTM F1241]

Doping n : the addition of impurities to a semiconductor to control the electrical resistivity. [SEMI M1-94 and ASTM F1241]

Drain n : one of the three major parts of a complementary metal oxide semiconductor transistor. [SEMATECH]

EBIC : See Electron Beam Induced Current.

Edge crown n : an increase of epitaxial layer thickness around the periphery of the wafer arising from differences in deposition rate. [SEMATECH]

EFM : See Electrostatic Force Microscopy

Electron Beam Induced Current (EBIC) : A failure analysis technique for electronic devices that displays the external current resulting from the electron hole pairs generated from the interaction of the electron beam and the PN junction. [ISTFA]

Electromagnetic Interference (EMI) n : any electrical signal in the non-ionizing (sub-optical) portion of the electromagnetic spectrum with the potential to cause an undesired response in electronic equipment. [SEMI E33-94]

Electrostatic discharge (ESD) n 1: a sudden electric current flow, such as between a human body and a metal oxide semiconductor, with potential damage to the component. [SEMATECH] 2 : the transfer of electrostatic charge between bodies at different electrostatic potentials. [SEMI E33-94]

Electrostatic force Microscopy (EFM) : A failure analysis tool that uses a repetitive single-pulse sampling approach to measure high frequency internal signals from ICs. [ISTFA]

Energy-dispersive X-ray spectrometer n : a detector used to determine which elements are present in a sample by analyzing X-ray fluorescence for energy levels that are characteristic of each element. [SEMATECH]

Epitaxial layer n : in semiconductor technology, a layer of a single crystal semiconducting material grown on a host substrate which determines its orientation. [SEMI M2-94 and ASTM F1241]

Epitaxy (epi) n : a silicon crystal layer grown on top of a silicon wafer that exhibits the same crystal structure orientation as the substrate wafer with a dissimilar doping type or concentration or both. Examples are p/p+, n/n+, n/p, and n/n. [SEMATECH] Also see epitaxial layer.

ESD : see electrostatic discharge.

Etch 1 n : a category of lithographic processes that remove material from selected areas of a die. Examples are nitride etches and oxide etches. [SEMATECH] 2 : in the manufacture of silicon wafers, a solution, a mixture of solutions, or a mixture of gases that attacks the surfaces of a film or substrate, removing material either selectively or nonselectively. [SEMI M1-94 and ASTM F1241] Also see anisotropic etch, preferential etch, dry plasma etch, reactive ion etch, and wet chemical etch.

Etchant n : an acid or base (in either liquid or gaseous state) used to remove unprotected areas of a wafer layer. Examples are potassium hydroxide, buffered oxide etch, and sulfur hexafluoride. [SEMATECH]

Etch pit n : a pit, resulting from preferential etching, localized on the surface of a wafer at a crystal defect or stressed region. [ASTM F1241]

Eutectic n : alloy or solution with components distributed in the proportions necessary to minimize the melting point. [SEMATECH] Also see azeotrope.

Excessive leakage n : in the testing of semiconductors, current that is above the specified limit for the particular test being conducted. [Sandia Labs]

Failure n : in failure analysis, an event where the semiconductor component does not function according to its intended use or specifications. [Sandia Labs]

Failure mechanism n : in failure analysis, a fundamental process or defect responsible for a failure. [SEMATECH]

Failure mode n : in failure analysis, the electrical symptoms by which a failure is observed to occur. Failure mode types include a catastrophic failure that is both sudden and complete and degraded failure that is gradual, partial, or both, as well as intermittent failures. [Sandia Labs]

Failure mode and effects analysis (FEMA) n : an analytically derived identification of the conceivable semiconductor failure modes and the potential adverse effects of those modes on the system and mission. [SEMATECH]

Failure pattern instance : a group of failing SRAM cells attributed to the same deforming event. [ISTFA]

Failure pattern type : A set of similar failure pattern instances. [ISTFA]

Fault n 1 : an accidental condition that causes a functional unit to fail to perform its required function. [SEMATECH] 2 : a defect causing out-of-spec operation of an integrated circuit. [SEMATECH] Also see exception condition and defect.

Fault coverage n : the percentage of a particular fault type that a test vector set will detect when applied to a chip. [1994 National Technology Roadmap for Semiconductors]

Fault dictionary n : a list of faults that a test vector will detect in a failing circuit, or a list of all such faults for each vector in a vector set. [1994 National Technology Roadmap for Semiconductors]

Fault model n : a model of the behavior of defective circuitry in an integrated circuit. Physical defects result in improper behavior in a circuit which must be modeled in order for test patterns to be a model, timing model, and bridging model. [1994 National Technology Roadmap for Semiconductors]

FET : see field-effect transistor.

FIB : see focused ion beam

Field-effect transistor (FET) n : a transistor consisting of a source, gate and drain, the action of which depends on the flow of majority carriers past the gate from the source to the drain. The flow is controlled by the transverse electric field under the gate. [SEMATECH]

Fishtails n : structures, attributed to substrate defects, either above or below the surface plane after epitaxial deposition; the "tails" are aligned in a particular crystallographic direction. [ASTM F1241] Also see crescents.

Fissure : see crack.

Flake n : material missing from one byt not the other side of a semiconductor wafer [SEMI M10-89]

Flake chip : see chip and peripheral chip.

Flaking : see peeling.

Flip-chip n : a leadless, monolithic structure that contains an integrated circuit designed to electrically and mechanically interconnect to a hybrid circuit. Connection is made to bump contacts covered with a conductive bonding agent on the face of the hybrid. [SEMATECH] Also called controlled collapse chip connect or C4

Fluorescence n : the emission of light as the result of, and only during, the absorption of radiation of shorter wavelengths. [IEEE]

Fluorescent Microthermographic Imaging (FMI) n : a failure analysis technique that uses a temperature dependent fluorescent compound and an optical pumping source to image temperature changes on a semiconductor device with near optical spatial resolution. [Sandia labs]

FMEA : see failure mode and effects analysis

FMI : see Fluorescent Microthermographic Imaging.

Focused ion beam (FIB) n : an imaging tool that can be used to deposit or etch materials on wafers. A focused ion beam is often used in the etch mode to selectively cleave structures for failure analysis. It is also used in photomask repair for removing or adding material, as necessary, to make the photomask defect free. [SEMATECH]

Fourier transform infrared spectroscopy: see Micro Fourier transform infrared spectroscopy.

FPI : see failure pattern instance.

FPT : see failure pattern type

Front end of line (FEOL) n 1: in semiconductor processing technology, all processes from wafer start through final contact window processing. [SEMATECH]

FTIR : see Micro Fourier transform infrared spectroscopy.

Functional pattern : see pattern, functional.

Functional probe n : the electronic testing of die on a wafer to determine conformance to specifications. [SEMATECH]

Functional test n : one or more tests to determine whether a circuit's logic behavior is correct. [1994 National Technology Roadmap for Semiconductors]

Galvanic corrosion: corrosion damage induced when two dissimilar metals coupled in a corrosive electrolyte. This can be seen in either wafer fab process or assembly process. [corrosiondoctor.org]

Gate n : an electrode that regulates the flow of current in a metal oxide semiconductor transistor. [SEMATECH]

Gate electrode n : the electrode of a metal oxide semiconductor field effect transistor (MOSFET); it controls the flow of electrical current between the source and the drain. [SEMATECH]

Gate oxide n : a thin, high –quality silicon dioxide film that separates the gate electrode of a metal oxide semiconductor transistor from the electrically conducting channel in the silicon. [SEMATECH]

Glass n : a deposited film of silicon dioxide with additives to adjust coefficient of thermal expansion, color, conductivity,

and melting point, generally doped with boron or phosphorous or both. [SEMATECH] Also see silicon dioxide.

Groove n : in a semiconductor wafer, a shallow scratch with rounded edges that is usually the remnant of a scratch not completely removed by polishing. [SEMI M1-94 and ASTM F1241]

Growth hillock : see pyramid.

Hermetic seal n : a coat applied in the final stage of thermal processing to seal the ceramic package and to protect the device from the external environment. [SEMATECH]

Hillock n : a defect caused by stress that raises portions of a metal (such as aluminum) film above the surface of the film. Localized stress within the metal film may elevate portions of the film through the adjacent dielectric layer, resulting in a metal extrusion and a short to the next metal layer. [SEMATECH] Also see pyramid.

Hole n 1 : of a semiconductor, a mobile vacancy in the electronic valence structure that acts like a positive electron charge with a positive electron charge with positive mass; the majority carrier in p-type. [SEMI M1-94 and ASTM F1241] 2 : in plastic and metal wafer carriers, the area through which a pin from another wafer carrier can enter for the transfer of wafers. [SEMI E1-86] Also see wafer carrier.

Hot carriers n : those carriers, which may be either electrons or holes that have been accelerated by the large traverse electric field between the source and the drain regions of a metal oxide semiconductor field - effect transistor (MOSFET). They can jeopardize the reliability of a semiconductor device when these carriers are scattered (that is, deflected) by phonons, ionized donors or acceptors, or other carriers. The scattering phenomenon can manifest itself as substrate current, gate current, or trapped charges. [SEMATECH] Also see trapped charges.

IC : see integrated circuit.

IDDQ : abbreviation for direct drain quiescent current. See static current test.

Impact test n : in component testing, a test performed to determine particle contribution as a result of mechanical shock to the component. [SEMATECH] Also called particle impact noise detection or PIND

Implant : see ion implantation.

Impurity n : a chemical or element added to silicon to change the electrical properties of the material. [SEMATECH] Also see dopant, ion implantation.

Inclusion n : discrete second phases (oxides, sulfides, carbides, intermetallic compounds) that are distributed in a metal matrix. [SEMATECH]

Indent n : on a semiconductor wafer, an edge defect that extends from the front surface to the back surface. [ASTM F1241] Contrast chip.

Insulator n : a substance that will not conduct electricity; for example, silicon dioxide and silicon nitride. [SEMATECH] Contrast conductor.

Integrated circuit (IC) n 1: two or more interconnected circuit elements on a single die. [SEMATECH] 2 : a fabrication technology that combines most of the components of a circuit on a single-crystal silicon wafer. [SEMI Materials, Vol. 3, Definitions for Semiconductor Materials]

Interference contrast microscope n : an optical microscope that reveals surface details of an object in which there is no appreciable absorption by using the interference between two beams of light. [Adapted from ASTM F1241] Also called Nomarski Interference Contrast

Interlevel dielectric n : an insulating film between two conductive film layers, as between poly and aluminum or between layers of aluminum. [SEMATECH]

Interstitial n : in a crystalline solid, an atom that is not located on a lattice site. [SEMATECH]

Intrinsic semiconductor n : a semiconductor in which the density of electrons and holes is approximately equal . [SEMATECH] Contrast extrinsic semiconductor

Ion implantation (I 2 , II) n : a high-energy process that injects an ionized species such as boron, phosphorus, arsenic, or other ions into a semiconductor substrate. [SEMATECH]

I/O pins n : connections to an integrated circuit through which input and/or output (I/O) signals pass. [1994 National Technology Roadmap for Semiconductors]

Isolation n : an electrical separation of regions of silicon on a wafer; for example, boron diffusion to isolate a transistor. [SEMATECH] Also see dielectric isolation.

Junction spiking n : the penetration of a junction by aluminum, which occurs when silicon near the junction dissolves in aluminum and migrates along the interconnect lines. Aluminum then replaces silicon at the junction. [SEMATECH]

Kirkendall void n : voids induced in a diffusion couple between two metals that have different interdiffusion coefficients. [SEMATECH]

Large scale integration (LSI) n : the placement of between 100 and 1000active devices on a single die. [SEMATECH]

Laser-scattering light event n: a signal pulse that exceeds a preset threshold, generated by the interaction of a laser beam

with a localized light scatterer (LLS) at a wafer surface as sensed by a detector. [ASTM F1241]

Laser Voltage Probe (LVP) : a technique that measures the photons absorbed within a junction area. [ISTFA]

Layout n 1 : the physical geometry of a circuit or die. [1994 National Technology Roadmap for Semiconductors] 2 : the process of creating the physical geometry of a circuit or die. [1994 National Technology Roadmap for Semiconductors] 3 : see composite drawing.

LDD : see lightly doped drain.

Life test n : in semiconductor reliability, a test designed to operate the semiconductor until it fails by elevating both temperature and voltage to accelerate the aging process. [Sandia Labs]

Lightly doped drain (LDD) n : a metal-oxide semiconductor (MOS) device design in which the drain doping is reduced to improve breakdown voltage. [SEMATECH]

Line defect : see dislocation.

Liquid-metal embrittlement: The brittle failure of a normally ductile metal when in the presence of a liquid metal and subsequently stressed in tension. [ISTFA]

Light Induced Voltage Alteration (LIVA): An advanced failure analysis technique that utilizes a laser scanning microscope and various electronics to produce localized photons within the semiconductor device circuitry. [ISTFA]

LSI : see large scale integration.

LVP : see laser voltage probe.

Metallization void n : the absence of a clad, evaporated, plated or screen printed metal layer or braze from a designated area. [SEMI G58-94] Also called metal void.

Metal void : see metallization void.

MFM-CCI : see magnetic current imaging.

Microchip : see die.

Micro Fourier transform infrared spectroscopy (FTIR) : a technique, considered to be non-destructive, based on infrared photon absorbtion typically used to determine polymers, plastics, fibers, and organic films with an approximate sensitivity of 1 - 10 ppm with a resolution down to ~15 µm. Also called Fourier transform infrared spectroscopy.

Miller Indices : a numeric convention that describes the orientation of the atomic planes in a crystal lattice, i.e., the crystal faces, of a crystalline material. Also called crystal indices. Also see crystallographic notation.

Moon crater n : on a semiconductor wafer, surface texture that results when a wafer floats during the initial stages of chemical polishing in a rotating cup etcher. [ASTM F1241]

Motorboat : see comet.

Mottled adj : pertaining to the existence on a wafer of a of material in a window that prevents the window from being properly opened. [SEMATECH]

Mound n : on a semiconductor wafer, an irregularly shaped projection on a semiconductor wafer surface with one or more irregularly developed facets. [ASTM F1241] Contrast pyramid. Also called mouse bite.

Nano hardness tester (NHT): a technique that places a weighted top on the film surface and measures the maximum load that can be applied without deforming the film. [ISTFA]

Nick : see chip

NHT : see Nano hardness tester

Notch n 1: an unexpected intrusion or reduction of line width in patterned geometries. May also be a V-shaped intrusion into the perimeter of a wafer. The intrusion is used to align the wafer during process. [SEMATECH] 2 : on a semiconductor wafer, an intentionally fabricated indent of specified shape and dimensions oriented such that the diameter passing through the center of the notch is parallel with a specified low index crystal direction. [SEMI M1-94 and STM F1241]

Oil canning n : in metal lid/perform assembly, lid concavity after sealing. [SEMI G53-92]

Optical Beam Induced Resistance CHange (OBIRCH): An advanced failure analysis technique that uses a laser scanning microscope to detect a change in the current under constant voltage conditions. [ISTFA]

Overcoat : see passivation

Oxide defect n : an area of missing oxide on the back surface of back-sealed wafers discernible to the unaided eye. [ASTM F1241]

Oxide etch n : an etch process in which unprotected areas of the oxide layer are eroded by use of a chemical to expose the underlying layer. [SEMATECH]

Parametric test n : wafer-level testing of discrete devices such as transistors and resistors. [SEMATECH]

Parasitics n : unwanted circuit components (for example, capacitors or resistors) present in a design. [1994 National Technology Roadmap for Semiconductors]

Particle n 1: a minute quantity of solid or liquid matter. [SEMATECH] Also called dirt. 2 : in the manufacture of photolithographic pellicles, material that can be distinguished from the film, whether on the film surface or embedded in the film. [SEMI P5-94] 3 : the re-plating step in which a catalytic material, often a palladium or gold compound, is absorbed on a surface to act as sites for initial stages of deposition. [ASTM B374-93]

Particulate 1 n: discrete particles of dirt or other material. [ASTM F1241] Also see dirt. 2 n (dust) : discrete particle of material that can usually be removed by (nonetching) cleaning. [SEMI M10-89] 3 adj : describes material in small, discrete pieces; anything that is not a fiber and has an aspect ratio of less than 3 to 1. Examples are dusts, fumes, smokes, mists, and fogs. [SEMATECH]

Particulate contamination n : on a semiconductor wafer, a particle or particles on the surface of the wafer. [ASTM F1241]

Passivation n : deposition of a scratch-resistant material, such as silicon nitride and/or silicon dioxide, to prevent deterioration of electronic properties caused by water, ions, and other external contaminants. The final deposition layer in processing. [SEMATECH] Also called overcoat and cap deposition.

Peeling n: any separation of a plated, vacuum deposited, or clad metal layer from the base metal of a leadframe, lead, pin, heat sink or seal ring from an under plate, or from a refractory metal on a ceramic package. Peeling exposes the underlying material. [SEMI G61-94] Also called flaking. Contrast blister metal.

Peripheral chip n 1 : crystallographic damage along the circumference formed in the periphery of the specimen through conchoidal fracture and resultant spalling. [ASTM F1241] Also called flake chip or surface chip.

Picosecond Imaging Circuit Analysis (PICA): A technique that measures time-dependent hot carrier induced light emission from the integrated circuit (IC) both spatially and temporally, thus enabling failure analysis and timing evaluation of a device. [ISTFA]

Pinhole n 1 : minute defect or void in a film, mask , or resist, usually the result of contaminants. [SEMATECH] 2 : a small opening that extends through a covering, such as a resist coating or an oxide layer on a wafer. [SEMI P2-86]

Pit n 1 : in a wafer surface, a depression in a wafer surface that has steeply sloped sides which meet the surface in a distinguishable manner, in contrast to the rounded sides of a dimple. [ASTM F1241] Also see slip and dislocations. 2 : in semiconductor packages, plastic or ceramic, or in the lead-frames, a shallow depression or crater. The bottom of the depression must be visible in order for the term to apply. A pit is formed during the component manufacture. [SEMI G61-94] Contrast chip. 3 : in flat panel display substrates, a small indentation on the glass substrate surface. [SEMI D9-94]

Point defect n : a localized crystal defect such as a lattice vacancy, interstitial atom, or substitutional impurity. [ASTM F1241] Contrast with localized light scatter,

Poly : see polycrystalline silicon.

Polycrystalline adj : describes a form of semiconductor material made up of randomly oriented crystallites and containing large angle grain boundaries, twin boundaries, or both. [SEMI M10-89 and ASTM F1241] Contrast single crystal. Also see amorphous silicon.

Polycrystalline silicon (poly) n 1 : a nonporous form of silicon made up of randomly oriented crystallites or domains, including glassy or amorphous silicon layers. [ASTM F399-88] Also called poly and poly silicon. Contrast amorphous silicon and single crystal.

Polysilicon: see polycrystalline silicon.

Precipitate n 1 : within a silicon lattice, a region of silicon oxide frequently manifested as an etch pit. [ASTM F1241] Also see crystal lattice and pit. 2 : in a gallium arsenide wafer, a localized concentration of dopant that is insoluble. Precipitate is formed during crystal growth and during any process in which the temperature is sufficient to provide the necessary impurity mobility. [SEMI M10-89]

Process-induced defect (PID) n : defect(s) added to the wafer as a result of a processing step. The PID wafer undergoes the same process sequence as a product wafer. PID wafer data is a closer approximation of actual process defect contributions than particles per wafer pass (PWP) wafer data. [SEMATECH]

Pyramid n : a structure displaying [111] facets that appears on surfaces after epitaxial growth. DISCUSSION – a pyramid originates at the interface of the substrate and the epi layer and is due to various imperfections at the beginning of epi growth. [ASTM F1241] Also called growth hillock. Also see hillock and mount.

Quiescent Signal Analysis (QSA) : an electrical test based diagnostic technique that uses I_{DDQ} measurements made at multiple chip supply pads as a means of locating shorting defects in the layout. [ISTFA]

QSA : see quiescent signal analysis.

Reactive Ion Etching (RIE) n : a dry-etch process using electrical discharge to ionize and induce ion bombardment of the wafer surface to obtain the required etch properties. [SEMATECH]

Registration n 1 : the accuracy of the relative position of all functional patterns on any reticle with the corresponding

patterns of any other reticle of a given device series when the reticles are properly superimposed. [ASTM F127-84] 2 : a vector quantity defined at every point on the wafer. It is the difference, R, between the vector position, P1, of a substrate geometry and the vector position of the corresponding point, P0, in reference grid. [SEMATECH] 3 : in the overlay capabilities of wafer steppers, a vector quantity defined at every point on the wafer. It is the difference, R, between the vector position, P1, of a substrate geometry, and the vector position of the corresponding point, P0, in a reference grid. [SEMI P18-92]

Residue n : any undesirable material that remains on a substrate after any process step. [ASTM F127-84 and SEMI P3-90]

Resistive Interconnection Localization (RIL): a scanning laser microscope analysis technique that directly and rapidly localizes defective vias, contacts, and conductors from the front side and backside. [ISTFA]

RIE : See Reactive Ion Etching

RIL : See resistive interconnection localization

Root cause n 1 : in failure analysis, the fundamental incident or condition that initially caused the failure to occur. [Sandia Lab]

SAM : see Auger electron spectroscopy

Saucer pits : see shallow etch pits.

Saw-blade defect n 1 : on semiconductor wafers, a roughened area visible after polishing with a pattern characteristic of the saw blade travel. [STM F1241] Also see saw marks. 2 : a depression in the wafer surface made by the blade, which may not be visible before polishing. [SEMI M10-89]

Saw exit chip n : in gallium arsenide technology, an edge fragment on a wafer broken off at the point at which the saw completed its cut the wafer. A saw exit chip is typically straight or arc shaped, not irregular, and sometimes can be confused with the orientation flats. [SEMI M10-89] Contrast saw exit mark.

Saw exit mark n : in silicon technology, a ragged edge at the periphery of a wafer consisting of numerous adjacent small adjoining edge chips resulting from saw blade exit. [ASTM F1241] Also see saw marks, saw exit chip.

Saw-kerf : see scribe line

Saw marks n : on a wafer, surface irregularities in the form of a series of alternating ridges and depressions in arcs, the radii of which are the same as those of the saw blade used for slicing [ASTM F1241] Also see saw exit mark.

Scanning Auger microprobe (SAM) : see Auger electron spectroscopy

Scanning electron microscope (SEM) n : a device that displays an electronically scanned image of a die or wafer for examination on a screen or for transfer onto photographic film; displays a higher magnification than an optical microscope. [SEMATECH]

Scanning Probe Microscopy (SPM) : Provides topographic imaging coupled with a variety of material characterization information such as thermal magnetic, electric, capacitance, resistance and current with nano-meter scale resolution. [ISTFA]

Scanning Thermal Microscope (SThM) : A failure analysis technique for electronic devices that uses a resistive probe to detect temperature distribution with an accuracy of 5mK and a local resolution of 50 nm. [ISTFA]

Scanning tunneling microscope (STM) n : an instrument for producing surface images with atomic scale lateral resolution, in which a fine probe tip is raster scanned over the surface and the resulting tunneling current is monitored. [SEMATECH]

Scratch n : on semiconductor wafers, a shallow groove or cut below the established plane of the surface, with a length to width ratio greater than 5:1 [ASTM F1241] Also see macroscratch, microscratch.

Scum n : resist residue located in a window or along the foot of patterned geometry. [SEMATECH]

Secondary Ion Mass Spectrometry (SIMS) n : an analytical tool that uses a focused ion beam that is directed to a solid surface, removing material in the form of neutral and ionized atoms and molecules. The secondary ions are then accelerated into a mass spectrometer and separated according to their mass-to-charge ratio. [Center for Microanalysis of Materials]

Seebeck Effect Imaging (SEI) n : An advanced failure analysis technique that utilizes infrared laser to isolate a defect within a die by thermally altering the defect's electrical characteristic without biasing the device. [ISTFA]

SEM : see scanning electron microscope.

Semiconductor n : an element that has an electrical resistivity in the range between conductors (such as aluminum) and insulators (such as silicon dioxide). Integrated circuits are typically fabricated in semiconductor materials such as silicon, germanium, or gallium arsenide. [SEMATECH]

Shallow etch pits n : on a wafer etch pits that are small and shallow in depth under high magnification greater than 200X [ASTM F1241] Also called saucer pits. Also see haze

Short : see bridge

Silicon (Si) : a brownish crystalline semimetal used to make the majority of semiconductor wafers. [SEMATECH]

Silicon dioxide (SiO_2) n : a passivation layer thermally grown or deposited on wafers. It is resistant to high temperatures. Oxygen or water vapor is used to grow silicon dioxide at temperatures above 900°C. Silicon dioxide is used as a masking layer as well as an insulator. [SEMATECH] Also called quartz. Also see glass.

Silicon nitride (Si_3N_4) (abbr. SiN) n : a passivation layer chemically deposited on a wafer at temperature of between 600°C and 900 ^{o}C to protect the wafer from contamination . Silicon nitride is also used as a masking layer and as an insulator. [SEMATECH]

Silicon on insulator (SOI) n : a novel substrate for high-performance, low-power, and radiation-hard CMOS applications that offers process simplification, improved scalability, latch-up free and soft-error free operation, improved sub threshold slope, and drastic reduction in parasitic capacitance. At this writing, there are two manufacturing-oriented techniques to build SOI : SIMOX and bonded. [SEMATECH].

SIMS : see Secondary Ion Mass Spectrometry

Slice : see wafer.

Slip n : in semiconductor wafers, a process of plastic deformation in which one part of a crystal undergoes a shear displacement relative to another in a manner that preserves the crystallinity of each part of the material. DISCUSSION – After preferential etching,, slip lines are evidenced by a pattern of one or more parallel straight lines of dislocation etch pits that do not necessarily touch each other. On [111] surface, group of lines are incline at 60° to each other; on [100] surfaces, they are inclined at 90° to each other. [SEMI M10-89 and STM F1241] Also see pit.

Slip line n : a step occurring at the intersection of a slip plane with the surface. [ASTM F1241]

Slip plane n : the crystallographic plane on which the dislocations forming the slip move. [ASTM F1241]

Small scale integration (SSI) n : the placement of between 2 and 10 active devices on a single die. [SEMATECH] Also see die.

Smudge n : dense local area of contamination usually caused by handling or fingerprints. [SEMI M1-94 and ASTM F1241] Also see dirt.

Snowball n : on a semiconductor wafer, a track with the appearance under magnification of a snowball rolled through snow. [ASTM F1241]

Soft defect localization : A laser scanning failure analysis methodology that combines localized heating and detection of changes in the pass/fail probability of a vector set. [ISTFA]

SOI : see silicon on insulator.

SOS : see silicon on sapphire.

Source n : one of the three major components of a CMOS transistor. [SEMATECH]

Spike n 1 : in a epitaxial wafer surface, a tall, thin dendrite or crystalline filament that often occurs at the center or recess. [ASTM F1241] 2 : an extreme structure that has a large ratio of height-to-base width and no apparent relation to epitaxial film thickness. [SEMATECH] Also see pyramid and mound.

SPM : see scanning probe microscopy.

SPP : See statistical post processing.

SQUID (Super-conducting quantum interference device) microscope n. : A tool that uses a highly sensitive magnetic field that works at cryogenic temperatures (90 K) to detect magnetic field position which is then be Fourier transformed to produce current density maps based on Biot-Savart law. [ISTFA]

SSI : see small scale integration.

Stacking fault n : in a crystal, a two-dimensional defect caused by a deviation from the normal stacking sequence of atoms. [ASTM F1241]

Stain n 1 : a solution applied to a cross-sectioned silicon device to reveal the location of various structures. [SEMATECH] 2 : contaminant in the form of streaks that are chemical in nature and cannot be removed except through further lapping or polishing. Examples are “white” stains that are seen after chemical etching as white or brown streaks. [SEMI Materials, Vol. 3, Definitions for Semiconductor Materials] 3 : a two-dimensional, contaminating foreign substance on a component surface. [SEMATECH] Also see contamination and foreign material. 4 : in flat panel display substrates, any erosion of the surface; generally cloudy in appearance, it sometimes exhibits apparent color. [SEMI D9-94] 5 : area contamination that is chemical in nature and cannot be removed except through further lapping or polishing [ASTM F1241]

Step coverage n : the ratio of thickness of film along the walls of a step to the thickness of the film at the bottom of a step. Good step coverage reduces electromigration and high resistance pathways. [SEMATECH]

Statistical post processing (SPP) : a statistical technique that separates the effects of defects from normal wafer-to-wafer variation. [ISTFA]

STM : see scanning tunneling microscope.

SThM : see scanning thermal microscope.

Stuck-at fault n : a fault in a manufactured circuit causing an electrical node to be stuck at a logical value of 1 or a logic value of 0, independent of the input to the circuit. [1994 National Technology Roadmap for Semiconductors]

Substrate n : in the manufacture of semiconductors, a wafer that is the basis for subsequent processing operations in the fabrication of semiconductor devices or circuit. [ASTM F1241]

Surface chip : see peripheral chip

Surface defects n 1 : in the manufacture of silicon on sapphire (SOS) epitaxial silicon wafers, mechanical imperfections, SIO_2 residual dust, and other imperfections visible on the wafer surface. Some examples of surface defects are: dimple, pits, particulates, spots, scratches, smears, hillocks, and polycrystalline regions. [SEMI M4-88] 2 : in flat panel display substrates, a marking, tearing or single line abrasion on the glass surface. [SEMI D9-94]

TCR : See Thermal coefficient of resistance

TDR : See Time domain reflectometry

Tester pattern generation (TPG) n : the generation of a program that runs on an integrated circuit hardware tester (integrated circuit tester). The purpose of this program is to permit test vectors to be applied to determine the performance of the integrated circuit. [1994 National Technology Roadmap for Semiconductors] Also called tester program generation.

Test pattern : see pattern, test.

Test techniques n : any methods used for the expressed purpose of testing integrated circuits. Examples include build-in self test (BIST), automatic test pattern generator (ATPG), static current test (IDDQ), and boundary scan. [1994 National Technology Roadmap for Semiconductors]

Test vectors n : sequences of signals applied to the pins of an integrated circuit to determine whether the integrated circuit is performing as it was designed. [1994 National Technology Roadmap for Semiconductors]

Thermal coefficient of resistance (TCR): A fundamental characteristic of metallization and is dependent only on the residual resistivity of the metal which is determined by differences in the structural order of a pure, bulk metal, that contribute to electron scattering. [ISTFA]

Time Domain Reflectometry (TDR): Analysis technique that uses a low voltage, low current, and very short rise time voltage pulse to determine the impedance of a signal trace as a function of time. [ISTFA]

Tin whiskers : hair-like growths of near-perfect single crystalline structures of tin that grow from some electroplated tin surfaces. [ISTFA]

Thermal Induced Voltage Alteration (TIVA) : An advanced failure analysis technique that uses a laser scanning microscope to locally heat a region to stimulate an alteration in the voltage under constant current conditions. [ISTFA]

Total reflection X-ray fluorescence (TXRF) n : an analytical method usually used to characterize the level of metallic (and non-metallic element) surface contamination. IN TXRF, an X-ray beam excites fluorescence from the contamination that is present on a silicon surface. Since the beam is incident at grazing angles, it totally reflects from the surface, thus maximizing the signal. [SEMATECH]

Trapped charges n : charges trapped either in the gate oxide or, in the case of a lightly doped drain (LDD) metal-oxide semiconductor field-effect transistor (MOSFET), in the spacer region. Trapped charges in the gate or the spacer lead to threshold voltage shift or to transconductance degradation, respectively. [SEMATECH]

TXRF : see total reflection X-ray fluorescence.

Undercutting n : the lateral etching into a substrate under a resistant coating, as at the edge of a resist image. [ASTM F127-84]

Unencapsulated thermal test chip n : an unpackaged, specially designed silicon die with standard test junctions that, after mounting into a package, may be used to thermally characterize that package. This technique is useful in determining the difference between various vendors' packages and package designs. [SEMATECH]

Via n : a connection between two conducting layers above the silicon surface that is created by a different material or deposition step [Sandia Lab]

Void 1 : see dielectric void. 2 : see glass void. 3 : see metallization void.

Wafer n : in semiconductor technology, a thin slice with parallel faces cut from a semiconductor crystal. [ASTM F1241] Also called a slice. Also see substrate.

Well n : a localized n-type region on a p-type wafer or a p-type region on a n-type wafer. [SEMATECH]

X-ray fluorescence n 1 : the property of atoms to absorb X rays and emit light of characteristic wavelengths. [SEMATECH] 2 : a material diagnostic technique that determines the surface concentration of contaminants. [SEMATECH]

X-ray laminography : a radiographic 3-D technology that provides a focused image slice at a selected plane and analyzes the images using a set of algorithms. [ISTFA]

XRL : see X-ray laminography

Author Index

A
Aitken, Rob 190
Alers, G.B. 99
Altmann, F. 330
Antoniou, Nicholas 607

B
Barton, Daniel L. 310
Beaudoin, F. 340
Beermann, Hubert 209
Birdsley, Jeff 16
Bjork, Roger 16
Boit, Christian 279, 340, 627
Breitenstein, O. 330

C
Canumalla, Sridhar 23
Carder, Darrell 218
Cargo, James 445
Chien, Han-Chung 573
Childs, Kenton 561
Cole, Edward I. Jr. 246
Colvin, J. 536

D
Das, Diganta 171
Davis, Brennan 440
De Wolf, Ingrid 52
Dellin, Theodore A. 149
Desplats, R. 340
Dias, Rajen 34

E
Eide, Geir 199
Engel, B. 417

F
Ferrier, M. Steven 1
Flatoff, D. 437
Frank, Steve 159

G
Gattiker, Anne 190
Gloor, Cary A. 239
Görlich, S. 627

H
Hartfield, Cheryl D. 362
Hawkins, Charles F. 128
Henderson, Christopher L. 121, 612
Henry, Leo G. 225
Herlinger, Lowell 440
Herrick, Robert W. 78
Hoffman, Steven 121
Hooghan, Kultaransingh (Bobby) 583
Huang, Y.S. 49

J
Jarausch, K. 536
Johnson, Richard 440

K
Karg, D. 330
Keim, Martin 199
Knauss, L.A. 301
Kolachina, Siva 349

L
Lane, Michael 16
Lee, J.H. 49
Levine, E. 417
Li, Susan Xia 40
Liou, Yung 573
Lundquist, Ted 594

M
McDonald, John 457
Moore, Thomas M. 362
Mount, Gary 573

N
Ng, William 354
Nigh, Phil 190

O
Ong, Ryan 635
Orozco A. 301

P
Palosh, Steve 218
Parente, Renee 440
Perdu, P. 340
Perungulam, Srikanth 397
Petrus, J. 417
Prejean, Seth 440

R
Rai, Raghaw S. 506
Raina, Rajesh 218
Roberts, S. 437
Rosenkranz, Ruediger 269
Ross, Richard J. 617

S
Santana, Mike 440
Scherer, Juergen 561
Schlenker, D. 627
Schmidt, C. 330
Schnabel, Patrick 561
Scholtens, K. 627
Segura, Jaume 128
Shore, A. 417
Silvus, Stan 111
Smolyansky, D. 383
Soden, Jerry 128
Sood, Bhanu 171
Su D.H. 49
Subramanian, Swaminathan 506
Sunter, Stephen 181

T
Tangyunyong, Paiboon 310
Thompson, Mark 594
Thong, John T.L. 263

V
Vallett, D. 292
Vanderlinde, W. 477, 497, 549
Venkataraman, Srikanth 199
Versen, Martin 104
Viswanadham, Puligandla 23

W
Walraven, Jeremy A. 52
Wang, Steve 529
Waterson, Bradley A. 52
Weiland, R. 627
Wills, Kendall Scott 397
Woods, S.I. 301
Wu, Huixian 445

Y
Young, Richard J. 583

Subject Index

A

acceleration models, 125
  Arrhenius model, 125
  Eyring model, 125
  power law model, 125
acceptable quality level (AQL), 122
acoustic microscopy, 362–382
  IC package inspection, 363–364(F)
  imaging flip chip devices, 374–378(F)
  nondestructive techniques, 380(F)
  package inspection, 369–374(F)
  reflected signal analysis, 364–366(F,T)
  scanning acoustic microscopy (SAM), 362–363
  stacked die packages, 378–379(F)
  time of flight imaging, 367–368(F)
  transducer selection, 368–369(T)
active voltage contrast (AVC), 273–275(F,T), 276–278(F)
analog BIST, principles of, 185–188
  serial digital bus, 187(F)
  undersample, 186(F)
analog circuit characterization, 163–170(F)
analog design for test and diagnosis, 181–189
analog device characterization, 159–163(F)
analog DFT methods, 181–182
  analog test bus, 182(F)
  loopback, 182(F)
analysis techniques
  destructive vs. non destructive, 19–20(F)
atomic force microscopy, 536–548
  scanning probe microscopy, 536–546(F)
    atom force probing (AFP), 545–546(F)
    C-AFM and passive voltage contrast, 544–545(F)
    contact mode (DC), 536–537(F)
    electric field, 540(F)
    implant imaging, 540–541
    lateral force, 537
    lift mode, 539(F)
    magnetic field, 539(F)
    nanoindentation, 539(F)
    phase contrast, 538–539(F)
    sample preparation techniques, 546
    scanning capacitance microscopy, 541–542(F)
    scanning spreading resistance microscopy
(SSRM), 542–543(F)
    scanning thermal microscopy (SThM), 540(F)
    tapping mode (AC), 537–538(F)
    tunneling mode, 543–544(F)
Auger electron spectroscopy (AES), 561–572
  applications, 566–571
    buried defects, 569–571(F)
    complex defects, 566–567(F)
    low z defects, 567(F)
    small particles, 566(F)
    thin flakes and residues, 568–569(F)
  Auger spectrum, 562(F)
  device design rules, 561(T)
  schematic diagram of, 561(F)
  submicron defect analysis, 562–565
    analysis volume, 563–565(F)
    chemical information, 565(F)
    compositional identification, 565
automatic test equipment, 195
average outgoing quality level (AOQL), 122(F)

B

backside accessibility, 45(F)
backside deprocessing, 440–444
  deprocessing methods, 441–443(F)
  packaged vs. unpackaged devices, 440(F)
"bathtub" curve for failure rates, 125(F)
beam-based defect localization, 246–262
BGA package, 40(F)
bipolar junction transistors, 161(F)
  characterization, 161–162(F)
BIST, 194
blackbody radiation, 310–312(F)
board level failure mechanisms and analysis, 23–33
burn-in, 126
business model for failure analysis service, 627–634
  classification of analysis jobs, 629–630(T)
  elements of a business model, 627–628
    business process, 627–628(F)
    FA reference lab, 628
    process flows as analysis guideline, 628
  integration of FA, 633
  metrics and target setting, 628–629
    key performance indicators (KPIs), 628–629(F)
  pipeline management, 630–633(F,T)

C

CAD navigation for fast fault isolation, 354–361
  FA and ATPG, 360(F)
  integration of simulation, 359(F)
  Merlin framework, 357–359(F)
  preparation with DRACULA, 355–356(F)
  preparation with Hercules, 356–358(F)
cause-effect relationships, 8–9(F), 14(F)
chip scale packages, 40–48
  categories of, 41(T)
  failure analysis challenges
    die related, 42–43(F)
    package related, 42(F)
circuit edit at first silicon, 594–606
  chemistry assisted deposition, 596–597
    conductor, 596(F)
    insulator, 596–597(F)
  chemistry assisted etching, 595–596

circuit edit at first silicon *(continued)*
aluminum, 595
copper, 595–596(F)
low-k dielectrics, 596
SiO_2, 595
edit operations, 597–598(F,T)
backside editing, 602–603
high aspect ratio milling, 598–599
lateral interconnects, 599–600
over-spray cleanup, 600–601(F)
probe-point creation, 603
reconstructive surgery, 602
trace cutting, 601–602(F)
vertical interconnects, 599(F)
mask/fabrication emulation, 594
science behind, 594–595
success statistics, 603–605(F)
design rules for diagnostics, 604–605
why edits fail, 604
circuit edit through bulk silicon, 607–611
areas of concern, 609–611
charge control, 611
design rules, 611
die distortion, 610
endpoint in silicon, 610
silicon on insulator, 610
backside editing, 608(F)
process flow, 607–609
dielectric deposition, 609
editing the circuit, 609, 611(F)
fine navigation, 609, 610(F)
navigation and trenching, 608, 609(F)
polishing and cleaning, 607–608(F)
thinning, 607
trenching with laser micro chemical technology, 608, 609(F)
CMOS
bridging defects, 135–137(F)
circuits, transient luminescence in, 293–294(F)
inverter, 133–134(F)
open defects, 137–139(F)
submicron devices, 149–158
coefficient of thermal mismatch, IC package, 363(F)
contrapositive test, 3
correlation list, 4–7(T), 10
cost factor, 21
cost saving, 21(F)
counterfeit electronic parts, 171–180
creation of, 173–174
refurbishing, 173
relabeling, 173
repackaging, 174
detection of, 174–175(T)
die inspection, 178(F)
external visual inspection, 175
incoming inspection, 175–176(F)
inspection methods, 179(T)
material characterization, 177–178(F)
packaging evaluation, 178
x-ray inspection, 176–177(F)
limitations of visual inspection, 173(T)
parts used to create counterfeits, 172(T)
processes used to create counterfeits, 174(T)
cross sectioning, 417–436
artifacts, 434–435(F)
cleaning/oxygen ashing, 423
delineation layer, 423(F)
final polish, 428–429(F)
focused ion beam
alternative, 434(F)
complement, 432–433(F)
laboratory logistics, 417–418
laser or FIB marking, 423
method (cleave or polish), 419
fracture cleaving techniques, 420–422(F)
mounting, 424–425(F)
optical inspection, 425
rough grind, 425–428(F)
sample decoration techniques, 431–432(F)
structures, 418–419(F)
surface preparation
final, 430–431(F)
pre cross-section, 422(F)
current imaging, magnetic sensors, 301–309
current-induced voltage alteration (CIVA), 249–251(F)
using low primary beam energies, 250–253(F)
low energy SEM resolution improvements, 253(F)
current sources, 165(F)
curve tracer, 209–216
pulse measurements, 215–216(F)

D

defect localization techniques, 104–110
delayering techniques, 397–416
copper etching, 413–414
copper wet etch recipes, 414(F)
plasma etched copper line, 414(F)
transene copper etch, 414(F)
cross section process, 407–413
ball bond, 408(F)
cross section, device needing delayering, 408(F)
cross section etch, 408(F)
device junctions, 408(F)
solder bump, 407(F)
delayering etching options, 399(F)
demetallization, 409–412(F)
dislocations, 412(F)
items to control in wet chemical delayering, 401(F)
new trends, 414–416(F)
parallel polishing, 405–407
flow, 407(F)
lapping artifacts, 406(F)
modified process flow, 407(F)

parallel lapping tool and fixtures, 407(F)
reasons to, 406(F)
plasma delayering processes, 403–404(F)
plasma deposition processes, 405(F)
plasma processing, 401–402(F)
possible metal etches, list of, 400(F)
SEM micrographs of backside, 413(F)
wet chemical etching vs. plasma etching, 399
delineation etching, 437–439
ion beam milling, 438–439(F)
plasma/reactive ion etching, 438(F)
wet chemical etching, 437–438(F)
deprocessing techniques, 445–456
backside failure analysis, 452–455
SOI, Cu/low k devices, 452–455(F)
copper interconnects, 446–449
CMP and EP, 446–447(F)
dry, wet chemical etching, 447–449(F)
gate oxide integrity and gate level, 450–452(F)
inter-level dielectric layers
reactive ion etching, 449–450(F)
design for test practices, 239
design of experiments (DoE), 122
DFT for common functions, 182–185
ADC and DAC, 184
miscellaneous analog, 185
PLL, 183(F)
RF, 184(F)
SerDes, 183(F)
DMOS power transistor, 209–217
drain-source leakage currents, 213(F)
DRAM failure analysis, 104
bitmapping, 106(F)
limits of, 107–108(F)
DRAM cell, schematic of, 105(F)
DRAM device, 104(F)
electrical measurement techniques, 105
soft defect localization, 108
set up, 109(F)

E

e-beam waveform measurements, 263–266
capacitive coupling voltage contrast, 265–266
open and closed loop measurements, 263–264
sampling and signal-to-noise issues, 264–265
E/C/C, 3–11. *See also* metaprocess
education and training, 612–616
current situation, 613– 614(F)
future of analysis training, 615, 616(F)
historical perspective, 612–613(F)
philosophy of training, 614
process training, 614–615
technique/tool training, 615
technology training, 615
electron beam induced current (EBIC) imaging, 246–248(F,T)
electron beam probing, 263–268
ATE interface, 267
DUT preparation, 266–267
preparation for, 266
test patterns for, 267
electronics and failure analysis, 128–148
embedded compression, 206–207(F)
designs with, 206
principle of, 206(F)
energy dispersive x-ray analysis, 549–560
EDS overview, 549–551(F,T)
properties of analytical techniques, 549(T)
qualitative analysis, 553–554
death and spatial resolution, 553–554(F,T)
peak identification, 553
quantitative x-ray analysis, 554–555
x-ray detectors, 551–553
EDS, 551–552
micro-calorimetry, 552
wavelength dispersive (WDS), 552
x-ray artifacts, 552
x-ray elemental mapping, 555–560
EDS elemental maps, 555–558(F)
STEM-EDS elemental mapping, 558–559(F)
x-ray map artifacts, 560, 559–560(F)
EOS and ESD failures, 225–238
differentiating between, 225, 232
EOS simulations, 232–236(F)
ESD failures, 227(F)
charged device model (CDM), 230–232(F)
human body model (HBM), 227– 229(F)
machine model (MM), 228–229(F)
junction failure, 225(F)
pin combination, 230(F,T)
externally induced voltage alteration (XIVA), 259, 260(F)

F

failure analysis
actionable vs. random, 17–18
framework, 16(F), 18(F)
process, 1–15. *See also* metaprocess
process flow, 24(F)
wafer level, 49–51(F)
process to determine goals and scope, 23(F)
system level, 16–22
integration, 16(F)
failure mechanisms
electrochemical environment, 25, 31
corrosion, 30(F), 31–32
electrochemical migration, 30(F), 32
mechanical loading, 25(F)
die fracture, 25– 26(F)
interposer level failure,25(F), 26
PWB related failure, 29–31
solder joint fracture, 25–29(F)
thermal loading, 25
failure selection
emerging issue, 17(F)

failure selection *(continued)*
excursion, 17(F)
improvement effort, 18(F)
fast fault isolation, CAD navigation, 354–361
fault isolation, 18–19
impact on cost, 19(F)
fault localization, with voltage contrast, 269–278
fault models, 192(F)
FBGA package, 40, 42(F)
flipchip packaged device, 349(F)
fluorescent microthermal imaging, 317–328
application of, 326–328(F)
EuTTA compound, 317–320(F)
image processing, 320–322(F)
photon shot noise and signal averaging, 323–325(F)
system hardware, 322–324(F)
ultraviolet film bleaching, 324–326(F)
focused ion beam (FIB) systems, 583–593
applications, 586–592
analytical/failure analysis, 591–592(F)
capabilities, 586–587(F)
configurations, 586(F)
electrical, 587–590(F)
electrical probe assistance, 590–591(F)
voltage contrast imaging, 591(F)
collimation and focusing, 584(F)
contrast, imaging, and resolution, 584–586(F)
ion beam generation, 583–584(F)
raster and imaging, 584

G

gate defects, 213
gate short circuits, 213–215(F)
gate oxide breakdown, 144
gate-source overvoltage, 213(F)
GIDEP, 173

I

IC optical micrograph, schematic, 354(F)
inductive proof, 3
infrared (IR) thermography, 312–313(F)
integrated circuit testing methods, 190–198
IR lock-in thermography, 330–339(F)

L

laser voltage probing, 349–353(F)
light induced voltage alteration (LIVA), 254–257(F)
linear region, 154
liquid crystals, 313–315(F)
for failure analysis, 316–317(F)
nematic, optical properties of, 315–316(F)
Lissajous figures, 209–212

M

magnetic current imaging, 301–303(F), 305–309(F)
magnetoresistive sensors, 303–304(F)
management principles and practices, 617–626
financial management, 623–624(F)
lab design and operations, 620–622
lab design, 620–621
lab operations, 621–622
lab management, 617–619
outline of, 617(F)
recruitment, 617–618(F)
retention, 618–619(F)
skills mix, 619
staffing, 617
training, 618(F)
lab organization, 619–620
management philosophy, 619–620
reporting structure, 619
specialization versus generalization, 620
world-wide operations, 620
metrics and measurements, 624–626(F)
strategic development, 622(F)
MCP package, 41, 43(F)
memory failure. *See also* signature analysis
metal mousebites, 143(F)
metal slivers, 142–143(F)
metaprocess, 1–15
analysis flow, 3–4, 11–14(F)
application of, 3
definition of, 2
loop for failure analysis, 1(F)
microelectromechanical systems (MEMS), 52–77
bulk micromachining, 52–53(F)
classification, 57–58
failure modes/mechanisms, 58–63
electrostatic discharge, 61–63(F)
fatigue, 59(F)
fracture, 59(F)
oxide charging, 63(F)
particles/obstruction, 58(F)
stiction, 61(F)
wear, 60–61(F)
integrated, 56–57(F)
LIGA, 55–56(F)
polymer, 57
surface micromachining, 53–55(F)
tools and techniques
0-level packaging hermeticity and scanning
acoustic microscopy (SAM), 72–74(F)
3-D motion analysis, 68–70(F)
atomic force microscopy (AFM), 73–74(F)
focused ion beam (FIB), 65–67(F)
IR confocal laser microscopy, 68–69(F)

photoemission microscopy (PEM), 71–73(F)
Raman spectroscopy, 69–72(F)
scanning electron microscopy (SEM), 63–64(F)
thermally-induced voltage alteration (TIVA), 67–68(F)
transmission electron microscopy (TEM), 64–65(F)
MOS
capacitor, 152–153(F)
transistor, 149(F), 162(F)
characterization, 163(F)
long channel, 153–154(F)
saturation region, 162
subthreshold region, 163(F)
triode region, 162, 163(F)
MOSFET
analysis of a complete circuit, 131–132(F)
operation and three bias states, 128–130(F,T)
short channel transistors, 132–133(F,T)
terminal characteristics, 130– 131(F)

N

non-ideal behavior, 155–157
leakage currents, 155–156
gate insulator tunneling currents, 156(F)
short channel effect, 155–156(F)
subthreshold current, 155(F)
reducing drive current, 156–157
mobility reduction and velocity saturation, 156
poly depletion and channel quantization, 157
source resistance, 156

O

OBIRCH, TIVA, and SEI, 256, 258–260(F)
op-amps, 167–168(F)
input bias current (IIB), 168, 169(F)
input offset voltage (VOS), 168–170(F)
optical beam induced current (OBIC), 254(F)
optical microscopy, 457–476(F)
aberrations, 465–466
chromatic, 465–466(F)
lens, 465
spherical, 466(F)
binocular eyepieces, 464–465
brightfield/darkfield, 468(F)
coma, astigmatism, distortion, 466
doped silicon backside thinning, 470–471(F)
field and aperture planes, 463–464(F)
field flatness, 465(F)
immersion, solid immersion lenses (SILs), 466–468(F)
index of refraction, 457
infrared microscopy, 469–470(F)
Kohler illumination, 464(F)
laser signal injection microscopy, 474–475(F)
magnification
downside to excess, 463
on resolution, role of, 462–463(F)
pixel pitch, 463(F)
microscope column, 458(F)
microscope objectives, 459–462(F)
numerical aperture, 460–462(F)
photoemission high NA macro lenses, 472(F)
photoemission infrared microscopy, 471(F)
polarized light microscopy, 469(F)
silicon transmission, 470
stereomicroscopes, 459 (F)
thermal infrared hot spot detection, 472–474(F)
UV microscopy, 469
optoelectronic devices, 78–98
characterization techniques, 87–89(F)
detector failure mechanisms, 85
determining cause of failure, 85–86
failure library, 85–86
failure tree, 86
product return history, 86
failure from overstress (EOS or ESD), 84–85(F)
LED failure mechanisms, 85
maverick failure mechanisms
epitaxial defects, 83–84(F)
maverick and wearout mechanisms, difference between, 81–82(F)
mechanical damage, 84(F)
optoelectronic FA tools, 86–87(F)
passive optoelectronics failure mechanisms, 85
physics of failure
climb dislocations, 80(F)
dark spot defects (DSDs), 81
glide dislocations, 80–81(F)
gradual degradation, 81(F)
product reliability, common steps to assure, 93–96(F)
scanning electron microscopy techniques, 89–91(F)
scanning laser techniques, 92–93(F)
screening, 96
semiconductor lasers
active regions, types of, 78
basics of, 79–80(F)
transmission electron microscopy, 91–92(F)
types of, 78
wearout failure mechanisms, 82
epitaxial contamination, 82
facet damage, 82
ionic contamination, 82
laser failure, 82
normal, 83(F)

P

package failure, 34–39
definition of, 34
FA flow, 34–39
open and high resistance failures, 34–35(F), 38(F)
shorts and leakage failures, 35–37(F), 39(F)
package fault isolation using TDR, 383–396
parametric failures, 139
extrinsic, 142
how to detect, 144
intrinsic, 140–142(F)
passive components, 111–120
aluminum electrolytic capacitor, 114–115(F), 118–119(F)
fluid-filled metallized-film capacitor, 112–114(F), 117–118(F)
surface-mount stacked-film capacitor, 112(F)
surface-mount thick-film resistor, 111(F), 115–117(F)
switch contact block, 120(F)
wound components, 114, 119(F)
passive voltage contrast (PVC), 276–278(F)
in FIB, 269–272(F,T)
in SEM versus FIB, 272–273(F,T)
phase interpolation, 183(F)
photon emission (PEM), 279–291(F,T)
emission sources, 284–288(F,T)
picosecond imaging circuit analysis (PICA), 292–300
output and optimization, 296–297(F)
systems and instrumentation, 294–296(F)
pn junction, 151–152(F)
diode, 159(F)
diode characterization, 159–161(F)
polymer grass, 403(F)
PoP stacking, 41(F)
precision decapsulation, 43–44(F)
product integration, 16(F)
product life cycle, shrinking, 17(F)

R

reliability and quality basics, 121–127
quality vs. reliability, 121
reliability engineering, 123
resistive contrast imaging (RCI), 248–249(F)
resistive vias, 142–143
RIE grass, 402(F)
RoHS, 171
root cause analysis, 19

S

saturation region, 154
scaling, 155(F)
scan analysis, high-volume, 218–224
correlating failure modes, 221–223(F)
scan pattern cleanup, 220–221(F)
volume data analysis, 219–220(F)
scan chain test, 201(T)
scan design, 193(F)
scan diagnosis, 199–208(T)
at-speed failures, 203
bridges, 203
cell internal defect or interconnect defect, 204
defects in scan chains, 204
opens, 202
tools, 202
scanning electron microscopy, 477–496
beam damage, 484
beam-sample interaction, 481–482(F)
beam voltage, 486–487(T,F)
brightness, 450
and contrast, 490–491(F)
dynamic focus and image correction, 490
electron optics, 480–481(F)
environmental SEM (ESEM), 495–496(F)
focus and astigmatism correction, 488–490(F)
image distortions, 483(F)
in-lens detection, 479
magnification, 479(F)
raster alignment, 490
sample
charging, 483–484(F)
mounting, 484–485
preparation, 485(F)
tilt and image composition, 487–488(F)
scanning principles, 477–478(F)
scan speed and image quality, 491
SEM cathodes, 480(T)
SEM/optical comparison, 477(T)
sputter coating, 485–486(F)
ultra-high resolution, 497–505
forward scattered electron imaging, 501–504(F)
resolution limits, 497–498(F,T)
scanning probe microscopy. *See also* atomic force microscopy, 536–548
scanning transmission electron microscopy (STEM), 515–528
applications of, 516
high resolution STEM, 516, 517(F)
interfacial defect, 516, 517(F)
challenges, 517
elemental analysis, 518
EELS vs. EDS, 520
electron energy loss spectroscopy (EELS), 519–520(F)
energy dispersive spectroscopy (EDS), 518–519(F)
elemental mapping, 520
energy-filtered TEM (EFTEM), 521
low-loss imaging, 522–523(F)
STEM EDS, 520–521(F)
STEM EELS, 521
with EFTEM, 523–524(F)

zero-loss imaging, 521–522(F)
illustration of, 515(F)
planar STEM-BF, 516(F)
off-axis electron holography, 525–526(F)
STEM-ADF, 516(F)
STEM-HAADF, 516(F)
STEM-in-SEM, 491–496(F,T), 518
STEM-in-TEM, 498–501(F)
scan tests, failure bucketing of, 201(F)
secondary ion mass spectrometry (SIMS), 573–582
applications, 576–582
III-IV materials, 579(F)
dielectric materials, 579–580(F)
failure analysis experimental design, 580–581
gate oxide breakdown, 580(F)
oxynitrile, 577(F)
silicon germanium (SiGe), 578–579(F)
surface and interface contamination, 581–582(F)
ultra-shallow implants, 577–578(F)
depth profiles, 574(F)
dynamic and static, 574–576(F)
secondary ion generation, 573–574(F)
ultra-thin film analysis, 576
SIMS depth resolution, 576(F)
vacuum compatibility, 574
signature analysis, 239–245
design for test practices, 239–240
physical failure analysis strategies, 244–245
post-test data analysis tools, 241–243(F)
root cause theorization, 243–244
test floor data collection, 240–241
with I_{DDQ} versus V_{DD}, 144–146(F)
silicon FETs, hot carriers in, 292–293
single transistor circuits, 163
common collector (CC) amp, 164
common drain (CD) amp, 165(F)
common emitter (CE) amp, 163–164(F,T)
common source (CS) amp, 165(F)
soft defect localization, 259–260(F)
solar photovoltaic module, 99–103
common failure for thin film PV modules, 99(T)
common failure mode for crystalline Si PV modules, 99(T)
electrical tests for failure, 99–100(F)
emission imaging, 101–102(F)
thermal imaging of modules, 100(F)
forward bias thermal imaging, 100(F)
lock-in thermal imaging, 101(F)
reverse bias thermal imaging and shunts, 101(F)
thermal reflectance imaging, 102–103(F)
solid immersion lenses (SILs), 260–261(F)
SQUID, 303(F)
standardization
delayering standard, considerations for, 398(F)
list of reasons for, 398(F)
“Standard Sam,” 397–398(F)
statistical distributions, 123–125
exponential distribution, 124(F)
lognormal distribution, 124(F)
normal distribution, 124(F)
Weibull distribution, 124–125(F)
statistical process control, 121(F)
submicron defects, analysis of. *See also* Auger electron spectroscopy, 561–572

T

Taguchi methodology, 123
terms and definitions, 635–650
test data compression architecture, 196(F)
testing
analog circuit, 195
functional and structural
definitions, 192
trade-offs, 193
memory, 194(F)
embedded, 194
reliability, 195
scan-based delay, 193
solving emerging problems, 196
defect-based test and statistical testing, 196
test data volume and embedded compression, 196
test power, 196
system-on-a-chip (SOC), 195
test flow, 190(F)
trends, 190
types of, 191–192
characterization, 192
contact power-up, 191
I_{DDQ}, 191(F)
I/O parametric, 191
I/O performance, 192
logical, 192
speed and power binning, 192
thermal defect detection, 310–329
thermal failure analysis, IR lock-in thermography, 330–339
thermal laser stimulation, 340–348
case studies, 345–347(F)
models, 342–345(F,T)
techniques, 339–342(F,T)
threshold voltage, 153(F)
time domain reflectometry (TDR), 383–396
comparative analysis, 390–391(F)
fundamentals, 383–385(F,T)
goals and methods, 388–390(F)
impedance profiles, 392–393(F)
measurements of “splits” and “stubs,” 387(F)
multiple reflection effects, 385–386(F)
plane-to-plane short, 394–395(F,T)
probing and fixturing, 387–388(F)
resolution and rise time, 386(F)

signal-to-ground short, 393–394(F)
  signature analysis, 390(F)
top die removal, 44–45(F)
transistor evolution, 157(F)
transmission electron microscopy (TEM), 506–528
  illustration of, 506
  imaging, 510
    applications of diffraction contrast, 511–512(F)
    blocked contacts, 513(F)
    defects in silicon substrate, 512–513(F)
    diffraction in silicon devices, 510–511(F)
    gate oxide breakdown, 513(F)
    parallel beam illumination, 510
    phase contrast, 514–515(F)
    stringers, 513(F)
    thickness-mass contrast, 514
    zone-axis imaging, 511(F)
  sample, 507–509
    cross section vs. planar, 507–508(F)
    FIB-induced damage, 509
    sample preparation procedures, 508–509(F)
    thickness of, 507(F)

V

voltage references, 166(F)
voltage regulators, 167(F)

W

wafer level failure analysis, 49–51(F)

X

x-ray imaging tools, 529–535
  attenuation length, 529(F)
  packaging failure analysis, 530–531(F)
  size scales of features, 530(F)
  transmission x-ray microscope (TXM) application, 532–534(F)

Y

yield enhancement loop, 49–50(F)