

SPONSORED BY THE
OPTICAL SOCIETY OF AMERICA

HANDBOOK OF OPTICS

FUNDAMENTALS, TECHNIQUES, & DESIGN

• SECOND EDITION •

VOLUME

I

MICHAEL BASS, EDITOR IN CHIEF
ERIC W. VAN STRYLAND • DAVID R. WILLIAMS • WILLIAM L. WOLFE, ASSOCIATE EDITORS

HANDBOOK OF OPTICS

Other McGraw-Hill Books of Interest

Hecht • THE LASER GUIDEBOOK

Manning • STOCHASTIC ELECTROMAGNETIC IMAGE PROPAGATION

Nishihara, Haruna, Suhara • OPTICAL INTEGRATED CIRCUITS

Rancourt • OPTICAL THIN FILMS USERS' HANDBOOK

Sibley • OPTICAL COMMUNICATIONS

Smith • MODERN OPTICAL ENGINEERING

Smith • MODERN LENS DESIGN

Waynant, Ediger • ELECTRO-OPTICS HANDBOOK

Wyatt • ELECTRO-OPTICAL SYSTEM DESIGN

To order, or to receive additional information on these or any other McGraw-Hill titles, please call 1-800-822-8158 in the United States. In other countries, please contact your local McGraw-Hill Office.

BC14BCZ

HANDBOOK OF OPTICS

Volume I
Fundamentals, Techniques,
and Design

Second Edition

Sponsored by the
OPTICAL SOCIETY OF AMERICA

Michael Bass Editor in Chief

*The Center for Research and
Education in Optics and Lasers (CREOL)
University of Central Florida
Orlando, Florida*

Eric W. Van Stryland Associate Editor

*The Center for Research and Education
in Optics and Lasers (CREOL)
University of Central Florida
Orlando, Florida*

David R. Williams Associate Editor

*Center for Visual Science
University of Rochester
Rochester, New York*

William L. Wolfe Associate Editor

*Optical Sciences Center
University of Arizona
Tucson, Arizona*

McGRAW-HILL, INC.

New York San Francisco Washington, D.C. Auckland Bogotá
Caracas Lisbon London Madrid Mexico City Milan
Montreal New Delhi San Juan Singapore
Sydney Tokyo Toronto

Library of Congress Cataloging-in-Publication Data

Handbook of optics / sponsored by the Optical Society of America ;
Michael Bass, editor in chief. — 2nd ed.

p. cm.

Includes bibliographical references and index.

Contents: 1. Fundamentals, techniques, and design — 2. Devices,
measurement, and properties.

ISBN 0-07-047740-X

1. Optics—Handbooks, manuals, etc. 2. Optical instruments—
Handbooks, manuals, etc. I. Bass, Michael. II. Optical Society
of America.

QC369.H35 1995

535—dc20

94-19339

CIP

Copyright © 1995 by McGraw-Hill, Inc. All rights reserved. Printed in the
United States of America. Except as permitted under the United States
Copyright Act of 1976, no part of this publication may be reproduced or
distributed in any form or by any means, or stored in a data base or
retrieval system, without the prior written permission of the publisher.

6 7 8 9 DOC/DOC 0 9 8 7 6 5 4 3 2 1

ISBN 0-07-047740-X

The sponsoring editor for this book was Stephen S. Chapman, the editing
supervisor was Peggy Lamb, and the production supervisor was Pamela A.
Pelton. It was set in Times Roman by The Universities Press (Belfast) Ltd.

Printed and bound by R.R. Donnelly & Sons Company.

This book was printed on acid-free paper.

Information contained in this work has been obtained by
McGraw-Hill, Inc. from sources believed to be reliable. How-
ever, neither McGraw-Hill nor its authors guarantees the
accuracy or completeness of any information published herein
and neither McGraw-Hill nor its authors shall be responsible for
any errors, omissions, or damages arising out of use of this
information. This work is published with the understanding that
McGraw-Hill and its authors are supplying information but are
not attempting to render engineering or other professional
services. If such services are required, the assistance of an
appropriate professional should be sought.

CONTENTS

Contributors	xvii
Preface	xix
Glossary and Fundamental Constants	xxi

Part 1. Geometric Optics 1.1

Chapter 1. General Principles of Geometric Optics *Douglas S. Goodman* 1.3

1.1. Glossary	/ 1.3
1.2. Introduction	/ 1.7
1.3. Fundamentals	/ 1.9
1.4. Characteristic Functions	/ 1.15
1.5. Rays in Heterogeneous Media	/ 1.20
1.6. Conservation of Etendue	/ 1.24
1.7. Skew Invariant	/ 1.25
1.8. Refraction and Reflection at Interfaces Between Homogeneous Media	/ 1.26
1.9. Imaging	/ 1.29
1.10. Description of Systems of Revolution	/ 1.35
1.11. Tracing Rays in Centered Systems of Spherical Surfaces	/ 1.39
1.12. Paraxial Optics of Systems of Revolution	/ 1.41
1.13. Images About Known Rays	/ 1.46
1.14. Gaussian Lens Properties	/ 1.48
1.15. Collineation	/ 1.60
1.16. System Combination—Gaussian Properties	/ 1.68
1.17. Paraxial Matrix Methods	/ 1.70
1.18. Apertures, Pupils, Stops, Fields, and Related Matters	/ 1.80
1.18. Geometric Aberrations of Point Images—Description	/ 1.82
1.20. References	/ 1.100

Part 2. Physical Optics 2.1

Chapter 2. Interference *John E. Greivenkamp, Jr.* 2.3

2.1. Glossary	/ 2.3
2.2. Introduction	/ 2.3
2.3. Waves and Wavefronts	/ 2.3
2.4. Interference	/ 2.5
2.5. Interference by Wavefront Division	/ 2.14
2.6. Interference by Amplitude Division	/ 2.19
2.7. Multiple Beam Interference	/ 2.29
2.8. Coherence and Interference	/ 2.36
2.9. References	/ 2.43

Chapter 3. Diffraction *A. S. Marathay* **3.1**

- 3.1. Glossary / 3.1
- 3.2. Introduction / 3.1
- 3.3. Light Waves / 3.2
- 3.4. Huygens-Fresnel Construction / 3.4
- 3.5. Cylindrical Wavefront / 3.13
- 3.6. Mathematical Theory of Diffraction / 3.19
- 3.7. Vector Diffraction / 3.27
- 3.8. References / 3.30

Chapter 4. Coherence Theory *William H. Carter* **4.1**

- 4.1. Glossary / 4.1
- 4.2. Introduction / 4.1
- 4.3. Some Elementary Classical Concepts / 4.2
- 4.4. Definitions of Coherence Functions / 4.4
- 4.5. Model Sources / 4.9
- 4.6. Propagation / 4.13
- 4.7. Spectrum of Light / 4.20
- 4.8. Polarization Effect / 4.23
- 4.9. Applications / 4.23
- 4.10. References / 4.25

Chapter 5. Polarization *Jean M. Bennett* **5.1**

- 5.1. Glossary / 5.1
- 5.1. Basic Concepts and Conventions / 5.2
- 5.2. Fresnel Equations / 5.4
- 5.3. Basic Relations for Polarizers / 5.12
- 5.4. Polarization by Nonnormal-Incidence Reflection (Pile of Plates) / 5.13
- 5.5. Polarization by Nonnormal-Incidence Transmission (Pile of Plates) / 5.16
- 5.6. Quarter-Wave Plates and Other Phase Retardation Plates / 5.22
- 5.7. Matrix Methods for Computing Polarization / 5.25
- 5.8. References / 5.28

Chapter 6. Scattering by Particles *Craig F. Bohren* **6.1**

- 6.1. Glossary / 6.1
- 6.2. Introduction / 6.2
- 6.3. Scattering: An Overview / 6.3
- 6.4. Scattering by Particles: Basic Concepts and Terminology / 6.5
- 6.5. Scattering by an Isotropic, Homogeneous Sphere: the Archetype / 6.12
- 6.6. Scattering by Regular Particles / 6.15
- 6.7. Computational Methods for Nonspherical Particles / 6.17
- 6.8. References / 6.18

Chapter 7. Surface Scattering *E. L. Church and P. Z. Takacs* **7.1**

- 7.1. Glossary / 7.1
- 7.2. Introduction / 7.1
- 7.3. Notation / 7.2
- 7.4. Scattering Theory / 7.3
- 7.5. Surface Models / 7.5

- 7.6. Wavelength Scaling / 7.7
- 7.7. Profile Measurements / 7.8
- 7.8. Finish Specification / 7.11
- 7.9. References / 7.12

Part 3. Quantum Optics 8.1

Chapter 8. Optical Spectroscopy and Spectroscopic Lineshapes *Brian Henderson* 8.3

- 8.1. Glossary / 8.3
- 8.2. Introductory Comments / 8.4
- 8.3. Theoretical Preliminaries / 8.5
- 8.4. Rates of Spectroscopic Transitions / 8.6
- 8.5. Lineshapes of Spectral Transitions / 8.8
- 8.6. Spectroscopy of 1-Electron Atoms / 8.10
- 8.7. Multielectron Atoms / 8.12
- 8.8. Optical Spectra and the Outer Electronic Structure / 8.14
- 8.9. Spectra of Tri-Positive Rare Earth Atoms / 8.15
- 8.10. Vibrational and Rotational Effects of Molecules / 8.21
- 8.11. Lineshapes in Solid State Spectroscopy / 8.25
- 8.12. References / 8.30

Chapter 9. Fundamental Optical Properties of Solids *Alan Miller* 9.1

- 9.1. Glossary / 9.1
- 9.2. Introduction / 9.4
- 9.3. Propagation of Light in Solids / 9.4
- 9.4. Dispersion Relations / 9.13
- 9.5. Lattice Interactions / 9.16
- 9.6. Free Electron Properties / 9.19
- 9.7. Band Structures and Interband Transitions / 9.24
- 9.8. References / 9.33

Part 4. Optical Sources 10.1

Chapter 10. Artificial Sources *Anthony LaRocca* 10.3

- 10.1. Glossary / 10.3
- 10.2. Introduction / 10.3
- 10.3. Laboratory Sources / 10.4
- 10.4. Commercial Sources / 10.11
- 10.5. References / 10.49

Chapter 11. Lasers *William T. Silfvast* 11.1

- 11.1. Glossary / 11.1
- 11.2. Introduction / 11.2
- 11.3. Laser Properties Associated with the Laser Gain Medium / 11.4

- 11.4. Laser Properties Associated with Optical Cavities or Resonators / 11.20
- 11.5. Special Laser Cavities / 11.27
- 11.6. Specific Types of Lasers / 11.32
- 11.7. References / 11.39

Chapter 12. Light-Emitting Diodes *Roland H. Haitz, M. George Craford, and Robert H. Weissman*

12.1

- 12.1. Glossary / 12.1
- 12.2. Introduction / 12.2
- 12.3. Light-Generation Processes / 12.2
- 12.4. Light Extraction / 12.7
- 12.5. Device Structures / 12.8
- 12.6. Materials Systems / 12.15
- 12.7. Substrate Technology / 12.21
- 12.8. Epitaxial Technology / 12.23
- 12.9. Wafer Processing / 12.24
- 12.10. LED Quality and Reliability / 12.27
- 12.11. LED Based Products / 12.31
- 12.12. References / 12.38

Chapter 13. Semiconductor Lasers *Pamela L. Derry, Luis Figueroa, and Chi-Shain Hong*

13.1

- 13.1. Glossary / 13.1
- 13.2. Introduction / 13.3
- 13.3. Applications for Semiconductor Lasers / 13.3
- 13.4. Basic Operation / 13.4
- 13.5. Fabrication and Configurations / 13.7
- 13.6. Quantum Well Lasers / 13.10
- 13.7. High-Power Semiconductor Lasers / 13.19
- 13.8. High-Speed Modulation / 13.32
- 13.9. Spectral Properties / 13.39
- 13.10. Surface-Emitting Lasers / 13.42
- 13.11. Conclusion / 13.46
- 13.12. References / 13.47

Chapter 14. Ultrashort Laser Sources *Xin Miao Zhao and Jean-Claude Diels*

14.1

- 14.1. Glossary / 14.1
- 14.2. Introduction / 14.2
- 14.3. Passively Mode-Locked Lasers / 14.2
- 14.4. Synchronous, Hybrid, and Double Mode Locking / 14.7
- 14.5. Active and Passive Negative Feedback / 14.11
- 14.6. Nonlinear Optical Sources / 14.12
- 14.7. Additive and Self-Mode-Locking / 14.14
- 14.8. Other Ultrashort Pulse Sources / 14.18
- 14.9. Amplification / 14.21
- 14.10. Diagnostic Techniques / 14.22
- 14.11. References / 14.25

Part 5. Optical Detectors **15.1**

Chapter 15. Photodetectors *Paul R. Norton* **15.3**

- 15.1. Scope / 15.3
- 15.2. Thermal Detectors / 15.4
- 15.3. Quantum Detectors / 15.5
- 15.4. Definitions / 15.8
- 15.5. Detector Performance and Sensitivity / 15.11
- 15.6. Other Performance Parameters / 15.15
- 15.7. Detector Performance / 15.19
- 15.8. References / 15.100

Chapter 16. Photodetection *Abhay M. Joshi and Gregory H. Olsen* **16.1**

- 16.1. Glossary / 16.1
- 16.2. Introduction / 16.2
- 16.3. Principles of Operation / 16.3
- 16.4. Applications / 16.12
- 16.5. Reliability / 16.13
- 16.6. Future Photodetectors / 16.16
- 16.7. Acknowledgment / 16.19
- 16.8. References / 16.19

Chapter 17. High-Speed Photodetectors *J. E. Bowers and Y. G. Wey* **17.1**

- 17.1. Glossary / 17.1
- 17.2. Introduction / 17.3
- 17.3. Photodetector Structures / 17.3
- 17.4. Speed Limitations / 17.6
- 17.5. PIN Photodetectors / 17.11
- 17.6. Schottky Photodiode / 17.17
- 17.7. Avalanche Photodetectors / 17.19
- 17.8. Photoconductors / 17.22
- 17.9. Summary / 17.25
- 17.10. References / 17.26

Chapter 18. Signal Detection and Analysis *John R. Willison* **18.1**

- 18.1. Glossary / 18.1
- 18.2. Introduction / 18.1
- 18.3. Prototype Experiment / 18.2
- 18.4. Noise Sources / 18.3
- 18.5. Applications Using Photomultipliers / 18.7
- 18.6. Amplifiers / 18.11
- 18.7. Signal Analysis / 18.13
- 18.8. References / 18.16

Chapter 19. Thermal Detectors *William L. Wolfe and Paul W. Kruse* **19.1**

- 19.1. Glossary / 19.1
- 19.2. Thermal Detector Elements / 19.1
- 19.3. Arrays / 19.8
- 19.4. References / 19.13

Part 6. Imaging Detectors 20.1

Chapter 20. Photographic Films *Joseph H. Altman* 20.3

- 20.1. Glossary / 20.3
- 20.2. Structure of Silver Halide Photographic Layers / 20.4
- 20.3. Grains / 20.5
- 20.4. Processing / 20.5
- 20.5. Exposure / 20.6
- 20.6. Optical Density / 20.6
- 20.7. D-Log H Curve / 20.9
- 20.8. Spectral Sensitivity / 20.11
- 20.9. Reciprocity Failure / 20.12
- 20.10. Development Effects / 20.13
- 20.11. Color Photography / 20.14
- 20.12. Microdensitometers / 20.16
- 20.13. Performance of Photographic Systems / 20.17
- 20.14. Image Structure / 20.18
- 20.15. Acutance / 20.19
- 20.16. Graininess / 20.21
- 20.17. Sharpness and Graininess Considered Together / 20.24
- 20.18. Signal to Noise Ratio and Detective Quantum Efficiency / 20.24
- 20.19. Resolving Power / 20.26
- 20.20. Information Capacity / 20.26
- 20.21. List of Photographic Manufacturers / 20.27
- 20.22. References / 20.27

Chapter 21. Image Tube Intensified Electronic Imaging *C. B. Johnson and L. D. Owen* 21.1

- 21.1. Glossary / 21.1
- 21.2. Introduction / 21.2
- 21.3. Optical Interface / 21.3
- 21.4. Image Intensifiers / 21.7
- 21.5. Image Intensified Self-Scanned Arrays / 21.20
- 21.6. Applications / 21.29
- 21.7. References / 21.31

Chapter 22. Visible Array Detectors *Timothy J. Tredwell* 22.1

- 22.1. Glossary / 22.1
- 22.2. Introduction / 22.2
- 22.3. Image Sensing Elements / 22.2
- 22.4. Readout Elements / 22.13
- 22.5. Sensor Architectures / 22.22
- 22.6. References / 22.37

Chapter 23. Infrared Detector Arrays *Lester J. Kozlowski and Walter F. Kosonocky* 23.1

- 23.1. Glossary / 23.1
- 23.2. Introduction / 23.4
- 23.3. Monolithic FPAs / 23.10
- 23.4. Hybrid FPAs / 23.15

- 23.5. Performance: Figures of Merit / 23.25
- 23.6. Current Status and Future Trends / 23.30
- 23.7. References / 23.25

Part 7. Vision **24.1**

Chapter 24. Optics of the Eye *W. N. Charman* **24.3**

- 24.1. Glossary / 24.3
- 24.2. Introduction / 24.5
- 24.3. Eye Models / 24.7
- 24.4. Ocular Transmittance and Retinal Illuminance / 24.9
- 24.5. Factors Affecting Retinal Image Quality / 24.13
- 24.6. Final Retinal Image Quality / 24.19
- 24.7. Depth-of-Focus and Accommodation / 24.26
- 24.8. Movements of the Eyes / 24.34
- 24.9. Two Eyes and Steropsis / 24.37
- 24.10. Conclusion / 24.40
- 24.11. References / 24.40

Chapter 25. Visual Performance *Wilson S. Geisler and Martin S. Banks* **25.1**

- 25.1. Glossary / 25.1
- 25.2. Introduction / 25.2
- 25.3. Optics, Anatomy, Physiology of the Visual System / 25.3
- 25.4. Visual Performance / 25.15
- 25.5. References / 25.44

Chapter 26 Colorimetry *David H. Brainard* **26.1**

- 26.1. Glossary / 26.1
- 26.2. Introduction / 26.1
- 26.3. Fundamentals / 26.3
- 26.4. Topics / 26.25
- 26.5. Appendix A. Matrix Algebra / 26.44
- 26.6. Acknowledgments / 26.48
- 26.7. References / 26.48

Chapter 27. Displays for Vision Research *William Cowan* **27.1**

- 27.1. Glossary / 27.1
- 27.2. Introduction / 27.3
- 27.3. Operational Characteristics of Color Monitors / 27.3
- 27.4. Colorimetric Calibration of Video Monitors / 27.21
- 27.5. An Introduction to Liquid Crystal Displays / 27.36
- 27.6. Acknowledgments / 27.43
- 27.7. References / 27.43

Chapter 28. Optical Generation of the Visual Stimulus *Stephen A. Burns and Robert H. Webb* **28.1**

- 28.1. Glossary / 28.1
- 28.2. Introduction / 28.1

28.3.	The Size of the Visual Stimulus /	28.1
28.4.	Free or Newtonian Viewing /	28.2
28.5.	Maxwellian Viewing /	28.4
28.6.	Building an Optical System /	28.8
28.7.	Light Exposure and Ocular Safety /	28.19
28.8.	Light Sources /	28.20
28.9.	Coherent Radiation /	28.20
28.10.	Detectors /	28.22
28.11.	Putting It Together /	28.23
28.12.	Conclusions /	28.27
28.13.	Acknowledgments /	28.27
28.14.	General References /	28.27
28.15.	References /	28.27

Chapter 29. Psychophysical Methods	<i>Denis G. Pelli and Bart Farell</i>	29.1
---	---------------------------------------	-------------

29.1.	Introduction /	29.1
29.2.	Definitions /	29.2
29.3.	Visual Stimuli /	29.4
29.4.	Adjustments /	29.4
29.5.	Judgments /	29.6
29.6.	Stimulus Sequencing /	29.10
29.7.	Conclusion /	29.10
29.8.	Tips from the Pros /	29.11
29.9.	Acknowledgments /	29.11
29.10.	References /	29.12

Part 8. Optical Information and Image Processing	30.1
---	-------------

Chapter 30. Analog Optical Signal and Image Processing	
<i>Joseph W. Goodman</i>	30.3

30.1.	Glossary /	30.3
30.2.	Introduction /	30.3
30.3.	Fundamental Analog Operations /	30.4
30.4.	Analog Optical Fourier Transforms /	30.5
30.5.	Spatial Filtering /	30.8
30.6.	Coherent Optical Processing of Synthetic Aperture Radar Data /	30.8
30.7.	Coherent Optical Processing of Temporal Signals /	30.10
30.8.	Optical Processing of Two-Dimensional Images /	30.14
30.9.	Incoherent Processing of Discrete Signals /	30.19
30.10.	Concluding Remarks /	30.22
30.11.	References /	30.23

Chapter 31. Principles of Optical Disk Data Storage	<i>Masud Mansuripur</i>	31.1
--	-------------------------	-------------

31.1.	Introduction /	31.1
31.2.	Preliminaries and Basic Definitions /	31.2
31.3.	The Optical Path /	31.7
31.4.	Automatic Focusing /	31.13
31.5.	Automatic Tracking /	31.15
31.6.	Thermomagnetic Recording Processes /	31.18
31.7.	Magneto-Optical Readout /	31.22
31.8.	Materials of Magneto-Optical Recording /	31.26

- 31.9. Concluding Remarks / 31.29
- 31.10. Further Information / 31.32
- 31.11. References / 31.32

Part 9. Optical Design Techniques **32.1**

Chapter 32. Techniques of First-Order Layout *Warren J. Smith* **32.3**

- 32.1. Glossary / 32.3
- 32.2. First-Order Layout / 32.4
- 32.3. Ray-Tracing / 32.4
- 32.4. Two-Component Systems / 32.5
- 32.5. Afocal Systems / 32.7
- 32.6. Magnifiers and Microscopes / 32.8
- 32.7. Afocal Attachments / 32.8
- 32.8. Field Lenses / 32.8
- 32.9. Condensers / 32.10
- 32.10. Zoom or Varifocal Systems / 32.11
- 32.11. Additional Rays / 32.12
- 32.12. Minimizing Component Power / 32.12
- 32.13. Is It a Reasonable Layout? / 32.13
- 32.14. Achromatism / 32.14
- 32.15. Athermalization / 32.15

Chapter 33. Aberration Curves in Lens Design *Donald C. O'Shea and Michael E. Harrigan* **33.1**

- 33.1. Glossary / 33.1
- 33.2. Introduction / 33.1
- 33.3. Transverse Ray Plots / 33.2
- 33.4. Field Plots / 33.4
- 33.5. Additional Considerations / 33.5
- 33.6. Summary / 33.6
- 33.7. References / 33.6

Chapter 34. Optical Design Software *Douglas C. Sinclair* **34.1**

- 34.1. Glossary / 34.1
- 34.2. Introduction / 34.2
- 34.3. Lens Entry / 34.3
- 34.4. Evaluation / 34.9
- 34.5. Optimization / 34.18
- 34.6. Other Topics / 34.22
- 34.7. Buying Optical Design Software / 34.23
- 34.8. Summary / 34.26
- 34.9. References / 34.26

Chapter 35. Optical Specifications *Robert R. Shannon* **35.1**

- 35.1. Glossary / 35.1
- 35.2. Introduction / 35.1
- 35.3. Preparation of Optical Specifications / 35.4

- 35.4. Image Specifications / 35.5
- 35.5. Element Description / 35.8
- 35.6. Environmental Specifications / 35.10
- 35.7. Presentation of Specifications / 35.10
- 35.8. Problems with Specification Writing / 35.12

Chapter 36. Tolerancing Techniques *Robert R. Shannon* **36.1**

- 36.1. Glossary / 36.1
- 36.2. Introduction / 36.1
- 36.3. Wavefront Tolerances / 36.3
- 36.4. Other Tolerances / 36.8
- 36.5. Starting Points / 36.8
- 36.6. Material Properties / 36.9
- 36.7. Tolerancing Procedures / 36.9
- 36.8. Problems in Tolerancing / 36.11

Chapter 37. Mounting Optical Components *Paul R. Yoder, Jr.* **37.1**

- 37.1. Glossary / 37.1
- 37.2. Introduction and Summary / 37.2
- 37.3. Mounting Individual Lenses / 37.2
- 37.4. Multicomponent Lens Assemblies / 37.14
- 37.5. Mounting Small Mirrors and Prisms / 37.20
- 37.6. References / 37.26

Chapter 38. Control of Stray Light *Robert P. Breault* **38.1**

- 38.1. Glossary / 38.1
- 38.2. Introduction / 38.1
- 38.3. Concepts / 38.2
- 38.4. Stray Light Software / 38.25
- 38.5. Methods / 38.28
- 38.6. Conclusion / 38.31
- 38.7. Sources of Information on Stray Light and Scattered Light / 38.32
- 38.8. References / 38.34

Chapter 39. Thermal Compensation Techniques *P. J. Rogers and M. Roberts* **39.1**

- 39.1. Glossary / 39.1
- 39.2. Introduction / 39.2
- 39.3. Homogeneous Thermal Effects / 39.2
- 39.4. Tolerable Homogeneous Temperature Change (No Compensation) / 39.5
- 39.5. Effect of Thermal Gradients / 39.6
- 39.6. Intrinsic Athermalization / 39.7
- 39.7. Mechanical Thermalization / 39.7
- 39.8. Optical Athermalization / 39.13
- 39.9. References / 39.16

Part 10. Optical Fabrication 40.1

Chapter 40. Optical Fabrication *Robert E. Parks* 40.3

- 40.1. Introduction / 40.3
- 40.2. Basic Steps in Optical Fabrication / 40.3
- 40.3. Plano Optical Surfaces / 40.6
- 40.4. Crystalline Optics / 40.6
- 40.5. Aspherics / 40.6
- 40.6. Diamond Turning / 40.7
- 40.7. Purchasing Optics / 40.7
- 40.8. Conclusions / 40.8
- 40.9. References / 40.8

Chapter 41. Fabrication of Optics by Diamond Turning *Richard L. Rhorer and Chris J. Evans* 41.1

- 41.1. Glossary / 41.1
- 41.2. Introduction / 41.1
- 41.3. The Diamond-Turning Process / 41.2
- 41.4. The Advantages of Diamond Turning / 41.2
- 41.5. Diamond-Turnable Materials / 41.3
- 41.6. Comparison of Diamond Turning and Traditional Optical Fabrication / 41.5
- 41.7. Machine Tools for Diamond Turning / 41.5
- 41.8. Basic Steps in Diamond Turning / 41.7
- 41.9. Surface Finish in Diamond-Turned Optics / 41.8
- 41.10. Measuring Diamond-Turned Surfaces / 41.10
- 41.11. Conclusions / 41.12
- 41.12. References / 41.12

Part 11. Optical Properties of Films and Coatings 42.1

Chapter 42. Optical Properties of Films and Coatings *J. A. Dobrowolski* 42.3

- 42.1. Glossary / 42.3
- 42.2. Introduction / 42.4
- 42.3. Theory and Design of Optical Thin-Film Coatings / 42.9
- 42.4. Thin-Film Manufacturing Considerations / 42.14
- 42.5. Measurements on Optical Coatings / 42.16
- 42.6. Antireflection Coatings / 42.19
- 42.7. Two-Material Periodic Multilayers—Theory / 42.34
- 42.8. Multilayer Reflectors—Experimental Results / 42.41
- 42.9. Cut-off, Heat-Control, and Solar-Cell Cover Filters / 42.54
- 42.10. Beam Splitters and Neutral Filters / 42.61
- 42.11. Interference Polarizers and Polarizing Beam Splitters / 42.68
- 42.12. Bandpass Filters / 42.73
- 42.13. Multilayer for Two or Three Spectral Regions / 42.94
- 42.14. Phase Coatings / 42.96
- 42.15. Interference Filters with Low Reflection / 42.98
- 42.16. Reflection Filters and Coatings / 42.101
- 42.17. Special-Purpose Coatings / 42.107
- 42.18. Acknowledgments / 42.109
- 42.19. References / 42.109

Part 12. Terrestrial Optics 43.1

Chapter 43. Optical Properties of Water *Curtis D. Mobley* 43.3

- 43.1. Introduction / 43.3
- 43.2. Terminology, Notation, and Definitions / 43.3
- 43.3. Radiometric Quantities Useful in Hydrologic Optics / 43.6
- 43.4. Inherent Optical Properties / 43.4
- 43.5. Apparent Optical Properties / 43.12
- 43.6. Optically Significant Constituents of Natural Waters / 43.14
- 43.7. Particle Size Distributions / 43.15
- 43.8. Electromagnetic Properties of Water / 43.17
- 43.9. Index of Refraction / 43.18
- 43.10. Measurement of Absorption / 43.20
- 43.11. Absorption by Pure Sea Water / 43.22
- 43.12. Absorption by Dissolved Organic Matter / 43.23
- 43.13. Absorption by Phytoplankton / 43.24
- 43.14. Absorption by Organic Detritus / 43.26
- 43.15. Bio-Optical Models of Absorption / 43.27
- 43.16. Measurement of Scattering / 43.30
- 43.17. Scattering by Pure Water and by Pure Sea Water / 43.31
- 43.18. Scattering by Particles / 43.33
- 43.19. Wavelength Dependence of Scattering; Bio-Optical Models / 43.35
- 43.20. Beam Attenuation / 43.42
- 43.21. Diffuse Attenuation and Jerlov Water Types / 43.44
- 43.22. Irradiance Reflectance and Remote Sensing / 43.48
- 43.23. Inelastic Scattering and Polarization / 43.51
- 43.24. Acknowledgments / 43.52
- 43.25. References / 43.52

Chapter 44. Atmospheric Optics *Dennis K. Killinger, James H. Churnside, and Laurence S. Rothman* 44.1

- 44.1. Glossary / 44.1
- 44.2. Introduction / 44.2
- 44.3. Physical and Chemical Composition of the Standard Atmosphere / 44.4
- 44.4. Fundamental Theory of Interaction of Light with the Atmosphere / 44.10
- 44.5. Prediction of Atmospheric Optical Transmission: Computer Programs and Databases / 44.21
- 44.6. Atmospheric Optical Turbulence / 44.25
- 44.7. Examples of Atmospheric Optical Remote Sensing / 44.36
- 44.8. Meteorological Optics / 44.39
- 44.9. Acknowledgments / 44.43
- 44.10. References / 44.44

Index follows Chapter 44 1.1

CHAPTER 12

LIGHT-EMITTING DIODES

Roland H. Haitz
M. George Craford
Robert H. Weissman

*Hewlett-Packard Co.,
San Jose, California*

12.1 GLOSSARY

c	velocity of light
E_g	semiconductor energy bandgap
h	Planck's constant
I_T	total LED current
J	LED current density
k	Boltzmann's constant
M	magnification
n_0	low index of refraction medium
n_1	high index of refraction medium
q	electron charge
T	temperature
V	applied voltage
η_i	internal quantum efficiency
θ_c	critical angle
λ	emission wavelength
τ	total minority carrier lifetime
τ_n	nonradiative minority carrier lifetime
τ_r	radiative minority carrier lifetime

12.2 INTRODUCTION

Over the past 25 years the light-emitting diode (LED) has grown from a laboratory curiosity to a broadly used light source for signaling applications. In 1992 LED production reached a level of approximately 25 billion chips, and \$2.5 billion worth of LED-based components were shipped to original equipment manufacturers.

This article covers light-emitting diodes from the basic light-generation processes to

descriptions of LED products. First, we will deal with light-generation mechanisms and light extraction. Four major types of device structures—from simple grown or diffused homojunctions to complex double heterojunction devices are discussed next, followed by a description of the commercially important semiconductors used for LEDs, from the pioneering GaAsP system to the AlGaInP system that is currently revolutionizing LED technology. Then processes used to fabricate LED chips are explained—the growth of GaAs and GaP substrates; the major techniques used for growing the epitaxial material in which the light-generation processes occur; and the steps required to create LED chips up to the point of assembly. Next the important topics of quality and reliability—in particular, chip degradation and package-related failure mechanisms—will be addressed. Finally, LED-based products, such as indicator lamps, numeric and alphanumeric displays, optocouplers, fiber-optic transmitters, and sensors, are described.

This article covers the mainstream structures, materials, processes, and applications in use today. It does not cover certain advanced structures, such as quantum well or strained layer devices, a discussion of which can be found in Chap. 13, "Semiconductor Lasers." The reader is also referred to Chap. 13 for current information on edge-emitting LEDs, whose fabrication and use are similar to lasers.

For further information on the physics of light generation, the reader should consult Refs. 1–11. Semiconductor material systems for LEDs are discussed in Refs. 13–24. Crystal growth, epitaxial, and wafer fabrication processes are discussed in detail in Refs. 25–29.

12.3 LIGHT-GENERATION PROCESSES

When a p - n junction is biased in the forward direction, the resulting current flow across the boundary layer between the p and n regions has two components: holes are injected from the p region into the n region and electrons are injected from the n region into the p region. This so-called minority-carrier injection disturbs the carrier distribution from its equilibrium condition. The injected minority carriers recombine with majority carriers until thermal equilibrium is reestablished. As long as the current continues to flow, minority-carrier injection continues. On both sides of the junction, a new steady-state carrier distribution is established such that the recombination rate equals the injection rate.^{1,2}

Minority-carrier recombination is not instantaneous. The injected minority carriers have to find proper conditions before the recombination process can take place. Both energy and momentum conservation have to be met. Energy conservation can be readily met since a photon can take up the energy of the electron-hole pair, but the photon doesn't contribute much to the conservation of momentum. Therefore, an electron can only combine with a hole of practically identical and opposite momentum. Such proper conditions are not readily met, resulting in a delay. In other words, the injected minority carrier has a finite lifetime τ_r before it combines radiatively through the emission of a photon.² This average time to recombine radiatively through the emission of light can be visualized as the average time it takes an injected minority carrier to find a majority carrier with the right momentum to allow radiative recombination without violating momentum conservation.

Unfortunately, radiative recombination is not the only recombination path. There are also crystalline defects, such as impurities, dislocations, surfaces, etc., that can trap the injected minority carriers. This type of recombination process may or may not generate light. Energy and momentum conservation are met through the successive emission of phonons. Again, the recombination process is not instantaneous because the minority carrier first has to diffuse to a recombination site. This nonradiative recombination process is characterized by a lifetime τ_{nr} .²

Of primary interest in design of light-emitting diodes is the maximization of the

radiative recombination relative to the nonradiative recombination. In other words, it is of interest to develop conditions where radiative recombination occurs fairly rapidly compared with nonradiative recombination. The effectiveness of the light-generation process is described by the fraction of the injected minority carriers that recombine radiatively compared to the total injection. The internal quantum efficiency, η_i can be calculated from τ_r and τ_n . The combined recombination processes lead to a total minority-carrier lifetime τ given by Eq. (1):

$$\frac{1}{\tau} = \frac{1}{\tau_r} + \frac{1}{\tau_n} \quad (1)$$

η_i is simply computed from Eq. (1) as the fraction of carriers recombining radiatively:²

$$\eta_i = \frac{\tau_n}{\tau_r + \tau_n} \quad (2)$$

Of interest are two simple cases: in the case of excellent material quality (large τ_n) or efficient radiative recombination conditions (small τ_r), the internal quantum efficiency approaches 100 percent. For the opposite case ($\tau_n \ll \tau_r$), we find $\eta_i \approx \tau_n/\tau_r \ll 1$. As discussed under "Material Systems," there are several families of III-V compounds with internal quantum efficiencies approaching 100 percent. There are also other useful semiconductor materials with internal quantum efficiencies in the 1 to 10 percent range.

To find material systems for LEDs with a high quantum efficiency, one has to understand the band structure of semiconductors. The band structure describes the allowed distribution of energy and momentum states for electrons and holes (see Fig. 1

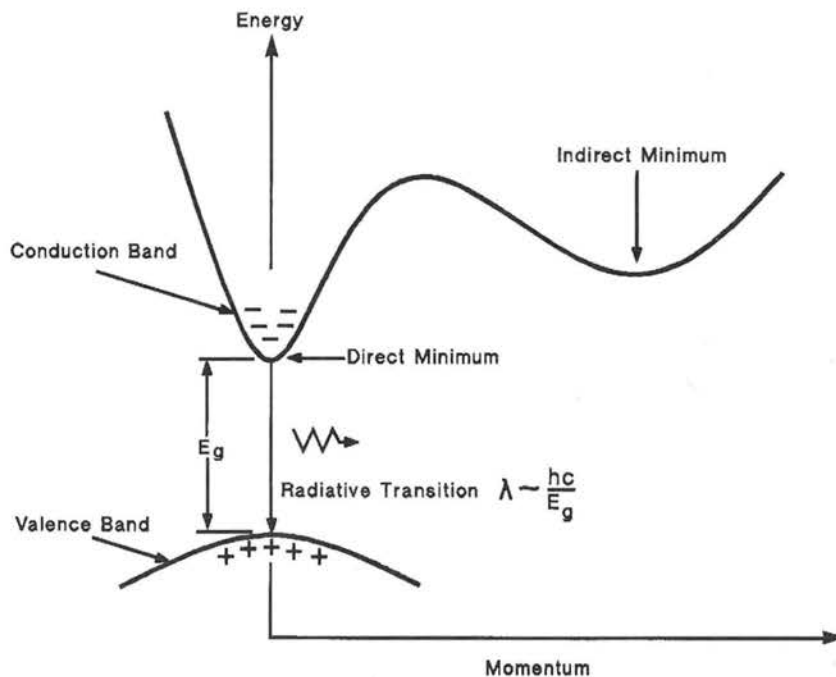


FIGURE 1 Energy band structure of a direct semiconductor showing radiative recombination of electrons in the conduction band with holes in the valence band.

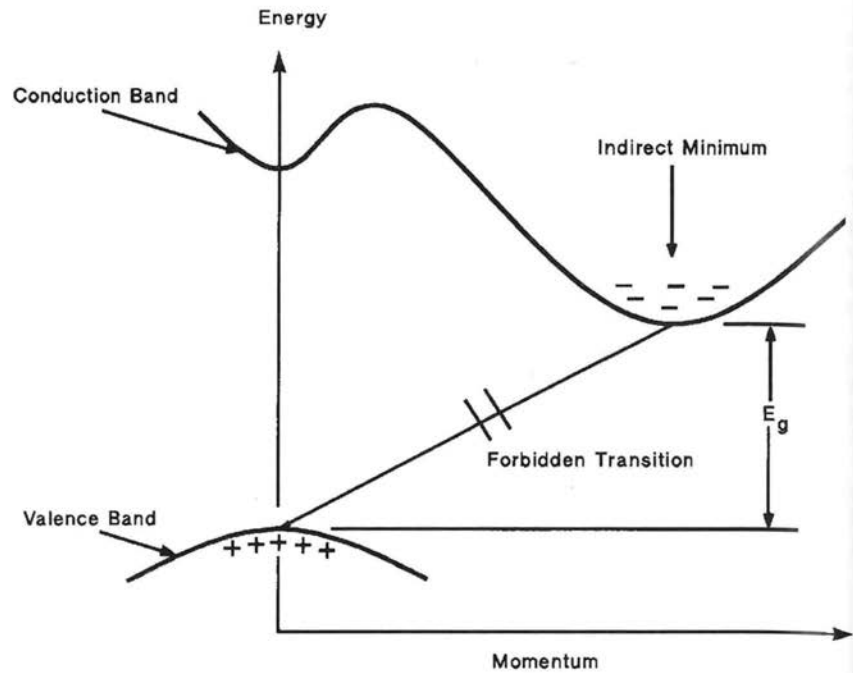


FIGURE 2 Energy band structure of an indirect semiconductor showing the conduction-band minima and valence-band maximum at different positions in momentum space. Radiative recombination of conduction-band electrons and valence-band holes is generally forbidden.

and Ref. 2). In practically all semiconductors the lower band, also known as the *valence band*, has a fairly simple structure, a paraboloid around the $\langle 0, 0, 0 \rangle$ crystalline direction. Holes will take up a position near the apex of the paraboloid and have very small momentum. The upper band, also known as the *conduction band*, is different for various semiconductor materials. All semiconductors have multiple valleys in the conduction band. Of practical interest are the valleys with the lowest energy. Semiconductor materials are classified as either *direct* or *indirect*.¹⁻³ In a direct semiconductor, the lowest valley in the conduction band is directly above the apex of the valence-band paraboloid. In an indirect semiconductor, the lowest valleys are not at $\langle 0, 0, 0 \rangle$, but at different positions in momentum/energy space (see Fig. 2). Majority or minority carriers mostly occupy the lowest energy states, i.e., holes near the top of the valence band paraboloid and electrons near the bottom of the lowest conduction-band valley.

In the case of a direct semiconductor the electrons are positioned directly above the holes at the same momentum coordinates. It is relatively easy to match up electrons and holes with proper momentum-conserving conditions. Thus, the resulting radiative lifetime τ_r is short. On the other hand, electrons in an indirect valley will find it practically impossible to find momentum-matching holes and the resulting radiative lifetime will be long. Injected carriers in indirect material generally recombine nonradiatively through defects.

In a direct semiconductor, such as GaAs, the radiative lifetime τ_r is in the range of 1–100 ns, depending on doping, temperature, and other factors. It is relatively easy to grow crystals with sufficiently low defect density such that τ_n is in the same range as τ_r .

For indirect semiconductors, such as germanium or silicon, the radiative recombination

process is extremely unlikely, and τ_r is in the range of seconds.¹ In this case, $\tau_r \gg \tau_n$, and practically all injected carriers recombine nonradiatively.

The wavelength of the photons emitted in a radiative recombination event is determined by the energy difference between the recombining electron-hole pair. Since carriers relax quickly to an energy level near the top of the valence band (holes), or the bottom of the conduction band (electrons), we have the following approximation for the wavelength λ of the emitted photon (see Fig. 1):

$$\lambda \approx hc/E_g \quad (3)$$

where h = Planck's constant, c = velocity of light, and E_g = bandgap energy.

This relation is only an approximation since holes and electrons are thermally distributed at levels slightly below the valence-band maximum and above the conduction-band minimum, resulting in a finite linewidth in the energy or wavelength of the emitted light. Another modification results from a recombination between a free electron and a hole trapped in a deep acceptor state. See Ref. 2 for a discussion of various recombination processes.

To change the wavelength or energy of the emitted light, one has to change the bandgap of the semiconductor material. For example, GaAs with a bandgap of 1.4 eV has an infrared emission wavelength of 900 nm. To achieve emission in the visible red region, the bandgap has to be raised to around 1.9 eV. This increase in E_g can be achieved by mixing GaAs with another material with a wider bandgap, for instance GaP with $E_g = 2.3$ eV. By adjusting the ratio of arsenic to phosphorous the bandgap of the resulting ternary compound, GaAsP, can be tailored to any value between 1.4 and 2.3 eV.³

The resulting band structure with varying As to P ratio is illustrated in Fig. 3. Note, the

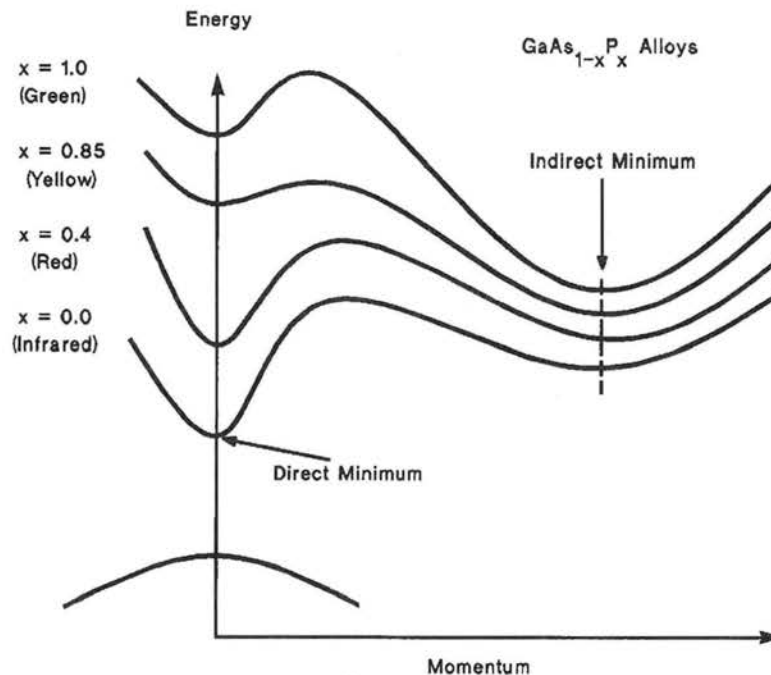


FIGURE 3 Energy band diagram for various alloys of the GaAs_{1-x}P_x material system showing the direct and indirect conduction-band minima for various alloy compositions.

two conduction-band valleys do not move upward in energy space at the same rate. The direct valley moves up faster than the indirect valley with increasing phosphorous composition. At a composition of around 40 percent GaP and 60 percent GaAs, the direct and indirect valleys are about equal in energy. When the valleys are approximately equal in energy, electrons in the conduction band can scatter from the direct valley into the indirect valley. While the direct valley electrons still undergo rapid radiative recombination, the indirect valley electrons have a long radiative lifetime and either have to be scattered back to the direct valley or they will recombine nonradiatively. In other words, near this crossover between direct and indirect valleys, the radiative efficiency drops off dramatically, and for compositions with greater than 40 percent phosphorous the direct radiative recombination is practically nonexistent.^{3,4}

The above discussion indicates that indirect semiconductors are not suitable for efficient generation of light through minority-carrier recombination. Fortunately, the introduction of so-called isoelectronic impurities can circumvent this limitation and introduces a new radiative recombination process.⁵ A commonly used isoelectronic trap is generated by substituting a nitrogen atom for phosphorous in the GaAsP system⁶⁻¹¹ (see Fig. 4). Since N and P both have five electrons in their outer shell, the trap is electrically neutral. However, the stronger electronegativity of N relative to P can result in capture of an electron from the conduction band. Since the electron is very tightly bound to the impurity atom, its wave function in momentum space is spread out and has reasonable magnitude at $(k=0)$ in momentum space.¹⁰ The negatively charged defect can attract a free hole to form a loosely bound electron-hole pair or "exciton." This electron-hole pair has a high probability to recombine radiatively. The energy of the emitted light is less than E_g . Another isoelectronic trap in GaP is formed by ZnO pairs (Zn on a Ga site and O on a P

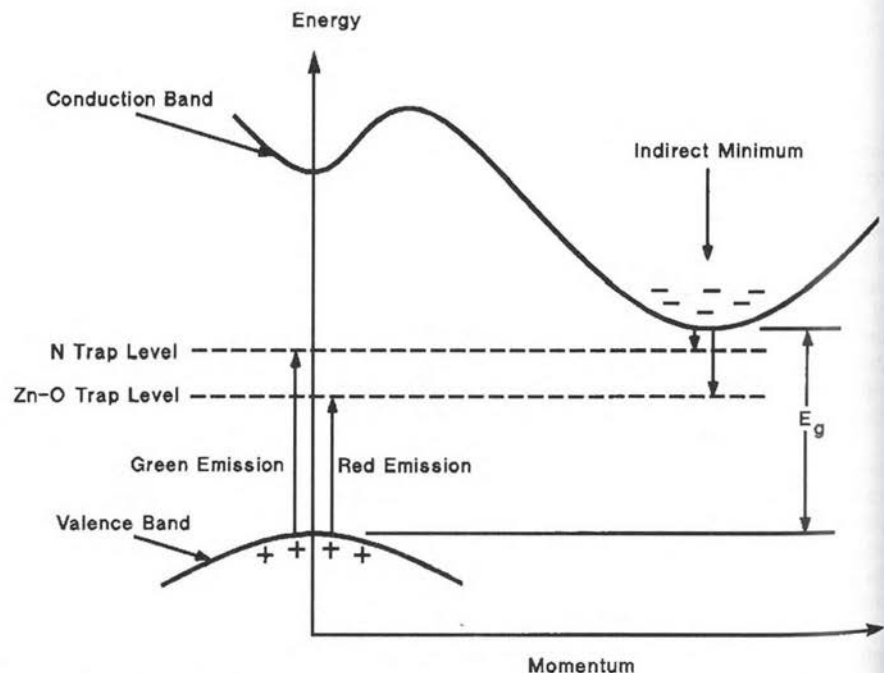


FIGURE 4 Formation of excitons (electron-hole pairs) by the addition of isoelectronic dopants N and ZnO to an indirect semiconductor. The excitons have a high probability to recombine radiatively.

site) (see Fig. 4). The ZnO trap is deeper than the N trap, resulting in longer wavelength emission in the red region of the spectrum.¹

The recombination process for exciton recombination is quite complex. For a detailed analysis, the reader is referred to Ref. 10. One result of this analysis is the recognition that the bound exciton has a relatively long lifetime in the range of 100–1000 ns. Light emission by exciton recombination is generally slower than emission due to direct band-to-band recombination.

12.4 LIGHT EXTRACTION

Generating light efficiently within a semiconductor material is only one part of the problem to build an efficient light source. The next challenge is the extraction of light from within the LED chip to the outside. The designer must consider total internal reflection.¹ According to Snell's law, light can escape from a medium of high index of refraction n_1 into a medium of low index refraction n_0 only if it intersects the surface between the two media at an angle from normal less than the critical angle θ_c with θ_c being defined by Eq. (4):

$$\theta_c = \arcsin n_0/n_1 \quad (4)$$

Most semiconductor LEDs have an isotropic emission pattern as seen from within the light-generating material. Assuming a cubic shape for the LED chip, because of internal reflections, only a small fraction of the isotropically emitted light can escape any of the six surfaces. As a case in point, let us calculate the emission through the top surface. For typical light-emitting semiconductors, n_1 is in the range of 2.9 to 3.6. If $n_1 = 3.3$ and $n_0 = 1.0$ (air), we find $\theta_c = 17.6^\circ$. The emission from an isotropic source into a cone with a half angle of θ_c is given by $(1 - \cos \theta_c)/2$. After correcting for Fresnel reflections, only 1.6 percent of the light generated escapes through the LED top surface into air. Depending on chip and p - n junction geometry, virtually all of the remaining light (98.4 percent) is reflected and absorbed within the LED chip.

The fraction of light coupled from chip to air is a function of the number of surfaces through which the chip can transmit light effectively. Most LED chips are called "absorbing substrate" (AS) chips. In such a chip, the starting substrate material (discussed later under "Substrate Technology") has a narrow bandgap and absorbs all the light with energy greater than the bandgap of the substrate. Consider the case of a GaAsP LED grown on a GaAs substrate. The emitted light ($E_g > 1.9$ eV) is absorbed by the GaAs substrate ($E_g = 1.4$ eV). Thus, a GaAsP-emitting layer on a GaAs substrate can transmit only through its top surface. Light transmitted toward the side surfaces or downward is absorbed.

To increase light extraction, the substrate or part of the epitaxial layers near the top of the chip has to be made of a material transparent to the emitted light. The "transparent substrate" (TS) chip is designed such that light transmitted towards the side surfaces within θ_c half-angle cones can escape. Assuming that there is negligible absorption between the point of light generation and the side walls, this increases the extraction efficiency by a factor of five (five instead of one escape cones).

In a TS chip, additional light can be extracted if the side walls are nonplanar, i.e., if light from outside an escape cone can be scattered into an escape cone. This process increases the optical path within the chip and is very dependent on residual absorption. In a chip with low absorption and randomizing side surfaces, most of the light should escape. Unfortunately, in practical LED structures, there are several absorption mechanisms left such as front and back contacts, crystal defects, and absorption in areas where secondary radiative recombination is inefficient.

TABLE 1 Extraction Efficiency into Air or Plastic for Three Types of Commonly Used LED Chips

Chip type	No. cones	Typical extraction efficiency	
		Air, %	Plastic, %
AS	1	1.5	4
Thick window	3	4.5	12
TS	5	7.5	18

A common approach is to use a hybrid chip with properties between AS and TS chips. These chips utilize a thick, transparent window layer above the light-emitting layer. If this layer is sufficiently thick, then most of the light in the top half of the cones transmitted toward the side surfaces will reach the side of the chip before hitting the substrate. In this case of hybrid chips, the efficiency is between that of AS and TS chips as shown in Table 1.

Another important way to increase extraction efficiency is derived from a stepwise reduction in the index of refraction from chip to air. If the chip is first imbedded in a material with an intermediate index, i.e., plastic with $n_2 = 1.5$, then the critical angle θ_c between chip and plastic is increased to 27° . The extraction efficiency relative to air increases by the ratio of $(n_2/n_0)^2$ plus some additional correction for Fresnel-reflection losses. The gain from plastic encapsulation is usually around $2.7\times$ compared to air. Chips with multipath internal reflection will result in lower gains. It is important to note that this gain can be achieved only if the plastic/air interface can accommodate the increased angular distribution through proper lenslike surface shaping or efficient scattering optics. Table 1 illustrates the approximate extraction efficiencies achieved by the three dominant chip structures in air and in plastic. The numbers assume only first-pass extraction, limited absorption, and no multiple reflections within the chip.

12.5 DEVICE STRUCTURES

LED devices come in a broad range of structures. Each material system (see following section) requires a different optimization. The only common feature for all LED structures is the placement of the p - n junction where the light is generated. The p - n junction is practically never placed in the bulk-grown substrate material for the following reasons:

- The bulk-grown materials such as GaAs, GaP, and InP usually do not have the right energy gap for the desired wavelength of the emitted light.
- The light-generating region requires moderately low doping that is inconsistent with the need for a low series resistance.
- Bulk-grown material often has a relatively high defect density, making it difficult to achieve high efficiency.

Because of these reasons, practically all commercially important LED structures utilize a secondary growth step on top of a single-crystal bulk-grown substrate material. The secondary growth step consists of a single-crystal layer lattice matched to the substrate.

This growth process is known as *epitaxial growth* and is described in a later section of this chapter.

The commonly used epitaxial structures can be classified into the following categories:

- Homojunctions
 - grown
 - diffused
- Heterojunctions
 - single confinement
 - double confinement

Grown Homojunctions

Figure 5 illustrates one of the simplest design approaches to an LED chip. An n -type GaAs layer with low to moderate doping density is grown on top of a highly doped n -type substrate by a vapor or liquid-phase epitaxial process (see "Epitaxial Technology"). After a growth of 5–10 μm , the doping is changed to p -type for another 5–10 μm . A critical dimension is the thickness of the epitaxial p -layer. The thickness should be larger than the diffusion length of electrons. In other words, the electrons should recombine radiatively in the epitaxially grown p layer before reaching the surface. The p -layer should be of sufficiently high quality to meet the condition for efficient recombination, i.e., $\tau_n \gg \tau_r$. In addition, the side surfaces may have to be etched to remove damage. Damage and other defects where the p - n junction intercepts the chip surface can lead to a substantial leakage current that reduces efficiency, especially at low drive levels.

The structure of Fig. 5 was used in some of the earliest infrared emitters (wavelength 900 nm). Efficiency was low, typically 1 percent. Modern infrared emitters use Si-doped

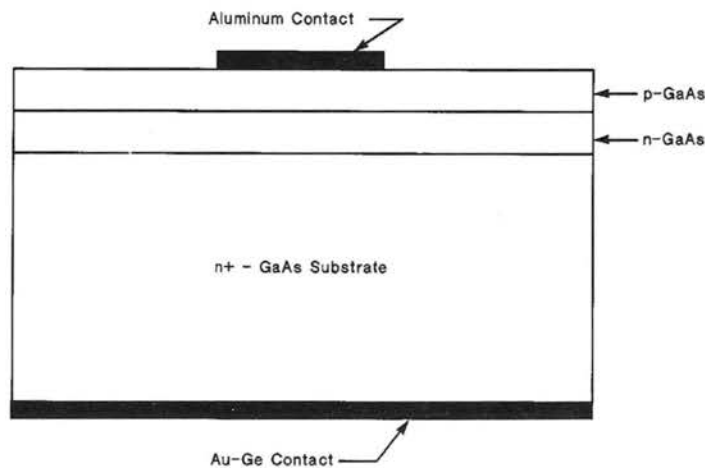


FIGURE 5 Cross section of an infrared LED chip. A p - n junction is formed by epitaxially growing n - and p -doped GaAs onto an n +-doped GaAs substrate.

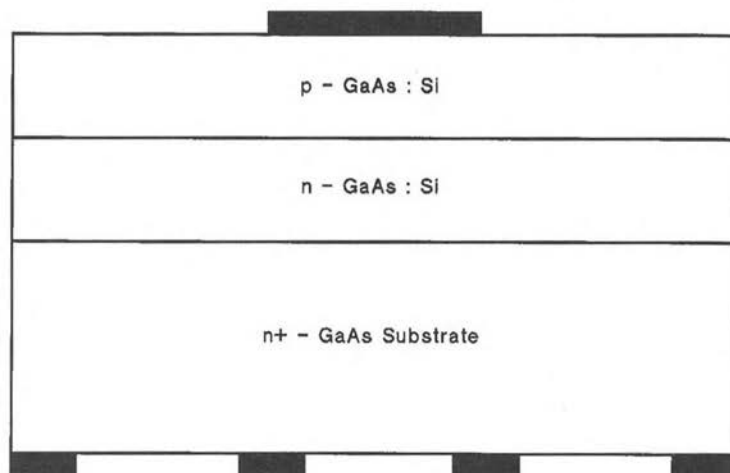


FIGURE 6 A high-efficiency IR LED made by LPE growth of GaAs that is doped with silicon on both the p and n sides of the junction. To increase external quantum efficiency, a partly reflective back contact is employed.

GaAs (see Fig. 6). The detailed recombination mechanism in GaAs:Si even today is quite controversial and goes beyond the scope of this publication. The recombination process has two important characteristics: (1) the radiative lifetime τ_r is relatively slow, i.e., in the range of $1 \mu\text{s}$; and (2) the wavelength is shifted to 940 nm . At this wavelength the GaAs substrate is partly transparent, making this device a quasi-TS structure with efficiencies into plastic of 5 to 10 percent.

Diffused Homojunction

The chip structure of Fig. 5 can also be produced by a zinc (Zn) diffusion into a thick n -layer. The commercially most significant structure of this type is shown in Fig. 7. By replacing 40 percent of the As atoms with P atoms, the bandgap is increased to 1.92 eV to make a GaAsP LED that emits visible red light. In this case, the p - n junction is diffused selectively by using a deposited layer of silicon-nitride as a diffusion mask. This structure of Fig. 7 has several advantages over the structure of Fig. 5. Lateral diffusion of Zn moves the intersection of the junction with the chip surface underneath the protecting Si_3N_4 layer. This layer protects the junction from contamination and adds to the long-term stability of the device. In addition, it is important for applications requiring more than one clearly separated light-emitting area. For instance, seven-segment displays, such as those used in the early hand-held calculators, are made by diffusing seven long and narrow stripes into a single chip of GaAsP material. (See Fig. 8.) This chip consists of eight (seven segments plus decimal point) individually addressable p -regions (anodes) with a common n -type cathode. Such a chip is feasible only in an AS-type structure because the individual segments have to be optically isolated from each other. A TS-type structure results in unacceptable levels of crosstalk.

Figure 7 shows another feature of practical LED devices. The composition with 40 percent P and 60 percent As has a lattice constant (atomic spacing) that is different from the GaAs substrate. Such a lattice mismatch between adjacent layers would result in a very high density of dislocations. To reduce this problem to an acceptable level, one has to slowly increase the phosphorous composition from 0 percent at the GaAs interface to 40

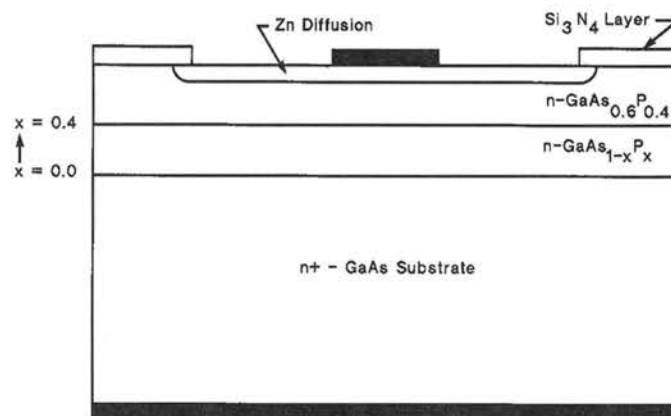


FIGURE 7 GaAsP LED which emits at 650 nm. On a GaAs substrate, a layer is grown whose composition varies linearly from GaAs to $\text{GaAs}_{0.6}\text{P}_{0.4}$, followed by a layer of constant composition. Zinc is selectively diffused using an Si_3N_4 mask to form the light-emitting junction.

percent over a 10–20- μm -thick buffer layer. Typically, the buffer layer is graded linearly. The phosphorous composition is increased linearly from bottom to top. The thicker the buffer layer, the lower is the resulting dislocation density. Cost constraints keep the layer in the 10–15- μm range. The layer of constant composition (40 percent P) has to be thick enough to accommodate the Zn diffusion plus the diffusion length of minority carriers. A thickness range of 5–10 μm is typical.

Another variation of a homojunction is shown in the TS-chip structure of Fig. 9. Instead of an absorbing GaAs substrate, one starts with a transparent GaP substrate. The graded layer has an inverse gradient relative to the chip shown in Fig. 7. The initial growth is 100 percent GaP phasing in As linearly over 10–15 μm . At 15 percent As, the emission is in the yellow range (585 nm), at 25 percent in the orange range (605 nm), and at 35 percent in the red range (635 nm). Figure 3 shows the approximate band structure of this material system. The composition range mentioned above has an indirect band structure. To obtain efficient light emission, the region of minority-carrier injection is doped with nitrogen forming an isoelectronic recombination center (see exciton recombination in the first section).

Figure 9 shows an important technique to increase extraction efficiency. In a TS-chip, the major light loss is due to free carrier absorption at the alloyed contacts. Rather than covering the entire bottom surface of the chip with contact metal, one can reduce the contact area either by depositing small contact islands (see Fig. 6) or by placing a dielectric mirror (deposited SiO_2) between the substrate and the unused areas of the back contact (Fig. 9). This dielectric mirror increases the efficiency by 20 to 50 percent at the expense of higher manufacturing cost.

Single Heterojunctions

Heterojunctions introduce a new variable: local variation of the energy bandgap resulting in carrier confinement. Figure 10 shows a popular structure for a red LED emitter chip. A p -type layer of GaAlAs with 38 percent Al is grown on a GaAs substrate. The GaAlAs alloy system can be lattice-matched to GaAs; therefore, no graded layer is required as in the GaAsP system of Fig. 7. Next, an n -type layer is grown with 75 percent Al. The variation of E_g from substrate to top is illustrated in Fig. 11. Holes accumulate in the

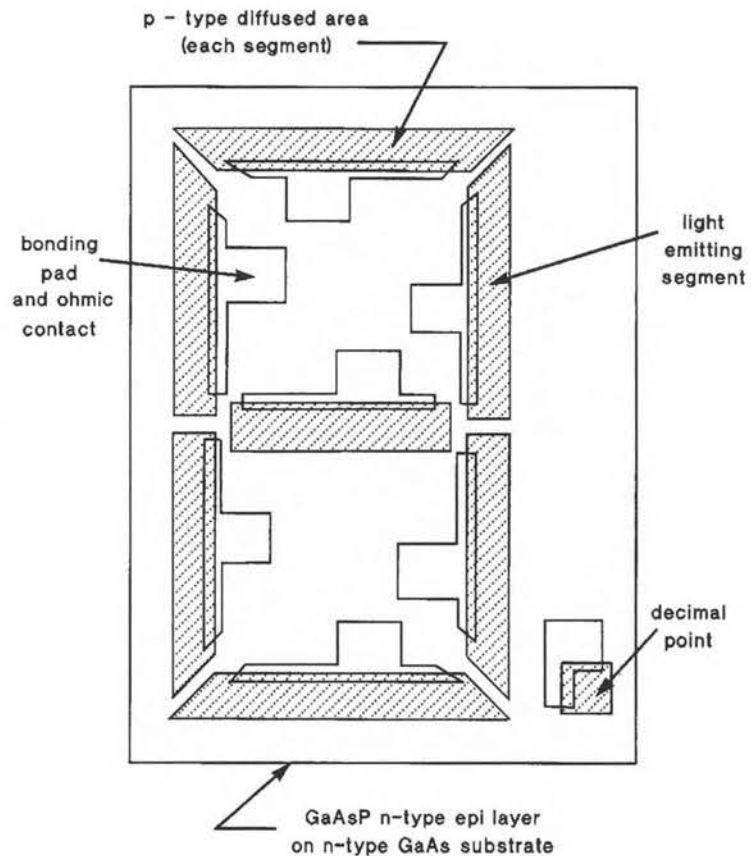


FIGURE 8 Monolithic seven-segment display chip with eight separate diffused regions (anodes) and a common cathode (the GaAs substrate).

GaAlAs p -layer with the narrower bandgap. Electrons are injected from the n -layer into this p -layer. The holes have insufficient energy to climb the potential barrier into the wide-bandgap material. Holes are confined to the p layer. In the p layer, the radiative recombination time is very short because of the high concentration of holes. As a result, the internal quantum efficiency is quite high. A variation of this structure is a widely used infrared emitter that emits at 880 nm.

Double Heterostructures

The double heterostructure shown in Fig. 12 repeats the p -side confinement of Fig. 10 on the n side. An n -type buffer layer is grown on the GaAs substrate to create a high-quality surface onto which the first n -type GaAlAs confinement layer with 75 percent Al is grown. The active or light-generation layer is a 3- μm -thick p -type layer with 38 percent Al. The top p -type confinement layer again uses 75 percent Al. This structure with the energy-band diagram of Fig. 13 has two advantages: (1) There is no hole injection into the n -type layer with reduced efficiency and a slow hole recombination in the lowly doped n layer. (2) The high electron and hole density in the active layer reduces τ_r , thus increasing device speed.

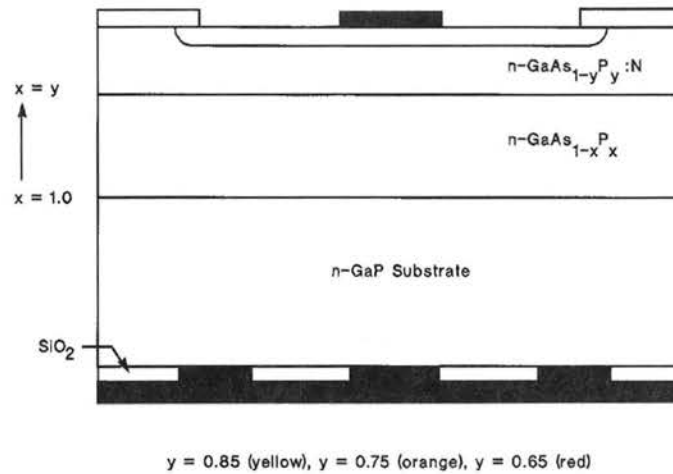


FIGURE 9 Cross section of a $\text{GaAs}_{1-x}\text{P}_x$ LED which, by changing the composition “x,” produces red, orange, or yellow light. The top layer is doped with nitrogen to increase the quantum efficiency. The GaP substrate is transparent to the emitted light. Also shown is a reflective back contact made by depositing the contact metallization on top of SiO_2 .

and efficiency. The increased speed is quite important for LED sources in fiber-optic communication applications. (See “Fiber Optics” subsection later in this chapter.)

The double heterostructure of Fig. 12 represents a one-dimensional containment of injected carriers. Injection and light emission occurs across the entire lateral dimension of the chip. For fiber-optic applications, the light generated over such a large area cannot be effectively coupled into small-diameter fiber. A rule of thumb for fiber coupling requires that the light-emitting area be equal to or, preferably, smaller than the fiber core diameter. This rule requires lateral constraint of carrier injection. The localized diffusion of Fig. 7 is not applicable to grown structures such as those in Fig. 12. The preferred solution inserts

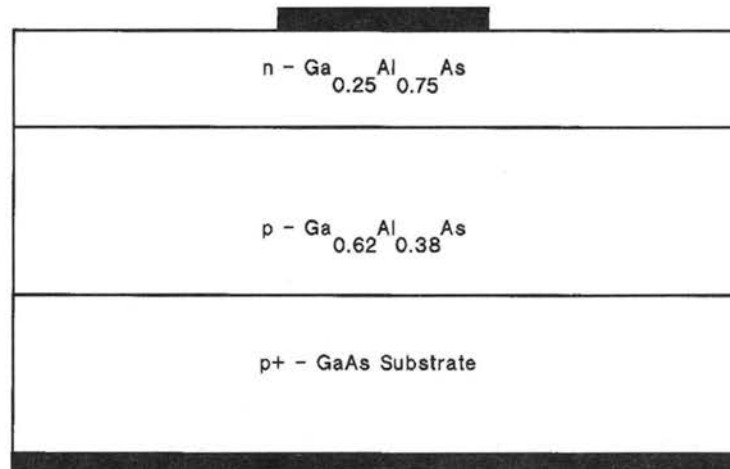


FIGURE 10 Cross section of a single heterostructure LED.

er into
into the
diative
result,
ly used

g. 10 on
-quality
grown.
Al. The
gy-band
pe layer
(2) The
ze speed

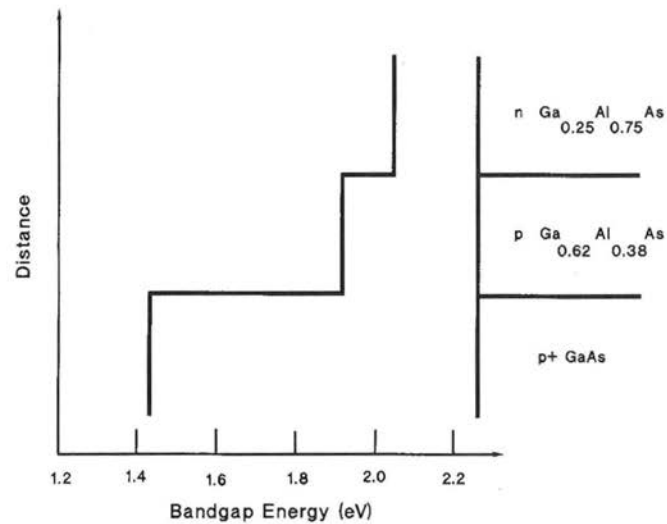


FIGURE 11 Variation in energy bandgap for the various layers in the GaAlAs LED shown in Fig. 10.

an n layer between buffer and lower confinement layer (see Fig. 14). A hole etched into the n layer allows current flow. Outside of this hole, the p - n junction between n layer and lower confinement layer is reverse-biased, thus blocking any current flow. The disadvantage of this approach is a complication of the growth process. In a first growth process, the n layer is grown. Then a hole is etched into the n layer using standard photolithographic etching techniques. Finally, a second epitaxial growth is used for the remaining layers.

Another technique to constrain current injection utilizes a small ohmic contact.¹² It is used frequently in conjunction with InP-based fiber-optic emitters (see Fig. 15). A SiO_2 layer limits contact to a small-diameter (typically $25\text{ }\mu\text{m}$) hole that results in a relatively

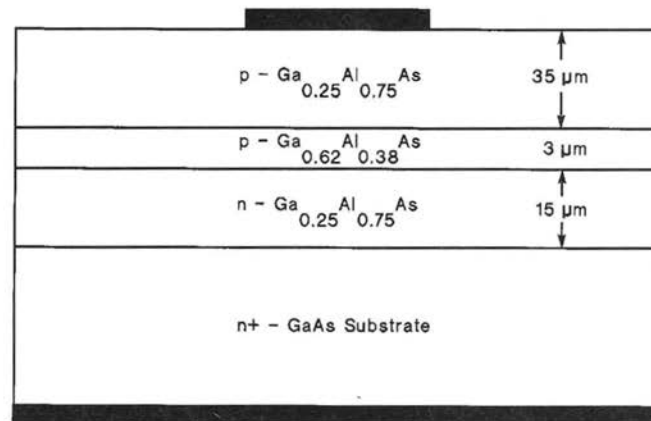


FIGURE 12 Structure of a double heterostructure (DH) GaAlAs LED. The DH is composed of a $\text{Ga}_{0.62}\text{Al}_{0.38}\text{As}$ layer surrounded on either side by a $\text{Ga}_{0.25}\text{Al}_{0.75}\text{As}$ layer. The thick top layer acts as window to increase light extraction through the side walls of the chip.

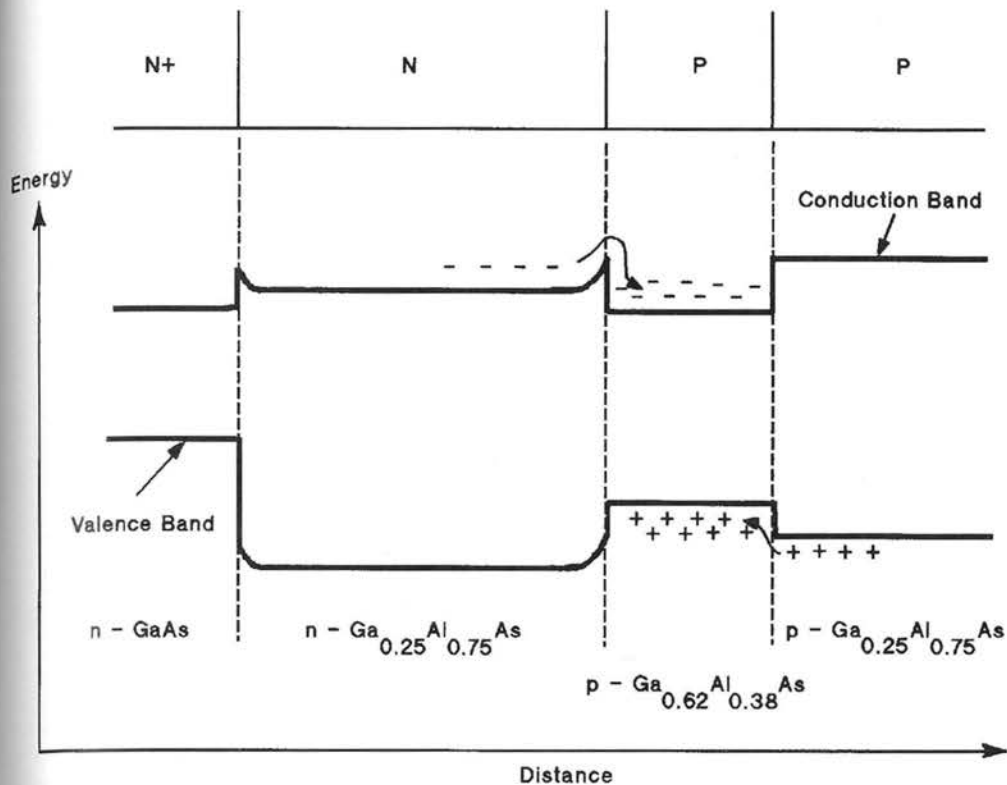


FIGURE 13 Energy band diagram of the GaAlAs LED shown in Fig. 12. The LED is forward-biased. Electrons and holes are confined in the p -doped $\text{Ga}_{0.62}\text{Al}_{0.38}\text{As}$ layer, which increases the radiative recombination efficiency.

small light-emitting area. The etched lens shown for this structure helps to collimate the light for more efficient coupling onto the fiber.¹²

12.6 MATERIALS SYSTEMS

The $\text{GaAs}_{1-x}\text{P}_x$ System

The most widely used alloy for LEDs is the ternary $\text{GaAs}_{1-x}\text{P}_x$ system, including its two binary components GaAs and GaP. This system is best described by the composition parameter " x " with $0 \leq x \leq 1$. For $x = 0$, we have GaAs and for $x = 1$ the composition is GaP. For $x \leq 0.4$, the alloy has a direct bandgap. GaAs was developed in the early 1960s as an infrared emitter with a wavelength of 910 nm and an efficiency in the range of 1 percent. This emitter was soon followed by a Si-doped variety. As discussed earlier, this leads to an emission wavelength of 940 nm, a wavelength at which the GaAs substrate is partly transparent. The resulting efficiency is increased substantially and, depending on configuration, is in the 5 to 10 percent range. However, the recombination process is quite slow, resulting in rise and fall times in the 0.5–1.0 μs range (see Table 2). One other drawback is caused by the low absorption coefficient of Si detectors at 940 nm. To absorb 90 percent of the light requires a detector thickness of 60–70 μm . Conventional photo

12.16 OPTICAL SOURCES

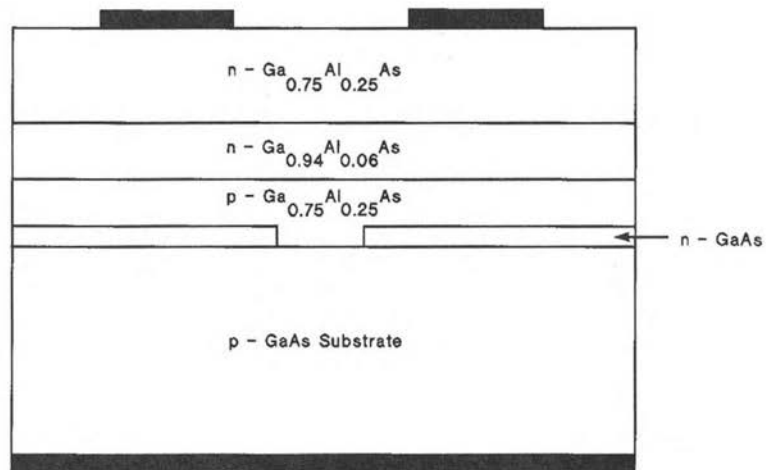


FIGURE 14 Cross section of an LED with three-dimensional carrier confinement. A DH structure is used to confine injected carriers in the $\text{Ga}_{0.94}\text{Al}_{0.06}\text{As}$ layer (direction perpendicular to the junction). The patterned, n -type GaAs layer is used to limit current flow in the lateral direction. The small emitting area and the 820-nm emission of this LED makes it ideal for fiber-optic applications.

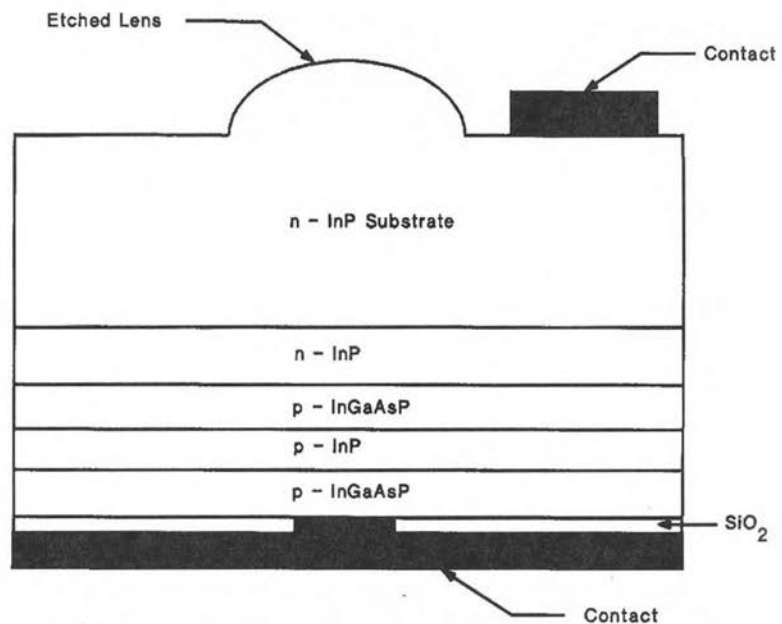


FIGURE 15 Structure of a 1300-nm LED used for optical fiber communications. The cross section shows the DH-layer configuration, limited area back contact for emission-size control, and an etched lens at the top of the chip which magnifies ($M = 2$) the source area and collimates the light for effective coupling into a fiber.

TABLE 2 Performance Summary of LED Chips in the $\text{GaAs}_{1-x}\text{P}_x$ System

x	Growth process*	Isoel. dopant	Substrate†	Dominant wavelength (nm)	Color‡	Luminous efficacy§ (lm/A)	Quantum efficiency§ (%)	Speed (ns)
0	LPE	—	AS	910	IR		1	50
0	LPE	—	TS	940	IR		10	1000
0.3	VPE	—	AS	700	IR		0.5	50
0.4	VPE	—	AS	650	Red	0.2		50
0.65	VPE	N	TS	635	Red	2.5		300
0.75	VPE	N	TS	605	Orange	2.5		300
0.85	VPE	N	TS	585	Yellow	2.5		300
1.0	LPE	N	TS	572	Y/green	6		300
1.0	LPE	—	TS	565	Green	1		
1.0	LPE	ZnO	TS	640	Red	1		

* LPE = liquid phase epitaxy; VPE = vapor phase epitaxy.

† AS = absorbing substrate; TS = transparent substrate.

‡ IR = infrared.

§ Into plastic (index = 1.5).

transistors are quite suitable as detectors. Integrated photo ICs with their 5–7- μm -thick epitaxial layers are very inefficient as detectors at 940 nm.

To shift the wavelength toward the near-infrared or into the visible spectrum, one has to grow a ternary alloy, a mixture between GaAs and GaP. Of commercial interest are two alloys with $x = 0.3$ and $x = 0.4$ grown on a GaAs substrate (see Table 2). The $x = 0.4$ alloy was the first commercially produced material with a wavelength in the visible range of the spectrum. Grown on an absorbing substrate, it has a modest luminous efficacy of around 0.2 lm/A. [Luminous efficacy is the luminous (visible) flux output measured in lumens divided by the electrical current input.] The absorbing substrate allows the integration of multiple light sources into a single chip without crosstalk. Such monolithic seven-segment displays became the workhorse display technology for hand-held calculators from 1972 to 1976.^{9,13} Today this alloy is used in LED arrays for printers.¹⁴

The $x = 0.3$ alloy with a wavelength of 700 nm became important in the mid 1970s as a light source in applications using integrated photodetectors. It has three to five times the quantum efficiency of the $x = 0.4$ alloy (see Table 2), but has a lower luminous efficacy because of the much-reduced eye sensitivity at 700 nm.

For $x > 0.4$, the GaAsP material system becomes indirect (see earlier under "Light-Generation Process" and Fig. 3). The quantum efficiency decreases faster than the increase in eye sensitivity.⁴ The only way to achieve a meaningful efficiency is through the use of isoelectronic dopants as described earlier. The choices for isoelectronic dopants that have been successful are N for GaAsP¹¹ and N or ZnO for GaP. Nitrogen doping is used widely for alloys with $x = 0.65$ to $x = 1.0$.^{6–10} The resulting light sources cover the wavelength range from 635 nm to approximately 565 nm (see Table 2). Since these alloys are either GaP or very close in composition to GaP, they are all grown on GaP substrates. The resulting transparent-substrate chip structure increases the luminous efficacy.

In the case of the binary GaP compound, the dominant wavelength depends on N concentration. With low concentrations, practically all N atoms are isolated in single sites. With increasing concentration, N atoms can arrange themselves as pairs or triplets. The resulting electron traps have lower energy states which shift the emitted light toward longer wavelength. Phonon coupling can also reduce the emission energy. Commercially significant are two compositions: (1) undoped GaP which emits at 565 nm (dominant wavelength) with a reasonably green appearance and a low efficiency of around 1 lm/A,

and (2) GaP with a nitrogen concentration in the range of 10^{19} cm^{-3} with a substantially higher luminous efficiency of around 6 lm/A at 572 nm. At this wavelength, the color appearance is yellow-green, often described as chartreuse.

Three nitrogen-doped ternary alloys of GaAsP are commercially important for red, orange, and yellow. The red source with $x = 0.65$ has an efficiency in the range of 2–3 lm/A. With increasing bandgap or decreasing wavelength, the drop in quantum efficiency is compensated by an increase in eye sensitivity, resulting in a practically wavelength-independent luminous efficiency for the range of 635 to 585 nm.^{8,15}

ZnO-doped GaP is an interesting material. The quantum efficiency of such chips is relatively high, around 3 percent. However, the linewidth is quite broad. The quantum efficiency peaks at 700 nm, but the luminous efficiency peaks at 640 nm (dominant wavelength). In other words, most of the photons are emitted at wavelengths with low eye sensitivity. Another problem of GaP:ZnO is saturation. The deep ZnO electron trap causes very slow exciton recombination. At high injection currents, all traps are saturated and most of the injected carriers recombine nonradiatively. At low injection levels ($\leq 1 \text{ A/cm}^2$), the efficacy is relatively high, 3–5 lm/A. At a more useful density of 10–30 A/cm², the emission saturates, resulting in an efficacy of around 1 lm/A.

The $\text{Al}_x\text{Ga}_{1-x}\text{As}$ System

The $\text{Al}_x\text{Ga}_{1-x}\text{As}$ material system has a direct bandgap for $0 \leq x \leq 0.38$. This system has one very significant advantage over the GaAsP system described above, the entire alloy range from $x = 0$ to $x = 1$ can be lattice-matched to GaAs. In other words, every alloy composition can be directly grown on any other alloy composition without the need for transition layers. This feature allows the growth of very abrupt heterojunctions, i.e., abrupt transition in composition and bandgap. These heterojunctions add one important property not available in the GaAsP system: carrier containment (see earlier under “Device Structures”). Carrier containment reduces the movement of injected carriers in a direction perpendicular to the junction. Thus, carrier density can be increased beyond the diffusion-limited levels. This results in increased internal quantum efficiency and higher speed. Another benefit is reduced absorption and improved extraction efficiency (under “Light Extraction”).

Of practical significance are two compositions: $x = 0.06$ and $x = 0.38$ (see Table 3). Both compositions exist in single and double heterojunction variations (see under “Device Structures”). The double heterojunctions usually have a 1.5–2.0-times advantage in efficiency and speed. In all cases, the efficiency strongly depends on the thickness of the window layer and, to a lesser degree, on the thickness of the transparent layer between active layer and absorbing substrate (see Fig. 12). Chips with a transparent substrate have an additional efficiency improvement of 1.5–3.0 times again, depending on layer thickness

TABLE 3 Performance Summary of LED Chips in the $\text{Al}_x\text{Ga}_{1-x}\text{As}$ System

x	λ (nm)	Substrate*	Structure†	Efficiency or efficacy	Speed (ns)
0.06	820	AS, TW	DH	8%	30
0.06	820	TS	DH	15%	30
0.38	650	AS, TW	SH	4 lm/A	
0.38	650	AS, TW	DH	8 lm/A	
0.38	650	TS	DH	16 lm/A	

* AS = absorbing substrate; TW = thick window layer; TS = transport substrate.

† DH = double heterostructure; SH = single heterostructure.

and contact area. The efficiency variation is best understood by counting exit cones as described in the text in conjunction with Table 1. For $x = 0.06$, the internal quantum efficiency of a double heterojunction approaches 100 percent. For $x = 0.38$, the direct and indirect valleys are practically at the same level and the internal quantum efficiency is reduced to the range of 50 percent, again dependent upon the quality of the manufacturing process.

The best compromise for efficiency and speed is the $x = 0.06$ alloy as a double heterostructure. Depending on layer thickness, substrate, and contact area, these devices have efficiencies of 5 to 20 percent and rise/fall times of 20–50 ns. This alloy is becoming the workhorse for all infrared applications demanding power and speed. A structural variation as shown in Fig. 14 is an important light source for fiber-optic communication.

The $x = 0.38$ alloy is optimized for applications in the visible spectrum. The highest product of quantum efficiency and eye response is achieved at $x = 0.38$ and $\lambda = 650$ nm. The single heterostructure on an absorbing substrate has an efficacy of around 4 lm/A. The equivalent double heterostructure is in the 6–8 lm/A range. On a transparent GaAlAs substrate, the efficacy is typically in the 15–20 lm/A range and results of as high as 30 lm/A have been reported in the literature.¹⁶ The major application for these red LEDs is in light-flux-intensive applications, such as message panels and automotive stoplights. A variation optimized for speed is widely used for optical communication using plastic fiber.

The AlInGaP System

The AlInGaP system has most of the advantages of the AlGaAs system with the additional advantage that it has a higher-energy direct energy gap of 2.3 eV that corresponds to green emission at 540 nm. AlInGaP can be lattice-matched to GaAs substrates. Indium occupies about half of the Group III atomic sites. The ratio of aluminum to gallium can be changed without affecting the lattice match, just as it can in the AlGaAs material system, since AlP and GaP have nearly the same lattice spacing. This enables the growth of heterostructures that have the efficiency advantages described in the previous section.

Various AlInGaP device structures have been grown. A simple DH structure with an AlInGaP active layer surrounded by higher bandgap AlInGaP confining layers has been effective for injection lasers, but has not produced efficient surface-emitting LEDs.¹⁷ The main problem has been that AlInGaP is relatively resistive and the top AlInGaP layer is not effective in distributing the current uniformly over the chip. This is not a problem with lasers since the top surface is covered with metal and the light is emitted from the edge of the chip.

The top layer must also be transparent to the light that is generated. Two window layers used are AlGaAs or GaP on top of the AlInGaP heterostructure.^{18,19} AlGaAs has the advantage that it is lattice-matched and introduces a minimum number of defects at the interface, but it has the disadvantage that it is somewhat absorptive to the yellow and green light which is generated for high-aluminum compositions. The highest AlInGaP device efficiencies have been achieved using GaP window layers.^{18,20} GaP has the advantage that it is transparent to shorter wavelengths than AlGaAs and that it is easy to grow thick GaP layers, using either VPE or LPE, on top of the AlInGaP DH that was grown by MOVPE. Both VPE and LPE have substantially higher growth rates than MOCVD. The various growth techniques are discussed later under "Epitaxial Technology."

AlInGaP devices with 45- μ m-thick GaP window layers have achieved external quantum efficiencies exceeding 5 percent in the red and yellow regions of the emission spectrum.²¹ This is more than twice as bright as devices that have thinner AlGaAs window layers.

Green-emitting AlInGaP devices have also been grown which are brighter than the conventional GaP and GaP:N green emitters.^{18,20,21} Substantial further improvement in green is expected since the quantum efficiency is not as high as would be expected based on the energy position of the transition from a direct to an indirect semiconductor.

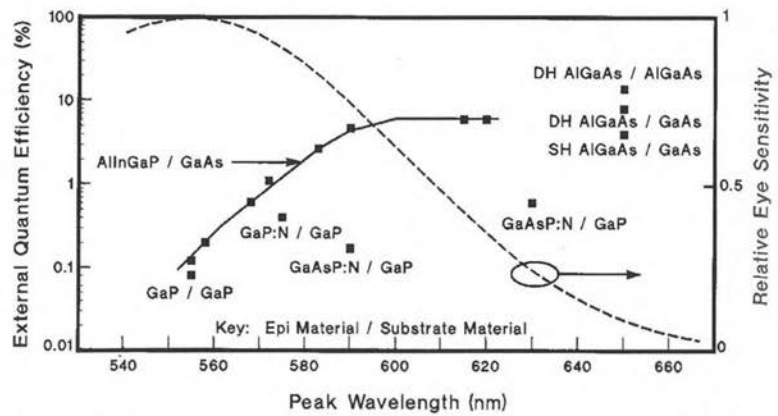


FIGURE 16 External quantum efficiency as a function of peak wavelength for various types of visible LEDs. Below 590 nm the efficiency of AlInGaP LEDs decreases due to the approaching transition from direct to indirect semiconductor. The human eye sensitivity curve is also shown. Since the eye response increases sharply from 660 to 540 nm, it partially makes up for the drop in AlInGaP LED efficiency. The resulting luminous performance is shown in Fig. 17.

The performance of AlInGaP LEDs compared to the most important other types of visible emitters is shown in Figs. 16 and 17. GaAsP on a GaAs substrate and GaP:ZnO are not shown since their luminous efficacy are off the chart at the bottom and the lower-right-hand corner, respectively. It is clear from Fig. 17 that the luminous efficacy of AlInGaP is substantially higher than the other technologies in all color regions except for red beyond 640 nm. Since forward voltage is typically about 2V, the lumen/watt value for a given device is about one-half the lumen/ampere value that is given in Tables 2 and 3. The quantum efficiency of AlInGaP is also better than all of the other technologies except for the highest-performance AlGaAs devices operating at about 650 nm. Because of the eye

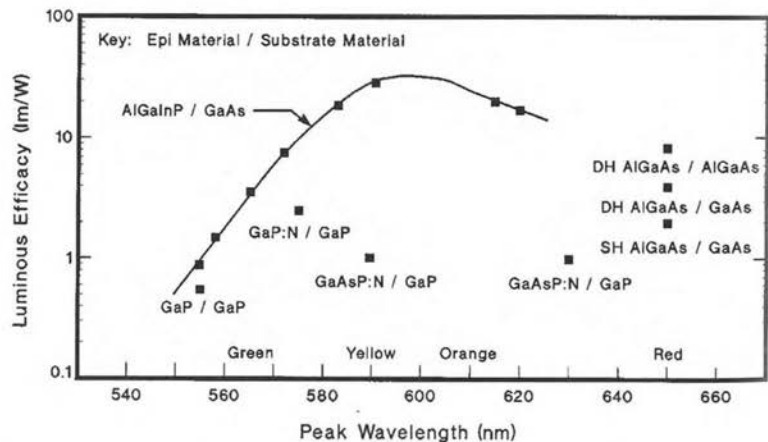


FIGURE 17 Luminous efficacy for AlInGaP LEDs versus wavelength compared to other LED technologies. AlInGaP LEDs are more than an order of magnitude brighter in the orange and yellow regions than other LEDs. AlInGaP LEDs compare favorably to the best AlGaAs red LEDs.

sensitivity variations (see C.I.E. curve in Fig. 16), the 620 nm (red/orange) AlInGaP devices have a higher luminous efficiency than 650 nm AlGaAs LEDs (see Fig. 17).

Blue LED Technology

Blue emitters have been commercially available for more than a decade, but have only begun to have a significant impact on the market in the last few years. SiC is the leading technology for blue emitters with a quantum efficiency of about 0.02 percent and 0.04 lm/A luminous performance. SiC devices are not much used due to their high price and relatively low performance efficiency.

Other approaches for making blue LEDs are the use of II-VI compounds such as ZnSe or the nitride system GaN, AlGaIn, or AlGaInN. It has been difficult to make good *p-n* junctions in these materials. Recently improved *p-n* junctions have been demonstrated in both ZnSe^{22,23} and GaN.²⁴ Device performance is still in the 0.1 lm/A range and reliability is unproven. However, this recent progress is very encouraging for blue-emission technology and could lead to a high-performance device in the next few years. Both ZnSe and the nitride system have a major advantage over SiC because they are direct bandgap semiconductors, so a much higher internal quantum efficiency is possible. However, it is difficult to find suitable lattice-matched substrates for these materials.

12.7 SUBSTRATE TECHNOLOGY

Substrate Criteria

There are several requirements for substrates for LEDs. The substrate must be as conductive as possible both thermally and electrically to minimize power loss. In order to minimize defects it should match the epitaxial layers as closely as possible in atomic lattice spacing, and in the coefficient of thermal expansion. The substrate should also have a low defect density itself. Finally, the substrate should, if possible, be transparent to the light generated by the LED structure since this will enhance the external quantum efficiency.

Substrate Choices

The substrates used for nearly all visible LEDs are GaAs and GaP. GaAs or InP is used for infrared devices, depending upon the device structure required. Substrate parameters are summarized in Table 4, along with Si and Ge for comparison. GaAs is used for the AlGaAs and AlInGaP material systems since they can be lattice-matched to it. GaAs is

TABLE 4 Properties of Common Semiconductor Substrates

Substrate	Lattice parameter	Energy gap @ 300 K (eV)		Melting point (°C)
GaAs	5.653	1.428	Direct	1238
GaP	5.451	2.268	Indirect	1467
InP	5.868	1.34	Direct	1062
Si	5.431	1.11	Indirect	1415
Ge	5.646	0.664	Indirect	937

also used for the $\text{GaAs}_{1-x}\text{P}_x$ system for $x \leq 0.4$ because it is more nearly lattice-matched and, since it is absorbing, it is useful in multiple-junction devices where optical crosstalk must be minimized (see under "Diffused Homojunction" and Fig. 8). GaP is used for compositions of $x > 0.6$ due to its transparency and closer lattice match. However, neither GaP nor GaAs are well matched to GaAsP, and grading layers are required to grow epitaxial layers with decent quality. InP is the choice for long wavelength emitters made using the InGaAs or InGaAsP materials systems, due to the better lattice match.

Substrate Doping

Generally, substrates are n -type and are doped with Te, S, or Si, although sometimes Se and Sn are also used. In some cases, particularly for some AlGaAs LEDs and laser structures, p -type substrates are required and Zn is nearly always the dopant. The doping levels are typically in the 10^{18}-cm^{-3} range. Basically, the substrates are doped as heavily as possible to maximize conductivity. However, the doping must be below the solubility limit to eliminate precipitates and other defects. In the case of substrates that are transparent to the light which is generated, such as GaP, with a $\text{GaAs}_{1-x}\text{P}_x$ epitaxial layer, the doping should also remain below the level at which substantial free carrier absorption occurs.

Growth Techniques

Substrates can be grown by either the Bridgeman or Liquid Encapsulated Czochralski (LEC) technique. The LEC technique is the most widely used. Both techniques are described in detail elsewhere and will be only briefly summarized here.²⁵

The LEC technique for GaAs consists of a crucible containing a molten GaAs solution, into which a single crystal "seed" is dipped. The temperature is carefully controlled so that the molten GaAs slowly freezes on the seed as the seed is rotated and raised out of the molten solution. By properly controlling the temperature, rotation rate, seed lift rate, etc., the seed can be grown into a single crystal weighing several kilograms and having a diameter of typically 2–4 inches for GaAs and for 2–2.5 inches for GaP. At the GaAs and GaP melting points As and P would rapidly evaporate from the growth crucible if they were not contained with a molten boric oxide layer covering the growth solution. This layer is the reason for the name "liquid encapsulated." The seed is dipped through the boric oxide to grow the crystal. The growth chamber must be pressurized to keep the phosphorus and arsenic from bubbling through the boric oxide. The growth pressure for GaP is 80 atm, and for GaAs is 20 atm or less, depending upon the approach for synthesis and growth. The LEC technique, used for GaAs, GaP, and InP, is similar to the Czochralski technique used for silicon, but the silicon process is much simpler since encapsulation is not required and the growth can be done at atmospheric pressure.

The Bridgeman and the gradient-freeze technique, which is a variation of the Bridgeman technique, can also be used to grow compound semiconductors. In this technique the growth solution and seed are contained in a sealed chamber so liquid encapsulation is not required. Growth is accomplished by having a temperature gradient in the solution, with the lowest temperature at the melting point in the vicinity of the seed. Growth can be accomplished by lowering the temperature of the entire chamber (gradient freeze), or by physically moving the growth chamber relative to the furnace (Bridgeman technique) to sweep the temperature gradient through the molten solution. The growing crystal can be in either a vertical or horizontal position.

GaAs for LEDs is commonly grown using either LEC or horizontal gradient freeze, also called "boat grown," but sometimes a vertical Bridgeman approach is used. GaP and InP are almost always grown using LEC but sometimes a vertical Bridgeman approach has also been used.

12.8 EPITAXIAL TECHNOLOGY

Growth Techniques Available

Epitaxial layers are grown using one of several techniques, depending upon the material system. The most common techniques are liquid phase epitaxy (LPE), which is primary used to grow GaAs, GaP, and AlGaAs; and vapor phase epitaxy (VPE), which is used to grow GaAsP. Metalorganic vapor phase epitaxy (MOCVD) is also used to grow AlGaAs, GaInAsP, and AlInGaP. Molecular beam epitaxy (MBE) is used for lasers and high-speed devices but is not used for high-volume commercial LEDs at this time. It has been used to grow blue ZnSe-based lasers and LEDs and could be important in the future. All of these epitaxial techniques have been discussed extensively elsewhere and will be only briefly described here.

LPE

LPE growth consists of a liquid growth solution, generally gallium, which is saturated with the compound to be grown.²⁶ The saturated solution is placed in contact with the substrate at the desired growth temperature, and cooled. As the solution cools, an epitaxial film is grown on the substrate. The technique has the advantage that it is relatively easy to grow high-quality epitaxial layers, and materials containing aluminum (such as AlGaAs) can be readily grown. The disadvantages are that composition control can be difficult. Also, the growth of epitaxial structures involving more than two or three layers, particularly thin layers, can be mechanically complicated since each layer requires a separate growth solution that must be carefully saturated and sequentially brought into contact with the substrate.

One important use of the LPE technique has been for the growth of GaP:ZnO and GaP:N for red and green LEDs, respectively. These devices each consist of two relatively thick layers: an *n*-type layer, followed by the growth of a *p*-type layer. While other growth techniques can be used to grow GaP LEDs, the best results have been obtained using LPE. As a result of a high-volume, low-cost production technology has evolved, which produces more visible LED chips than any other technique. Another major use of LPE is for the growth of GaAs:Si for infrared emitters. LPE is the only technique with which it has been possible to produce the recombination center that gives rise to the 940 nm emission characteristic of this material. The GaAs:Si structures are generally grown from a single growth solution. At high temperatures the silicon is incorporated on Ga sites and the layers are *n*-type. As the solution cools the silicon becomes preferentially incorporated on the As sites and the layer becomes *p*-type.

AlGaAs devices for both visible (red) and IR devices are also generally grown by LPE. AlGaAs devices can also be grown by MOCVD, but LPE has the advantage that thick layers can be more easily grown. This is important for high extraction efficiency (see under "Device Structures" and Fig. 12). In the case of visible devices at 650 nm, the internal quantum efficiency is also higher using LPE than MOCVD. This is not understood, but the result is that virtually all of the visible AlGaAs LEDs are produced using LPE.

VPE

VPE is the other major commercial epitaxial technology for LEDs.^{9,27} VPE consists of a quartz chamber containing the substrate wafers at the appropriate growth temperature. The reactants are transported to the substrates in gaseous form. The technique is mainly used for the growth of GaAsP which, along with GaP, dominates the high-volume visible LED market. In this case HCl is passed over Ga metal to form gallium chlorides, and AsH₃ and PH₃ are used to provide the As and P compounds. Appropriate dopant gases are

12.24 OPTICAL SOURCES

added to achieve the *n*- and *p*-type doping. NH_3 is used to achieve nitrogen doping for the growth of GaAsP:N. The VPE technique has the advantages that it is relatively easy to scale up the growth chamber size so large quantities of material can be grown and layer composition and thickness can be easily controlled by adjusting the flow conditions. A limitation of VPE is that it has not been possible to grow high-quality compounds containing aluminum because the aluminum-bearing reactants attack the quartz chamber resulting in contaminated films. Thus, AlGaAs and AlInGaP, the emerging high-brightness technologies, cannot be grown using this technique.

MOCVD

MOCVD growth, like conventional VPE, uses gases to transpose the reactants to the substrates in a growth chamber.²⁸ However, in this case metallorganic compounds such as trimethylgallium (TMG) are used for one or more of the reactants. A major difference between VPE and MOCVD is that in the case of MOCVD the decomposition of the source gas (e.g., TMG) occurs as a reaction at or near the substrate surface, and the substrate is in the hottest area of the reactor such that the decomposition occurs on the substrate instead of the walls of the growth chamber. The walls of the growth chamber remain relatively cool. This is the key factor that makes MOCVD suitable for the growth of aluminum-bearing compounds which, unlike the VPE situation, do not react significantly with the cooler reactor walls. Thus AlGaAs and AlInGaP can be readily grown with MOCVD, and this technology is widely used for infrared AlGaAs LEDs and lasers, and for the emerging visible AlInGaP laser and LED technology.

MOCVD is also used for the growth of GaN and AlGaN that are candidates for blue emission, and for the growth of II–VI compounds, such as ZnSe, that is also a potential blue emitter. However, at this time the key limitation in obtaining blue ZnSe emitters is the growth of low-resistivity *p*-type ZnSe. For reasons that are not yet understood, low-resistivity *p*-type ZnSe has been grown using MBE only.

MBE

MBE is a high vacuum growth technique in which the reactants are essentially evaporated onto the substrates under very controlled conditions.²⁹ MBE, like MOCVD, can be used to deposit compounds containing aluminum. The growth rates using MBE are generally slower than the other epitaxial techniques, so MBE is most suitable for structures requiring thin layers and precise control of layer thickness. MBE equipment is somewhat more expensive than the equipment used for the other types of epitaxial growth, so it has not been suitable for the high-volume, low-cost production that is required for most types of LEDs. MBE has generally been utilized for lasers and high-speed devices where control of complicated epitaxial structures is critical and where relatively low volumes of devices are required. One advantage that MBE has over the other growth technologies is that the reactants utilized are generally less hazardous. Consequently, MBE equipment is often cheaper and easier to install since there are less safety issues and safety-code restrictions to deal with.

12.9 WAFER PROCESSING

Wafer Processing Overview

Wafer processing of compound semiconductors for LED applications has many of the same general steps used to process silicon integrated-circuit wafers, namely passivation, diffusion, metallization, testing, and die fabrication. The LED device structures are much simpler, so fewer steps are required; but, due to the materials involved, the individual steps are generally different and sometimes more complicated.

Compound semiconductor processing has been described in detail elsewhere, so only a brief summary is discussed here.³⁰ Some types of LED structures require all of the processing steps listed here, but in many cases fewer process steps are required. An example is a GaP or AlGaAs device with a grown *p-n* junction. For these devices no passivation or diffusion is required.

Passivation

Some types of LED structures, particularly multijunction structures, require a passivation layer prior to diffusion, as shown in Fig. 7. This layer must be deposited relatively free of pinholes, be patternable with standard photolithographic techniques, and must block the diffusing element, generally zinc. In the case of silicon, a native oxide is grown which is suitable for most diffusions. Unfortunately, the compound semiconductors do not form a coherent native oxide as readily as silicon. Silicon nitride (Si_3N_4) is the most widely used passivation layer for LEDs. Si_3N_4 is grown by reacting silane (SiH_4) and ammonia at high temperature in a furnace. Si_3N_4 blocks zinc very effectively, and is easily grown, patterned, and removed. Sometimes an SiO_2 layer is used in conjunction with Si_3N_4 for applications such as protecting the surface of the compound semiconductor during high-temperature processing. Silicon oxynitride can also be used instead of or in addition to pure Si_3N_4 . Silicon oxynitride is somewhat more complicated to deposit and control, but can have superior properties, such as a better match of coefficient of expansion, resulting in lower stress at the interface.

Diffusion

Generally, only *p*-type impurities, usually Zn, are diffused in compound semiconductors. *N*-type impurities have prohibitively small diffusion coefficients for most applications. Zn is commonly used because it diffuses rapidly in most materials and because it is nontoxic in contrast to Be, which also diffuses rapidly. Mg is another reasonable *p*-type dopant, but it diffuses more slowly than zinc. Diffusions are generally done in evacuated and sealed ampoules using metallic zinc as source material. A column V element such as As is also generally added to the ampoule to provide an overpressure that helps to prevent decomposition of the semiconductor surface during diffusion. Diffusion conditions typically range from 600–900°C for times ranging from minutes to days, depending upon the material and device involved. Junction depths can range from a fraction of a μm to a more than 10 μm .

Open-tube diffusions have also been employed but have generally been harder to control than the sealed ampoule approach, often because of surface decomposition problems. Open-tube diffusions have the advantage that one does not have to deal with the expense and hazard of sealing, breaking, and replacing quartz ampoules.

A third type of diffusion that has been used is a "semisealed" ampoule approach in which the ampoule can be opened and reused. The diffusion is carried out at atmospheric pressure and the pressure is controlled by having a one-way pressure relief valve on the ampoule.

Contacting

The contacts must make good ohmic contact to both the *p*- and *n*-type semiconductor, and the top surface of the top contact must be well suited for high-speed wire bonding. Generally, multilayer contacts are required to meet these conditions. Evaporation, sputtering, and e-beam deposition are all employed in LED fabrication. The *p*-type

contact generally uses an alloy of either Zn or Be to make the ohmic contact. An Au-Zn alloy is the most common due to the toxicity of Be. The Au-Zn can be covered with a layer of Al or Au to enable high-yield wire bonding. A refractory metal barrier layer may be included between the Au-Zn and top Al or Au layer to prevent intermixing of the two layers and the out-diffusion of Ga, both of which can have a deleterious effect on the bondability of the Al or Au top layer. The *n*-type contact can be similar to the *p*-type contact except that an element which acts as an *n*-type dopant, commonly Ge, is used instead of Zn. An Au-Ge alloy is probably most frequently used to form the *n*-type ohmic contact since it has a suitably low melting point. If the *n*-type contact is the top, or bonded, contact it will be covered by one or more metallic layers to enhance bondability.

Testing

The key parameters that need to be tested are light output, optical rise and fall times, emission wavelength, forward voltage, and leakage current. The equipment used is similar to that used to test other semiconductor devices except that a detector must be added to measure light output. Rise and fall times and wavelength are generally measured on only a sample basis and not for each device on a wafer. In order to test the individual LEDs, the devices must be isolated on the wafer. This occurs automatically for LEDs that are masked and diffused, but if the LEDs are sheet diffused or have a grown junction the top layer must be processed to isolate individual junctions. This can be accomplished by etching or sawing with a dicing saw. Generally, sawing is used, followed by an etch to remove saw damage, because the layers are so thick that etching deep ($>10\text{ }\mu\text{m}$) grooves is required. It is advisable to avoid deep groove etching because undercutting and lateral etching often occur and the process becomes hard to control. In many cases LED junctions are not 100 percent tested. This is particularly true of GaP and AlGaAs red-emitting devices in which the top layer may be $30\text{ }\mu\text{m}$ thick. Wafers of this type can be sampled by "coring" through the top layer of the wafer in one or more places with an ultrasonic tool in order to verify that the wafer is generally satisfactory. Later, when chips are fully processed, chips can be selected from several regions of the wafer and fully tested to determine if the wafer should be used or rejected.

Die Fab

Die fab is the process of separating the wafer into individual dice so they are ready for assembly. Generally, the wafer is first mounted on a piece of expandable tape. Next the wafer is either scribed or sawed to form individual dice. Mechanical diamond scribing or laser scribing were the preferred technologies in the past. Mechanical scribing has zero kerf loss, but the chips tend to have jagged edges and visual inspection is required. Laser scribing provides uniform chips but the molten waste material from between the chips damages neighboring chips, and in the case of full function chips the edges of the junction can be damaged by the laser. As a result of the limitations of scribing, sawing (using a dicing saw with a thin diamond impregnated blade) has become the technology of choice for most LEDs.

The kerf loss for sawing has been reduced to about $40\text{ }\mu\text{m}$ and the chip uniformity is excellent such that a minimum of inspection and testing is required. For most materials a "cleanup" etch is required after sawing to remove work damage at the edges of the chips which can both affect the electrical performance and absorb light. The wafer remains on the expandable tape during the sawing process. After sawing, the tape is expanded so that the chips are separated. The tape is clamped in a ring that keeps it expanded and the chips aligned. In this form the chips are easily individually picked off the tape by the die-attach machine that places the chips in the LED package.

12.10 LED QUALITY AND RELIABILITY

LEDs offer many advantages over other types of light sources. They have long operating life, they operate over a wide temperature range, and they are unaffected by many adverse environmental conditions. LED devices also are mechanically robust, making them suitable for applications where there is high vibration, shock, or acceleration. Excellent quality and reliability are obtainable when an LED product is properly designed, fabricated, packaged, tested, and operated.

Product quality is defined as "fitness for use" in a customer's application. Quality is measured in units of the average number of defective parts per million shipped (i.e., ppm), and is inferred from product sampling and testing. LED product quality is assured by: (1) robust chip and product design, (2) high-quality piece parts, (3) well-controlled fabrication processes, (4) use of statistical process control during manufacturing, (5) careful product testing, and (6) proper handling and storage. Most III-V LEDs are comparable in quality to the best silicon devices manufactured today. Well-designed LED products have total defect levels well below 100 ppm.

Reliability measures the probability that a product will perform its intended function under defined use conditions over the useful life of the product. Probability of survival is characterized by a failure rate, which is calculated by dividing the number of failures by the total number of operating hours (number of products tested per x hours operated). Common measures for reliability are percent failures per 1000 hours (percent/khr) and number failures per 10^9 hours (FITS). LED failure rates typically are better than 0.01 percent per khr at 50°C.

The reliability of an LED product is dependent on the reliability of the LED semiconductor chip and on the robustness of the package into which the chip is placed. Interactions between the chip and package can affect product reliability as well. Aspects of LED packaging and LED chip reliability are discussed in the following paragraphs.

LED Package Reliability

The package into which the LED chip is assembled should provide mechanical stability, electrical connection, and environmental protection. To evaluate package integrity, stress tests such as temperature cycling, thermal and mechanical shock, moisture resistance, and vibration are used to establish the worst-case conditions under which a product can survive. Generally, product data sheets contain relevant information about safe conditions for product application and operation.

Plastic materials are commonly used to package LEDs (see under "LED-Based Products"). Thermal fatigue is a limiting factor in plastic-packaged LEDs. Take the case of the plastic LED lamp shown later in Fig. 21. Because of the different materials used (epoxy plastic, copper lead frame, gold wire, III-V LED, etc.) and the different coefficients of expansion of these materials, temperature changes cause internal stresses. If the package is not well designed and properly assembled, thermal changes can cause cracking, chip-attach failure, or failure of the wire bond (open circuits). Careful design can reduce these problems to negligible levels over wide temperature ranges. Today's high-quality plastic lamps are capable of being cycled from -55 to +100°C for 100 cycles without failure.

Long-term exposure to water vapor can lead to moisture penetration through the plastic, subjecting the chips to humidity. High humidity can cause chip corrosion, plastic delamination, or surface leakage problems. Plastic-packaged LEDs are typically not harmed when used under normal use conditions. Accelerated moisture resistance testing can be used to test the limits of LED packages. Plastic materials have been improved to

the point that LED products can withstand 1000 hr of environmental testing at elevated temperatures and high humidity (i.e., 85°C and 85 percent RH).

The thermal stability of plastic packaging materials is another important parameter. Over normal service conditions, the expansion coefficient of plastic is relatively constant. Above the so-called glass transition temperature T_g the coefficient increases rapidly. Reliable operation of plastic-packaged LEDs generally requires operation at ambient temperatures below T_g . Failures associated with improper soldering operations, such as too high a soldering temperature or for too long a time can cause the package to fail. Excessive storage temperatures also should be avoided. When an LED product is operated, internal ohmic heating occurs; hence, the safe operating maximum temperature is generally somewhat lower than the safe storage temperature.

LED Chip Reliability

The reliability characteristics of the LED chip determine the safe limits of operation of the product. When operated at a given temperature and drive current there is some probability that the LED will fail. In general, LED failure rates can be separated into three time periods: (1) infantile failure, (2) useful life, and (3) the wearout period (see Fig. 18). During the infant mortality period, failures occur due to weak or substandard units. Typically, the failure rate decreases during the infantile period until no weak units remain. During the "useful life" period, the failure rate is relatively low and constant. The number of failures that do occur are random in nature and cannot be eliminated by more testing. The useful life of an LED is a function of the operating temperature and drive conditions. Under normal use conditions, LEDs have useful lives exceeding 100,000 hours. The wearout period is characterized by a rapidly rising failure rate. Generally, wearout for LEDs is not a concern, as the useful life far exceeds the useful life of the product that the LEDs are designed into.

The principle failure mode for LED chips is light-output degradation. In the case of a

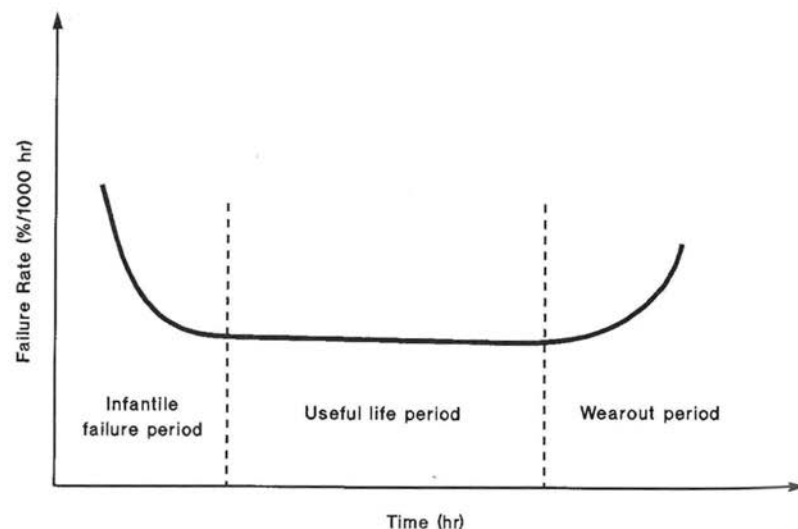


FIGURE 18 Plot of LED failure rate versus time, showing the infantile failure period (decreasing rate), the useful life period (constant, low rate) and the wear-out period (rapidly rising rate).

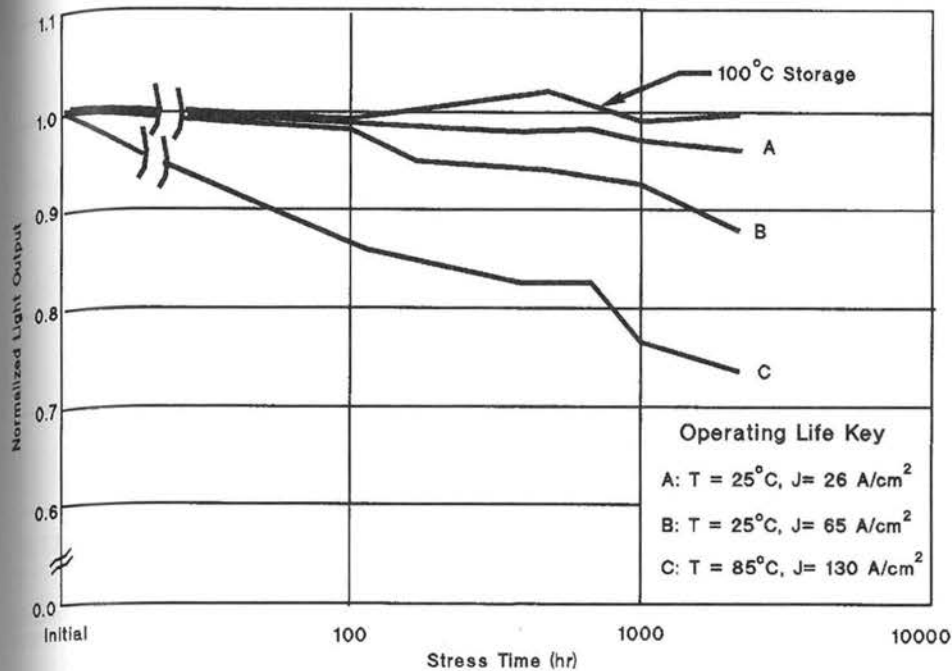


FIGURE 19 Curves of light output versus time for GaAsP indicator lamps stressed under various conditions. Light output is normalized to the initial value. Each curve shows the average degradation of 20 lamps.

visible lamp or display, failure is typically defined as a 50 percent decrease in light output from its initial value, since that is the level where the human eye begins to observe a noticeable change. For infrared emitters or visible LEDs where flux is sensed by a semiconductor detector, failure is commonly defined as a 20 to 30 percent decrease in flux output.

Figure 19 shows degradation curves for a direct bandgap GaAsP LED packaged in a 5-mm plastic lamp.³¹ Current must flow for degradation to occur, as negligible change is observed after 1 khr of 100°C storage. Degradation is a function of the temperature at the p - n junction and the junction current density. As shown in Fig. 19, a larger decline in light output is observed as junction current density and/or temperature at the junction increases. The dependence of degradation on current density J is superlinear, varying as J^x with $1.5 < x < 2.5$. Hence, accelerated-aging tests typically use high currents and temperatures to shorten the time needed to observe LED degradation. The maximum stress level shown in Fig. 19 is 200 percent of the maximum allowable drive current specified in the data sheet.

Light-output degradation in GaAsP LEDs is due to an increase in the nonradiative space-charge recombination current. Total current flowing through the LED is made up of the sum of diffusion current and space-charge recombination current, as shown in the following equations:

$$I_T(V, t) = A(t)e^{qV/kT} + B(t)e^{qV/2kT} \quad (5)$$

where q is electron charge, k is Boltzmann's constant, and T is temperature.³² The first term is diffusion current that produces light output, while the second term is space-charge

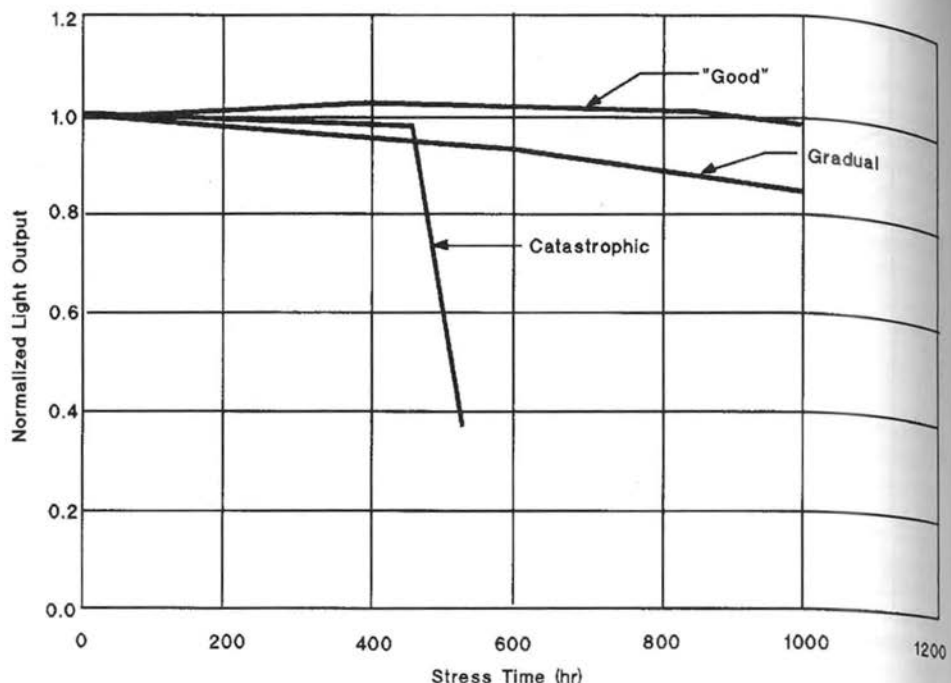


FIGURE 20 Light output degradation of AlGaAs LEDs. Three modes are shown: "good" devices with negligible light-output decrease over time, devices which degrade gradually over time, and "catastrophic" degradation devices where the flux rapidly decreases over a short time period.

recombination that is nonradiative. At fixed I_r , if $B(t)$ increases, then the diffusion term must decrease and, hence, the light output decreases. The reason for the increase in space-charge recombination in GaAsP LEDs is not fully understood.

The degradation characteristics of GaAlAs LEDs differ from those of GaAsP LEDs. Typical curves of normalized light output versus time for GaAlAs LEDs are shown in Fig. 20. "Good" units show negligible decrease in light output when operated under normal service conditions. Gradual degradation may occur, but it is relatively uncommon. The predominate failure mode in GaAlAs LEDs is catastrophic degradation. The light emission decreases very rapidly over a period of less than 100 hr and simultaneously nonradiative regions ("dark-spot" or "dark-line defects") are observed to form. The catastrophic degradation mechanism is described in detail in Refs. 33 and 34. In brief, the dark regions are caused by nonradiative recombination at dislocation networks that grow rapidly from a crystal dislocation located in the light-emitting region of the LED. Network formation depends on carrier recombination, both nonradiative, which creates mobile point defects, and radiative recombination, which enhances the movement of the point defects to the growing network.

Formation of dark-line defects is enhanced by mechanical stress either present in the LED chip or occurring during assembly. Properly designed products reduce such stress by minimizing bending caused by the different coefficients of expansion in the LED, and by stress-free die attach, wire bond, and encapsulation of the LED during assembly.

Failures due to dark-spot or dark-line defects can be effectively screened out by operating the LED at high current and temperature. Units with defects typically fail within the first few hundred hours. GaAlAs LEDs with dark-spot or dark-line defects are screened out by means of visual inspection and/or by eliminating units with large

decreases in light output. GaAlAs LEDs used for fiber-optic applications have small emitting areas (see under "Fiber Optics" discussion and Fig. 14). Due to the high current densities present in such devices, high temperature and current burn-in is used extensively for these types of LEDs.

Another degradation mode in LEDs is change in the reverse breakdown characteristics over time. The reverse characteristics become soft and the breakdown voltage may decrease to a very low value. Several mechanisms have been observed in LEDs. One cause is localized avalanche breakdown due to microplasma formation at points where electric fields are high. Microplasmas have been observed in GaAs and GaAsP LEDs.

Damage or contamination of the edges of an LED chip can cause increased surface leakage and reduced reverse breakdown. Incomplete removal of damage during die separation operations and damage induced during handling and assembly are known to cause reverse breakdown changes. Die-attach materials also can unintentionally contaminate the edges of LEDs. Copper, frequently found in LED packaging materials, can diffuse into the exposed surfaces of the LED, causing excess leakage and, in some cases, light-output degradation.³⁵ Chips whose p - n junction extends to the edges (i.e., Fig. 6) are very susceptible to damage and/or contamination.

12.11 LED BASED PRODUCTS

Indicator Lamps

The simplest LED product is an indicator lamp or its infrared equivalent. The most frequently used lamp is shown in Fig. 21. A LED chip with a typical dimension of

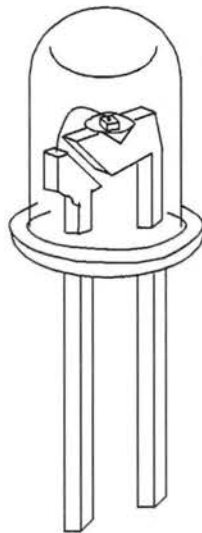


FIGURE 21 Plastic indicator lamp. The LED chip is placed in a reflector coined into the end of one electrode lead. The top of the chip is connected with a gold wire to the second electrode. The electrodes are encapsulated in plastic to form a mechanically robust package.

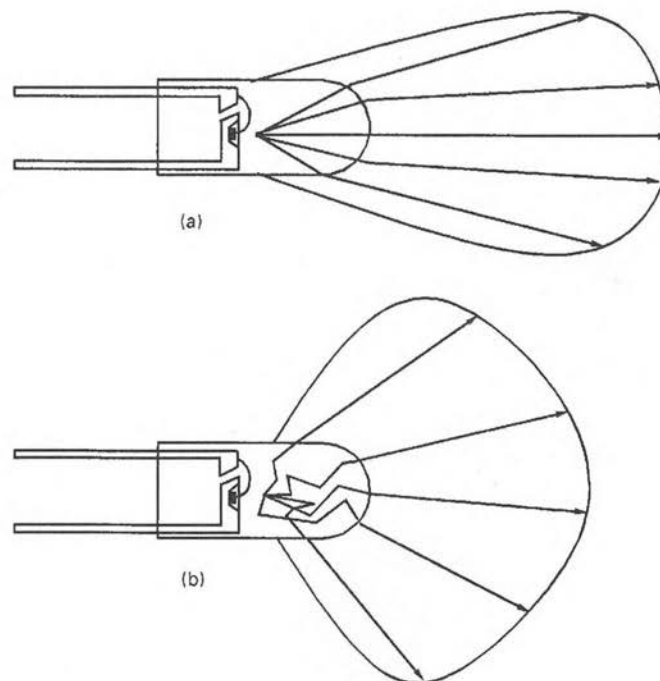


FIGURE 22 Radiation pattern of two types of LED indicator lamps: (a) lamp with clear plastic package with a narrow beam; (b) lamp with a diffusing plastic package (glass powder added) with a broader radiation pattern.

$250 \times 250 \mu\text{m}$ is attached with conductive silver-loaded epoxy into a reflective cavity coined into the end of a silver-plated copper or steel lead frame. The top of the LED chip is connected with a thin $25\text{-}\mu\text{m}$ gold wire to the second terminal of the lead frame. The lead frame subassembly is then embedded in epoxy. The epoxy serves several functions: (1) it holds the assembly together and protects the delicate chip and bond wire; (2) it increases the light extraction from the chip (see under "Light Extraction," discussed earlier); and (3) it determines the spatial light distribution.

There are a large number of variations of the lamp shown in Fig. 21. Besides the obvious variation of source wavelength, there are variations of size, shape, radiation pattern, etc. The cross section of the plastic body ranges from 2–10 mm. The radiation pattern is affected by three factors: the shape of the dome, the relative position of the chip/reflector combination, and by the presence of a diffusant in the epoxy. Figure 22a shows the radiation pattern of a lamp using clear plastic as an encapsulant. The rays emanating from chip and reflector are collimated into a narrow beam. For many indicator lamps a broader viewing angle is desired such as that shown in the radiation pattern of Fig. 22b. This effect is achieved by adding a diffusant, such as glass powder, to the clear plastic. Another variation is shape. Common shapes are round, square, rectangular, or triangular.

Another variation is achieved by placing two different chips into the reflector cup. For instance, a lamp with green and red chips connected to the second post in an antiparallel fashion can operate as a red indicator if the reflector post is biased positive, and as a green indicator if it is biased negative. Rapidly switching between the two polarities one can

achieve any color between the two basic colors, i.e., yellow or orange, depending on current and duty cycle.

A number of chips can be combined in a single package to illuminate a rectangular area. These so-called annunciator assemblies range in size from one to several cm. They typically use 4 chips per cm^2 . By placing an aperture-limiting symbol or telltale in front of the lit area, these structures are cost-effective means to display a fixed message, such as warning lights in an automotive dashboard.

Numeric Displays

Numeric displays are usually made up of a nearly rectangular arrangement of seven elongated segments in a figure-eight pattern. Selectively switching these segments generates all ten digits from 0–9. Often decimal point, colon, comma, and other symbols are added.

There are two main types of LED numeric displays: (1) monolithic displays and (2) stretched segment displays. All monolithic displays are based on GaAs-GaAsP technology. Seven elongated p -doped regions and a decimal point are diffused into an n -type epitaxial layer of a single or monolithic chip (see Fig. 8). Electrically, this is a structure with eight anodes and one common cathode. This monolithic approach is relatively expensive. For arms-length viewing, a character height of 3–5 mm is required. Adding space for bonding pads, decimal point, and edge separation, such a display consumes around 10 mm^2 of expensive semiconductor material per digit. One way to reduce material and power consumption is optical magnification. Viewing-angle limitation and distortion limit the magnification M to $M \leq 2.0$. Power consumption is reduced by M^2 —an important feature for battery-powered applications.

For digits $>5 \text{ mm}$, a stretched segment display is most cost effective. The design of Fig. 23 utilizes a $250 \times 250 \text{ }\mu\text{m}$ chip to generate a segment with dimensions of up to $8 \times 2 \text{ mm}$ for a 20-mm digit height. This corresponds to a real magnification of $M^2 = 256$, or an equivalent linear magnification of $M = 16$. This magnification is achieved without a reduction in viewing angle by using scattering optics. An LED chip is placed at the bottom

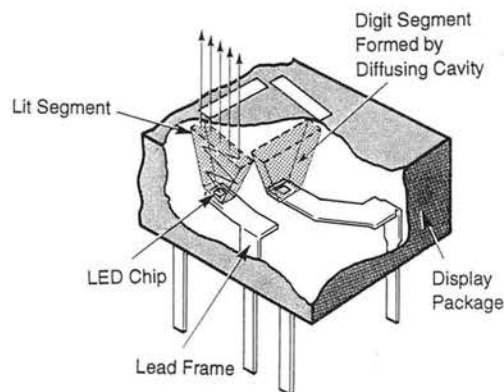


FIGURE 23 Cutaway of a seven-segment numeric LED display, showing how light from a small LED chip is stretched to a large character segment using a diffusing cavity. Characters 0–9 are created by turning on appropriate combinations of segments.

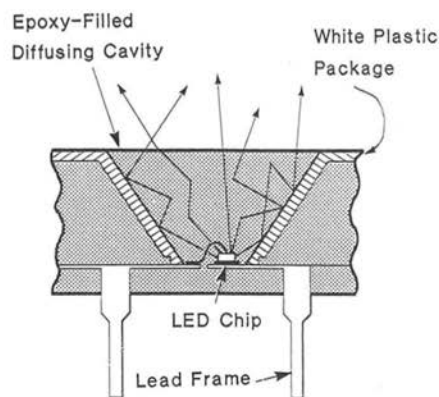


FIGURE 24 Cross section through one segment of the seven-segment numeric display shown in Fig. 23. A LED chip is placed at the bottom of a highly reflective, white plastic cavity which is filled with clear plastic containing a diffusant. Light is scattered within the cavity to produce uniform emission at the segment surface.

of a cavity having the desired rectangular exit shape and the cavity is filled with a diffusing plastic material (see Fig. 24). The side and bottom surfaces of the cavity are made as reflective as possible. Highly reflective white surfaces are typically used. A good white plastic surface measured in air may have a reflectivity of 94 percent compared with 98 percent for Ag and 91 percent for Al. Ag and Al achieve this reflectivity only if evaporated on a specularly smooth surface. Measured in plastic, the reflectivity of the white surface increases from 94 to 98 percent, the Ag surface remains at 98 percent, while the Al surface decreases to 86 percent. Both metallic surfaces have substantially lower reflectivities if they are evaporated onto a nonspecular surface, or if they are deposited by plating. Practically all numeric LED displays above 5-mm character heights are made using white cavity walls and diffusing epoxy within the cavity.

This case of magnification by scattering does not result in a power saving as in the case of magnification of monolithic displays. Since there is practically no reduction in emission angle, the law of energy conservation requires an increased light flux from the chip that equals the area magnification plus reflection losses in the cavity.

Alphanumeric Displays

There are two ways LEDs are used to display alphanumeric information: either by using more than seven elongated segments, i.e., 14, 16, or 24; or by using an array of LED chips in a 5×7 dot matrix format. The multiple segment products are similar in design to the monolithic or stretched segment numeric displays described above.

In the case of small monolithic characters, the number of input terminals quickly exceeds conventional pin spacing. These products are usually clusters of 4 to 16 characters combined with a decoder/driver integrated circuit within the same package. To reduce cost and power, some modest optical magnification is usually used. The segmented displays are usually larger, i.e., 12–25 mm and limited to 14 segments per character. At this size, there is no pin density constraint and the decoder is usually placed outside the package.

The most frequently used alphanumeric LED display is based on a 5×7 matrix per character. For small characters in the range of 5–8 mm, the LED chips are directly viewed

and pin density limitations require an on-board decoder/driver IC. Products are offered as end-stackable clusters of 4, 8, and 16 characters.

For larger displays, the LED chip is magnified by the same optical scattering technique described above for numerics. Exit apertures per pixel have a diameter of 2–5 mm. Products are offered as 5×7 single characters or end-stackable 20×20 tiles for large message- or graphics-display panels. At this size, pin density is not a limitation.

Optocouplers

An optocoupler is a device where signal input and signal output have no galvanic connection. It is mainly used in applications as the interface between the line voltage side of a system and the low-voltage circuit functions, or in systems where the separate ground connection of interconnected subsystems causes magnetic coupling in the galvanic loop between signal and ground connections. By interrupting the galvanic loop with an optical signal path, many sources of signal interference are eliminated.

The oldest optocouplers consist of an IR LED and a photodetector facing each other in an insulating tube. The second generation utilized the so-called dual-in-line package widely used by logic ICs. In this package an IR emitter and a phototransistor are mounted face-to-face on two separate lead frames. The center of the package between emitter and detector is filled with a clear insulating material. The subassembly is then molded in opaque plastic to shield external light and to mechanically stabilize the assembly (Fig 25). The second generation optocouplers have limited speed performance for two reasons: (1) the slow response time of the GaAs:Si LED and (2) the slow response of the phototransistor detector because of the high collector-base capacitance.

The third generation of optocouplers overcomes the speed limitation. It uses an integrated photodetector and a decoupled gain element. Integration limits the thickness of the effective detection region in silicon to $5\text{--}7\text{ }\mu\text{m}$. This thickness range forces a shift of the source to wavelengths shorter than the 940 nm sources used in the second-generation couplers. Third-generation couplers use GaAsP (700 nm) or AlGaAs emitters (880 nm).

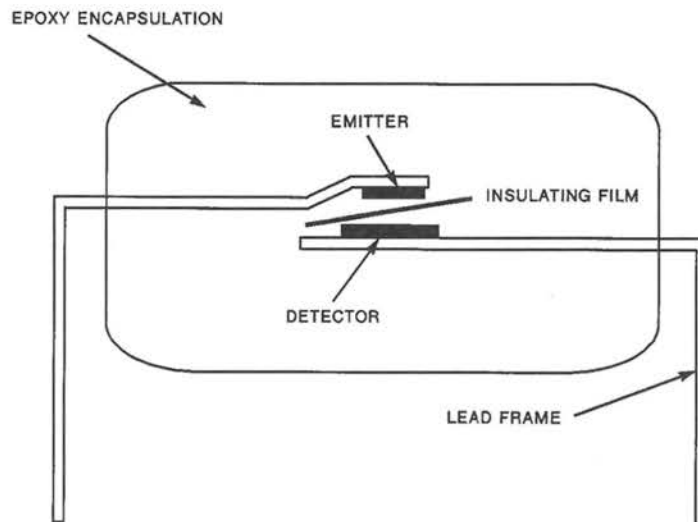


FIGURE 25 Optocoupler consisting of face-to-face emitter and detector chips. An insulating film is placed between the chips to increase the ability of the optocoupler to withstand high voltages between input and output electrodes.

Within the last decade, the optocoupler product family has seen further proliferation by adding features on the input or output side of the coupler. One proliferation resulted in couplers behaving like logic gates. Another variation used MOS FET devices on the output side, eliminating the offset voltages of bipolar devices. These couplers are comparable to the performance of conventional relays and are classified as solid-state relays. Other types use a CMOS input driver and CMOS output circuitry to achieve data transfer rates of 50 Mb/s and CMOS interface compatibility.

Fiber Optics

LEDs are the primary light source used in fiber-optic links for speeds up to 200 Mb/s and distances up to 2 km. For higher-speed and longer distances, diode lasers are the preferred source.

For fiber-optic applications, LEDs have to meet a number of requirements that go well beyond the requirements for lamps and displays. The major issues are: minimum and maximum flux coupled into the fiber, optical rise and fall times, source diameter, and wavelength. Analysis of flux budget, speed, fiber dispersion, wavelength, and maximum distance is quite complex and goes far beyond the scope of this work. A simplified discussion for the popular standard, Fiber Distributed Data Interface (FDDI) will highlight the issues. For a detailed discussion, the reader is referred to Ref. 36.

Flux budget. The minimum flux that has to be coupled into the fiber is determined by receiver sensitivity (-31 dBm), fiber attenuation over the 2-km maximum distance (3 dB), connector and coupler losses (5 dB), and miscellaneous penalties for detector response variations, bandwidth limitations, jitter, etc. (3 dB). With this flux budget of 11 dB, the minimum coupled power has to be -20 dBm. Another flux constraint arises from the fact that the receiver can only handle a maximum level of power before saturation (-14 dBm). These two specifications bracket the power level coupled into the fiber at a minimum of -20 dBm (maximum fiber and connector losses) and at a maximum of -14 dBm (no fiber or connector losses for the case of very short fiber).

Speed. The transmitter speed or baud rate directly translates into a maximum rise and fall time of the LED. The 125 Mdb FDDI specifications call for a maximum rise and fall time of 3.5 ns. As a rule of thumb, the sum of rise and fall time should be a little shorter than the inverse baud period (8 ns for 125 Mbd).

Source Diameter and Fiber Alignment. Efficient coupling of the LED to the fiber requires a source diameter that is equal to or preferably smaller than the diameter of the fiber core. For sources smaller than the core diameter, a lens between source and fiber can magnify the source to a diameter equal to or larger than the core diameter. The magnification has two benefits: (1) It increases the coupling efficiency between source and fiber. The improvement is limited to the ratio of source area to core cross-sectional area. (2) It increases the apparent spot size to a diameter larger than the fiber core. This effect relaxes the alignment tolerance between source and core and results in substantially reduced assembly and connector costs.

Wavelength. The LEDs used in fiber-optic applications operate at three narrowly defined wavelength bands 650, 820–870, and 1300 nm, as determined by optical fiber transmission characteristics.

650 nm. This band is defined by an absorption window in acrylic plastic fiber. It is a very narrow window between two C-H resonances of the polymer material. The bottom of the window has an absorption of approximately 0.17 dB/m. However, the effective absorption

is in the 0.3–0.4 dB/m range because the LED linewidth is comparable to the width of the absorption window and the LED wavelength changes with temperature. The 650-nm LEDs use either $\text{GaAs}_{1-x}\text{P}_x$ with $x = 0.4$ or $\text{GaAl}_x\text{As}_{1-x}$ with $x = 0.38$. Quantum efficiencies are at 0.2 and 1.5 percent, respectively. Maximum link length is in the range of 20–100 m, depending on source efficiency, detector sensitivity, speed, and temperature range.

820–870 nm. This window was chosen for several reasons. GaAlAs emitters (see Fig. 14) and Si detectors are readily available at this wavelength. Early fibers had an absorption peak from water contamination at approximately 870 nm. As fiber technology improved, the absorption peak was eliminated and the wavelength of choice moved from 820- to the 850–870-nm range. Fiber attenuation at 850 nm is typically 3 dB/km. Maximum link length in the 500–2000 m range, depending on data rate. In the 850–870-nm window the maximum link length is limited by chromatic dispersion. GaAlAs emitters have a half-power linewidth of approximately 35 nm. The velocity of light in the fiber is determined by the index of refraction of the fiber core. The index is wavelength-dependent resulting in dispersion of the light pulse. This dispersion grows with distance. The compounded effect of LED linewidth and fiber dispersion is a constant distance-speed product. For a typical multimode fiber and GaAlAs LED combination, this product is in the range of 100 Mbd-km.³⁶

1300 nm. At this wavelength, the index of refraction as a function of wavelength reaches a minimum. At this minimum, the velocity of light is practically independent of wavelength, and chromatic dispersion is nearly eliminated. The distance-speed limitation is caused by modal dispersion. Modal dispersion can be envisioned as a different path length for rays of different entrance angles into the fiber. A ray going down the middle of the core will have a shorter path than a ray entering the fiber at the maximum acceptance angle undergoing many bends as it travels down the fiber. The resulting modal dispersion limits multimode fibers to distance bandwidth products of approximately 500 Mbd-km. LED sources used at this wavelength are GaInAsP emitters, as shown in Fig. 15. At this wavelength, fiber attenuation is typically <1.0 dB/km and maximum link length is in the range of 500–5000 m, depending on data rate and flux budget.

Sensors

LED/detector combinations are used in a wide range of sensor applications. They can be grouped into three classes: transmissive, reflective, and scattering sensors.

Transmissive Sensors. The most widely used transmissive sensor is the slot interrupter. A U-shaped plastic holder aligns an emitter and detector face-to-face. It is used widely for such applications as sensing the presence or absence of paper in printers, end-of-tape in tape-recorders, erase/overwrite protection on floppy disks and many other applications where the presence or absence of an opaque obstruction in the light path determines a system response.

A widely used slot interrupter is a two- or three-channel optical encoder. A pattern of opaque and transmissive sections moves in front of a fixed pattern with the same spatial frequency. Two optical channels positioned such that they are 90° out-of-phase to each other with regard to the pattern allow the measurement of both distance (number of transmissive/opaque sequences) and direction (phase of channel A with regard to channel B). Such encoders are widely used in industrial control applications, paper motion in printers, pen movement in plotters, scales, motor rotation, etc. A third channel is often used to detect an index pulse per revolution to obtain a quasi-absolute reference.

Reflective Sensors. In a reflective sensor an LED, detector, and associated optical

elements are positioned such that the detector senses a reflection when a reflective surface (specular or diffuse) is positioned within a narrow sensing range. A black surface or nonaligned specular surface or the absence of any reflective surface can be discriminated from a white surface or properly aligned specular surface. Applications include bar-code reading (black or white surface), object-counting on a conveyer belt (presence or no presence of a reflecting surface), and many others. Many transmissive sensor applications can be replaced by using reflective sensors and visa versa. The choice is usually determined by the optical properties of the sensing media or by cost. An emerging application for reflective sensors is blood gas analysis. The concentration of O_2 or CO_2 in blood can be determined by absorption at two different LED wavelengths, i.e., red and infrared.

Scattering Sensors. One design of smoke detectors is based on light scattering. The LED light beam and the detector path are crossed. In the absence of smoke, no light from the LED can reach the detector. In the presence of smoke, light is scattered into the detector.

12.12 REFERENCES

1. A. A. Bergh and P. J. Dean, "Light-Emitting Diodes," *Proc. IEEE*, vol. 60, 1972, pp. 156-224.
2. K. Gillessen and W. Shairer, *LEDs—An Introduction*, Prentice-Hall, 1987.
3. M. G. Craford, "Properties and Electroluminescence of the GaAsP Ternary System," *Progress in Solid State Chemistry*, vol. 8, 1973, pp. 127-165.
4. A. H. Herzog, W. O. Groves, and M. G. Craford, "Electroluminescence of Diffused $GaAs_{1-x}P_x$ Diodes with Low Donor Concentrations," *J. Appl. Phys.*, vol. 40, 1969, pp. 1830-1838.
5. R. A. Faulkner, "Toward a Theory of Isoelectronic Impurities in Semiconductors," *Phys. Rev.*, vol. 175, 1968, pp. 991-1009.
6. W. O. Groves, A. J. Herzog, and M. G. Craford, "The Effect of Nitrogen Doping on $GaAs_{1-x}P_x$ Electroluminescent Diodes," *Appl. Phys. Lett.*, vol. 19, 1971, pp. 184-186.
7. M. G. Craford, R. W. Shaw, W. O. Groves, and A. H. Herzog, "Radiative Recombination Mechanisms in GaAsP Diodes with and without Nitrogen Doping," *J. Appl. Physics*, vol. 43, 1972, pp. 4075-4083.
8. M. G. Craford, D. L. Keune, W. O. Groves, and A. H. Herzog, "The Luminescent Properties of Nitrogen Doped GaAsP Light Emitting Diodes," *J. Electron. Mats.*, vol. 2, 1973, pp. 137-158.
9. M. G. Craford and W. O. Groves, "Vapor Phase Epitaxial Materials for LED Applications," *Proc. IEEE*, vol. 61, 1973, pp. 862-880.
10. J. C. Campbell, N. Holonyak Jr., M. G. Craford, and D. L. Keune, "Band Structure Enhancement and Optimization of Radiative Recombination in $GaAs_{1-x}P_x:N$ (and $In_{1-x}Ga_xP:N$)," *J. Appl. Phys.*, vol. 45, 1974, pp. 4543-4553.
11. R. A. Logan, H. G. White, and W. Wiegmann, "Efficient Green Electroluminescence in Nitrogen-Doped GaP $p-n$ Junctions," *Appl. Phys. Lett.*, vol. 13, 1968, p. 139.
12. A. A. Bergh and J. A. Copeland, "Optical Sources for Fiber Transmission Systems," *Proc. IEEE*, vol. 68, 1980, pp. 1240-1247.
13. M. G. Craford, "Recent Developments in Light-Emitting Diode Technology," *IEEE Trans. Electron Devices*, vol. 24, 1977, pp. 935-943.
14. H. Nather, V. Nitsche, and W. Schairer, "High Resolution Printing Capability of LED-Based Print Heads," *Proceedings SPIE*, 1988, pp. 396-404.
15. M. G. Craford, "Light-Emitting Diode Displays," in *Flat-Panel Displays and CRTs*, L. E. Tannas Jr. (ed.), Van Nostrand Reinhold, 1985, pp. 289-331.
16. L. W. Cook, M. D. Camras, S. L. Rudaz, and F. M. Steranka, "High Efficiency 650 nm Aluminum Gallium Arsenide Light Emitting Diodes," *Proceedings of the 14th International Symposium on GaAs and Related Compounds*, Institute of Physics, Bristol, 1988, pp. 777-780.

17. J. M. Dallesasse, D. W. Nam, D. G. Deppe, N. Holonyak Jr., R. M. Fletcher, C. P. Kuo, T. D. Osentowski, and M. G. Craford, "Short-Wavelength ($<6400\text{ \AA}$) Room Temperature Continuous Operation of p - n InAlGaP Quantum Well Lasers," *Appl. Phys. Lett.*, vol. 19, 1988, pp. 1826-1828.
18. C. P. Kuo, R. M. Fletcher, T. D. Osentowski, M. C. Lardizabel, and M. G. Craford, "High Performance AlGaInP Visible Light-Emitting Diodes," *Appl. Phys. Lett.*, vol. 57, 1990, pp. 2937-2939.
19. H. Sugawara, M. Ishikawa, and G. Hatakoshi, "High-efficiency InGaAlP/GaAs Visible Light-Emitting Diodes," *Appl. Phys. Lett.*, vol. 58, 1991, 1010-1012.
20. R. M. Fletcher, C. P. Kuo, T. D. Osentowski, K. H. Huang, and M. G. Craford, "The Growth and Properties of High Performance AlGaInP Emitters Using a Lattice Mismatched GaP Window Layer," *Jour. Elec. Mats.*, vol. 20, 1991, pp. 1125-1130.
21. K. H. Huang, J. G. Yu, C. P. Kuo, R. M. Fletcher, T. D. Osentowski, L. J. Stinson, A. S. H. Liao, and M. G. Craford, "Twofold Efficiency Improvement in High Performance AlGaInP Light-Emitting Diodes in the 555-620 nm Spectral Region Using a Thick GaP Window Layer," *Appl. Phys. Lett.*, vol. 61, 1992, pp. 1045-1047.
22. M. A. Haase, J. Qiv, J. M. DePoydt, and H. Cheng, "Blue-green Laser Diodes," *Appl. Phys. Lett.*, vol. 58, 1991, pp. 1272-1275.
23. J. Jeon, J. Ding, A. V. Normikko, W. Xie, M. Kobayashi, and R. L. Gunshore, "ZnSe Based Multilayer p/n Junctions as Efficient Light Emitting Diodes for Display Applications," *Appl. Phys. Lett.*, vol. 60, 1992, pp. 892-894.
24. S. Nakamura, M. Senoh, and T. Mukai, "Highly P -typed Mg-doped GaN Films Grown with GaN Buffer Layers," *Jap Jour. of Appl. Phys.*, vol. 30, 1991, pp. L1701-L1711.
25. A. G. Fischer, "Methods of Growing Crystals under Pressure," in *Crystal Growth*, B. R. Pamplin (ed.), Pergamon Press, 1975, pp. 521-555.
26. R. L. Moon, "Liquid Phase Epitaxy," in *Crystal Growth*, 2d ed., B. R. Pamplin (ed.), Pergamon Press, 1980.
27. J. W. Burd, "A Multi-Wafer Growth System for the Epitaxial Deposition of GaAs and $\text{GaAs}_{1-x}\text{P}_x$," *Trans. Met. Soc. AIME*, 1969, pp. 571-576.
28. G. B. Stringfellow, *Organometallic Vapor Phase Epitaxial Growth of III-V Semiconductor: Theory and Practice*, Academic Press, 1989.
29. E. C. H. Parker (ed.), *Technology and Physics of Molecular Beam Epitaxy*, Plenum Press, 1985.
30. S. K. Ghandhi, *VLSI Fabrication Principles*, John Wiley and Sons, 1982.
31. Hewlett-Packard, *Optoelectronics/Fiber-Optics Application Manual*, 2d ed., McGraw-Hill, 1981, p. 82.
32. A. S. Grove, *Physics and Technology of Semiconductor Devices*, sec. 6.6, John Wiley, 1967.
33. O. Veda, *Material Research Society Symposium Proc.*, vol. 184, 1991, p. 125.
34. M. Fukuda, *Reliability and Degradation of Semiconductor Lasers and LEDs*, Artech House, 1991.
35. A. A. Bergh, "Bulk Degradation of GaP Red LEDs," *IEEE Trans. Electron Devices*, vol. 18, 1971, pp. 166-170.
36. D. C. Hanson, "Progress in Fiber Optic LAN and MAN Standards," *IEEE LCS Magazine*, vol. 1, 1990, pp. 17-25.

CHAPTER 19

THERMAL DETECTORS

William L. Wolfe

*Professor, Optical Sciences Center
The University of Arizona
Tucson, Arizona*

Paul W. Kruse

*Consultant
Edina, Minnesota*

19.1 GLOSSARY

<i>DTGS</i>	deuterated triglycine sulfate
<i>p</i>	pyroelectric coefficient
<i>R_e</i>	electrical resistance
<i>R_{th}</i>	thermal resistance
$\Re$	responsivity
<i>S</i>	Seebeck coefficient
<i>TGS</i>	triglycine sulfate
<i>Z</i>	figure of merit
τ_e	electrical time constant
τ_{th}	thermal time constant

19.2 THERMAL DETECTOR ELEMENTS¹

Introduction

Thermal detectors (transducers) of optical radiation are generally considered to be those devices that absorb the radiation, increase their own temperature, and provide a resultant electrical signal. There are several types, divided according to the physical mechanism that converts the temperature change to a resultant electrical one. The oldest are bolometers and thermocouples. The bolometer changes its electrical resistance as a result of the temperature increase; the thermocouple changes its contact potential difference. There are several different types of bolometers, including thermistor, semiconducting, superconducting, carbon, and metallic. They may also be subdivided according to whether they operate

at room or cryogenic temperature. Thermocouples vary according to the materials that are joined, and are sometimes connected in series to generate thermopiles. Pyroelectric detectors make use of the property of a change in the internal polarization as a function of the change in temperature, the pyroelectric effect. Golay cells and certain variations make use of the expansion of a gas with temperature. All of these detectors are governed by the fundamental equation of heat absorption in the material. Many reviews and two books of collected reprints² provide additional information.

Thermal Circuit Theory

In the absence of joulean heating of the detector element, the spectrum of the temperature difference $d\Delta\tilde{T}$ is given in terms of the spectrum of the absorbed power $\tilde{P}$ (the power is P)

$$d\Delta\tilde{T} = \frac{\epsilon\tilde{P}}{G(1+i\omega\tau)} \quad (1)$$

where G is the thermal conductance, given by the product of the thermal conductivity times the cross-sectional area of the path to the heat sink and divided by the length of the path to that heat sink. The time constant τ is the product of the thermal resistance and the heat capacitance. The thermal resistance is the reciprocal of the thermal conductance, while the thermal capacitance is the thermal capacity times the mass of the detector. In the absence of Joulean heating, this is a simple, single time constant thermal circuit, for which the change in temperature is given by

$$d\tilde{T} = \frac{\epsilon\tilde{P}}{G(1+i\omega\tau)} \quad (2)$$

The absorbed power is equal to the incident power times the absorptance α of the material:

$$\tilde{P} = \epsilon\tilde{P}_i \quad (3)$$

The absorptance α is usually written as ϵ (which is legitimate according to Kirchhoff's law) since α is also used for the relative temperature coefficient of resistance (some writers use η).

$$\alpha = \frac{1}{R} \frac{dR}{dT} \quad (4)$$

As radiation is absorbed, part of the heat is conducted to the sink. Some of it gives rise to an increase in temperature. Some is reradiated, but this is usually quite small and is ignored here. The dc responsivity of a thermal detector is proportional to the emissivity and to the thermal resistance. The greater proportion of radiation that is absorbed, the greater will be the responsivity. The less heat that is conducted to the sink, the greater will be the temperature rise. The time constant is a true thermal time constant, the product of thermal resistance and capacitance. The greater the heat capacitance, the more heat necessary for a given temperature increase, and the less heat conducted to the sink, the more available for temperature increase. A high absorptance is accomplished by the use

of a black coating, and a sufficient amount of it. Thus, there is a direct conflict between high speed and high responsivity.

The Ideal Thermal Detector³⁻⁶

The ideal thermal detector has a noise that is associated only with the thermal fluctuations of the heat loss to the heat sink, and this coupling is purely radiative. Then the noise equivalent power (NEP) is given by

$$\text{NEP} = \sqrt{16A\sigma kT^5/\epsilon} \quad (5)$$

where it is assumed that the detector is irradiated by a hemisphere of blackbody radiation at the same temperature T as the detector. The corresponding specific detectivity, assuming that the signal varies as the area and the noise as its square root, is

$$D^* = \frac{\epsilon^{1/2}}{4\sqrt{\sigma kT^5}} \quad (6)$$

where the detector and background are at the same temperature.

For circumstances in which the detector is in a cooled chamber, the total radiation from the sources at various temperatures must be calculated. Figure 1 shows the specific detectivity of a background limited ideal thermal detector as a function of the temperature of the surround.

No detector is ideal, and every one will be limited by the signal loss due to incomplete absorption at the surface and any transmission losses by the optical system that puts the radiation on the detector. The detector will also have noise that arises from its conductive coupling to the heat sink, and probably Johnson noise as well. The conductive mean square power fluctuation is given by

$$\langle P^2 \rangle = 4kT^2G \quad (7)$$

The Johnson noise power density is $4kT$. Therefore, the total mean square power fluctuation is given by

$$\langle P^2 \rangle = 4kT[GT + 4\epsilon A\sigma T^4 + 1] \quad (8)$$

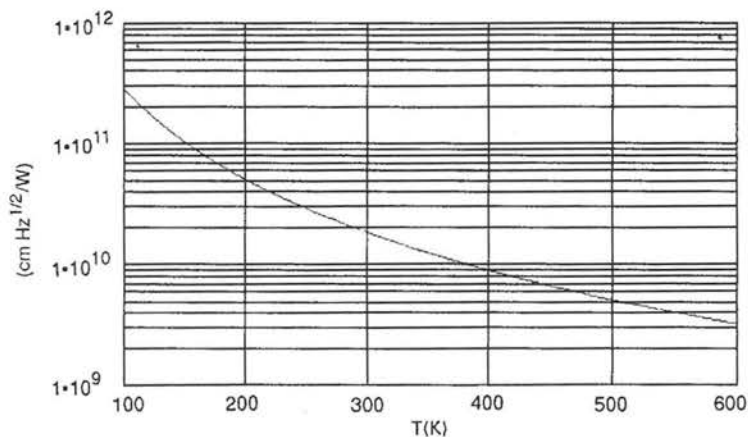


FIGURE 1 Theoretical specific detectivity for ideal thermal detectors.

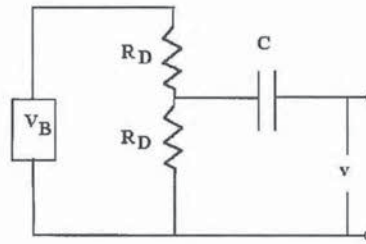


FIGURE 2 Balanced voltage divider circuit for a thermal detector.

Bolometers

Most single-element bolometers are connected in a voltage divider network, as shown in Fig. 2. A stable voltage supply is used to develop a current and consequent voltage drop across the two resistors. One is the detector, while the other should be a matching element to eliminate signals arising from a change in the ambient temperature. It should match the detector in both resistance and in the temperature coefficient of resistance. Usually another, but blinded, detector is used. The expression for power conservation is

$$\begin{aligned}
 C \frac{d \Delta T}{dt} + G \Delta T &= \frac{d(i^2 R)}{dT} \Delta T + P \\
 C \frac{d \Delta T}{dt} + G \Delta T &= \frac{V^2 (R_1 - R)}{(R_1 + R)^3} \frac{dR}{dT} \Delta T + P \\
 C \frac{d \Delta T}{dt} + \left[G - \frac{V^2 R \alpha}{(R_1 + R)^2 (R_1 + R)} \right] \Delta T &= P
 \end{aligned} \tag{9}$$

The solution to this is a transient that has an RC time constant, where the R is the reciprocal of the bracketed term, and C is the thermal capacitance, and the same steady-state term given above. The transient decays as long as G is greater than the rest of the bracket, but the detector burns up if not. This is still another reason for matching the resistances. The dc responsivity is a function of the construction parameters, including the path to the sink, the bias voltage, and the relative change of resistance with temperature. The different bolometers are divided according to how their resistances change with temperature. (R_1 and R represent slightly different values of R_D .)

Metal bolometers. These have a linear change in resistance with temperature that may be expressed as

$$R = R_0[1 + \gamma(T - T_0)] \tag{10}$$

Therefore the thermal coefficient is

$$\alpha = \frac{\gamma}{1 + \gamma(T - T_0)} \tag{11}$$

This coefficient always decreases with temperature, and burnout does not occur. The

coefficient is approximately equal to the inverse of the temperature, and is therefore never very high.

Semiconductor bolometers. These have an exponential change of resistance with temperature, given by

$$R = R_0 e^{\beta/T} \quad (12)$$

so that

$$\alpha = -\beta/T^2 \quad (13)$$

The value of β depends upon the particular material. These detectors can burn out. Two basic types exist: (1) those that are used at low temperatures and (2) those that are used at about room temperature.

The most-used low-temperature bolometer⁷ is germanium in a bath of liquid helium. Pure germanium is transparent in the infrared, but with enough compensated doping it becomes a good conductor with a high-temperature coefficient of resistance.⁸ Typical concentrations are about 10^{16} cm^{-3} of gallium and 10^{15} of indium. Even these are not sufficient at wavelengths shorter than $10 \mu\text{m}$ since the free-carrier absorption is proportional to wavelength. In such a case a black coating is sometimes used. Improvements have been made since Low's first work.⁹⁻¹¹

Superconducting bolometers. These make use of the extremely large thermal coefficient of resistance at the transition temperature.¹²⁻¹⁴ Originally they needed to be controlled very carefully, or a small change in ambient conditions (on the order of 0.01 K) could cause an apparent signal of appreciable magnitude. A more recent version¹⁵ incorporates an evaporated thin film on an anodized aluminum block that is coupled to a helium bath by a brass rod. The detector has a time constant of about $3 \mu\text{s}$ due to this high thermal conductance and a good NEP of about $10^{-13} \text{ WHz}^{-1/2}$. It still must be controlled to about 10^{-5} K , and this is accomplished with a heater current and control circuit.

Recently developed materials not only have high-temperature transition points but also have more gradual transitions, and provide a better compromise between good responsivity and the requirement for exquisite control.¹⁶

Carbon Bolometers. These are a form of semiconductor bolometers that have been largely superseded by germanium bolometers. They are made of small slabs of carbon resistor material, connected to a metal heat sink by way of a thin mylar film. Although their responsivities are comparable to germanium bolometers, their noise is several orders of magnitude higher.¹⁷

Thermocouples and Thermopiles^{18,19}

A *thermocouple* is made by simply joining two dissimilar conductors. A good pair has a large relative Seebeck coefficient and gives rise to a potential difference. The materials also have large electrical conductivities and small thermal ones, so there is little voltage drop across the length and a small thermal gradient. Although there are many different couples (many are not even used for radiation detection), those most often used for this application are bismuth telluride, copper, and constantan. The expression for the responsivity is given in terms of the relative Seebeck coefficient S_{12} (the difference in the

19.6 OPTICAL DETECTORS

voltage change with temperature between the two materials) and the expression derived above for the thermal circuit

$$\Re = \frac{S\epsilon}{G(1 + i\omega\tau)} \quad (14)$$

Good materials are those that have a large Seebeck coefficient, a high electrical conductivity, and a small thermal conductivity, and the figure of merit is often defined as

$$Z_{12} = \frac{S_{12}^2}{[\sqrt{G_1/\sigma_1} + \sqrt{G_2/\sigma_2}]^2} \quad (15)$$

Thermopiles are arrays of thermocouples connected in series. They are manufactured in two ways. Some are carefully wound wires with junctions aligned in the desired pattern, while others are evaporated with the pattern determined by masking operations. Most of the "bulk" thermopiles are wrapped on appropriate mandrels to obtain rigidity. Both kinds are obtainable in a variety of sizes and patterns that correspond to such things as spectrometer slits, centering annuli, and staggered arrays for moderate-sensitivity thermal imaging.

The Golay Cell²⁰

This detector is used mostly for laboratory operations, as it is slow and fragile, although it has high sensitivity. It is a gas-filled chamber that has a thin membrane at one end and a blackened detector area at the other. Light on the blackened surface causes the increase in temperature; this is transferred to the gas which therefore expands. The membrane bulges, and the amount of the bulge is sensed by some sort of optical lever²¹ or even change in capacity of an electrical element.²² Other versions do not use a blackened surface, but allow the radiation to interact with the gas directly, in which case they are spectral detectors that are "tuned" to the absorption spectrum of the gas.²³

Pyroelectric Detectors²⁴

Some crystals which do not have a center of symmetry experience an electric field along a crystal axis. This internal electric field results from an alignment of electric dipoles (known as polarization), and is related to the crystal temperature. In these ferroelectric crystals, this results in a charge being generated and stored on plates connected to the crystal. Polarization disappears above the so-called Curie temperature that is characteristic of each material. Thus, below the Curie temperature, a change in temperature results in a current. The equation for the response of a pyroelectric detector is

$$\Re = \frac{\omega p A_d \epsilon R_e R_{th}}{(1 + i\omega\tau_{th})(1 + i\omega\tau_e)} \quad (16)$$

where ω is the radian frequency, p is the pyroelectric coefficient, A_d is the detector area, R_e is the electrical resistance, R_{th} is the thermal resistance, τ_{th} is the thermal time constant, and τ_e is the electrical time constant. The relation is shown in Fig. 3, where the responsivity is plotted as a function of frequency. In the low-frequency region the responsivity rises directly as the frequency. This is a result of the ac operation of a pyroelectric. At the (radian) frequency that is the reciprocal of the slower (usually the thermal) time constant,

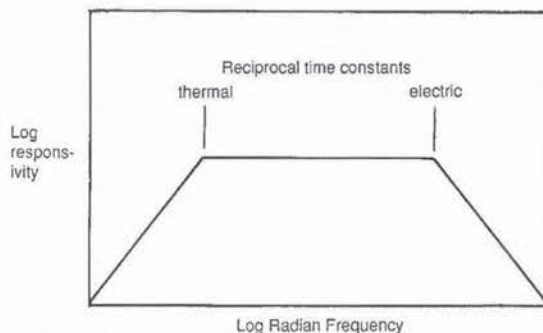


FIGURE 3 Responsivity asymptotes vs frequency.

the response levels off. This is the product of the ac rise and the thermal rolloff. Then, when the frequency corresponding to the shorter time constant is reached, the response rolls off.

Type II pyroelectric detectors work on a slightly different mechanism, which is still not fully understood. The electrodes are on the sensing surface of the detector and parallel to the polar axis. In these crystals, the temperature change is not uniform at the onset of radiation and the primary and secondary pyroelectric effects take place, thereby generating a body electric charge distribution.²⁵ Materials most often used for these detectors are TGS (triglycine sulfate), DTGS (deuterated TGS), Li_2SO_4 , LiNbO_3 , LiTaO_3 , and PLZT (lead lanthanum zirconate titanate). TGS is the most used for specialized sensor systems, but has a relatively low Curie point. For higher-temperature operation, usually LiTaO_3 or PLZT is used in the general laboratory environment.

The two advantages of the pyroelectric detector over the other thermal detectors, bolometers, and thermopiles, are its responsivity and its capability of rapid response. The response time and responsivity are traded by choice of the load resistor in the circuit. For instance, with a 100-M Ω load the time constant can be 1 ms and the responsivity 100 V/W, but with a 1-M Ω load the values would be 10 μs and 1.

Summary of Elemental Thermal Detector Properties

Although the user should contact suppliers for detailed information, this section provides overall property information about thermal detectors. There are several cautions about summary data. Most detectors can be tailored to have somewhat different properties. Improvements may have been made since the publication of these results. Not all parameters are available in all combinations. Table 1 does, however, give the general flavor of the performance of different thermal detectors.

19.3 ARRAYS

Introduction

As pointed out earlier, thermal detector response is governed by the thermal response time, which is the ratio of the pixel heat capacity C to the thermal conductance G of the heat leakage mechanism. High pixel responsivity is associated with high thermal isolation, i.e., low thermal conductance. Thermal detector design is driven by the thermal isolation structure. It is the structure which determines the extent to which the pixel performance

TABLE 1 General Properties of Thermal Detectors

Type	Operating Temperature (K)	$D^* \times 10^8$ (cmHz ^{1/2} W ⁻¹)	NEP $\times 10^{-10}$ (WHz ^{-1/2})	Time constant (m)	Size (mm ²)
Silicon bolometer	1.6		3×10^{-5}	8	0.25–0.70
metal bolometer	2–4	1		10	
Thermistor bolometer	300	1–6		1–8	0.01–10
Germanium bolometer	2–4		0.005	0.4	1.5
Carbon bolometer	2–4		0.03	10	20
Superconducting bolometer (NbN)	15		0.2	0.5	5×0.25
Thermocouples	300		2–10	10–40	0.1×1 to 0.3×3
Thermopiles	300			3.3–10	1–100
Pyroelectrics	300	2–5		10–100*	2×2
Golay cell	300	10	0.6	10–30	10

*Shorter values can be obtained at the expense of NEP (for laser detection).

can approach the temperature fluctuation noise limit and, ultimately, the background fluctuation noise limit. Given the value of G associated with the heat loss mechanism, the pixel heat capacity must be designed appropriately to attain the required thermal response time. Response times in the millisecond range are compatible with high thermal isolation; response times in the microsecond range are not. Thus, two-dimensional arrays of thermal detectors which operate at TV frame rates (30 Hz in the United States) are under development for applications in thermal imagers.

Noise Equivalent Temperature Difference (NETD)

Whereas elemental detectors are usually described by such figures of merit as NEP and D^* , arrays have been described by an NETD associated with their use in a camera under certain specific conditions. It is defined as the change in temperature of a blackbody which fills the field of view of a pixel of an infrared imaging system that gives rise to a change of unity in the signal-to-noise ratio at the output of the system. The measurement of the NETD should, however, be with the flooding of several pixels to avoid fringing effects and with an SNR (signal-to-noise ratio) well above 1 to obtain good accuracy. The pixel is defined as the subtense of a single element of the array. The NETD can be written in several different forms. Perhaps the simplest is

$$\text{NETD} = \frac{\sqrt{A_d B}}{D^* (dP_d/dT)} \quad (17)$$

where D^* is the specific detectivity, A_d is the area of a single pixel, B is the system bandwidth and (dP_d/dT) represents the change in power on the detector element per unit change in temperature in the spectral band under consideration. This form does not include the system noise, which is often included by the manufacturers in their calculations. In Eq. (17) the change in power with respect to temperature is

$$\frac{dP_d}{dT} = \frac{A_d \tau_o}{4FN^2} \int_{\lambda_1}^{\lambda_2} \frac{dM}{dT} d\lambda \quad (18)$$

where τ_o is the optics transmission, FN is the focal ratio (defined as the effective focal length divided by the entrance pupil diameter), and M is the radiant emittance of the source. This is almost the definition of the specific detectivity. The NETD can also be written in terms of the responsivity $\mathfrak{R}$, since the detectivity and responsivity are related in the following way:

$$D^* = \frac{\sqrt{A_d B}}{P} \frac{V_s}{V_N} = \sqrt{A_d B} \frac{\mathfrak{R}}{V_N} \quad (19)$$

where V_s is the signal voltage at the sensor and V_N is the rms noise voltage of a pixel in the bandwidth B . Therefore

$$\text{NETD} = \frac{V_N}{\mathfrak{R}(dP_d/dT)_{\lambda_1-\lambda_2}} \quad (20)$$

The power on the detector is related to the power on the aperture by the optical transmission τ_o . The expression can also be formulated in terms of the source radiance, L

$$\text{NETD} = \frac{4FN^2\sqrt{B}}{D^*\tau_a\tau_o\pi D \Delta\theta(dL/dT)_{\lambda_2-\lambda_1}} \quad (21)$$

where D is the diameter of the aperture, $\Delta\theta$ is the angular subtense of a pixel, L is the source radiance and τ_a is the atmospheric transmission. One last form can be generated by recognizing that, for an isotropic radiator, the radiance is the radiant emittance divided by π :

$$\text{NETD} = \frac{4FN^2V_N}{A_D\tau_a\tau_o\mathfrak{R}(T_s)(dM/dT)} \quad (22)$$

In this form of the expression for NETD, it is not necessary that the noise be white, nor is it necessary that the noise not include system noise. Whether or not system noise is included should be clearly stated.

Theoretical Limits

Figure 4 illustrates the theoretical limits of thermal arrays having the parameters shown and operating at 300 and 85 K as a function of thermal conductance. The performance of real thermal arrays with those parameters lies on or above the sloping line. As the conductance G is reduced (better thermal isolation), the noise equivalent temperature difference NETD is reduced (improves) until the background limit is reached, when radiant power exchange between the array and the background becomes the dominant heat transfer mechanism. Reducing the detector temperature to 85 K appropriate to a bolometer operating at the transition edge of the high-temperature superconductor $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ (YBCO) reduces the NETD by $\sqrt{2}$ and allows the limit to be reached with less thermal isolation (higher G value).

Arrays fall into two categories: monolithic and hybrid. Monolithic arrays are prepared on a single substrate, e.g., silicon, upon which the detecting material is deposited in the form of a thin film which is subsequently processed into an array. Hybrid arrays are prepared in two parts: (1) the readout electronics array, usually in silicon, and (2) the detecting material array, usually in wafer form which is thinned by lapping, etching, and polishing. These two arrays are mated by a technique such as flip-chip bonding. Here the interconnection at each pixel must have a sufficiently high electrical conductivity, yet a sufficiently low thermal conductivity—a difficult requirement. If array cost considerations are important, then the monolithic approach, especially in silicon, is the more desirable.

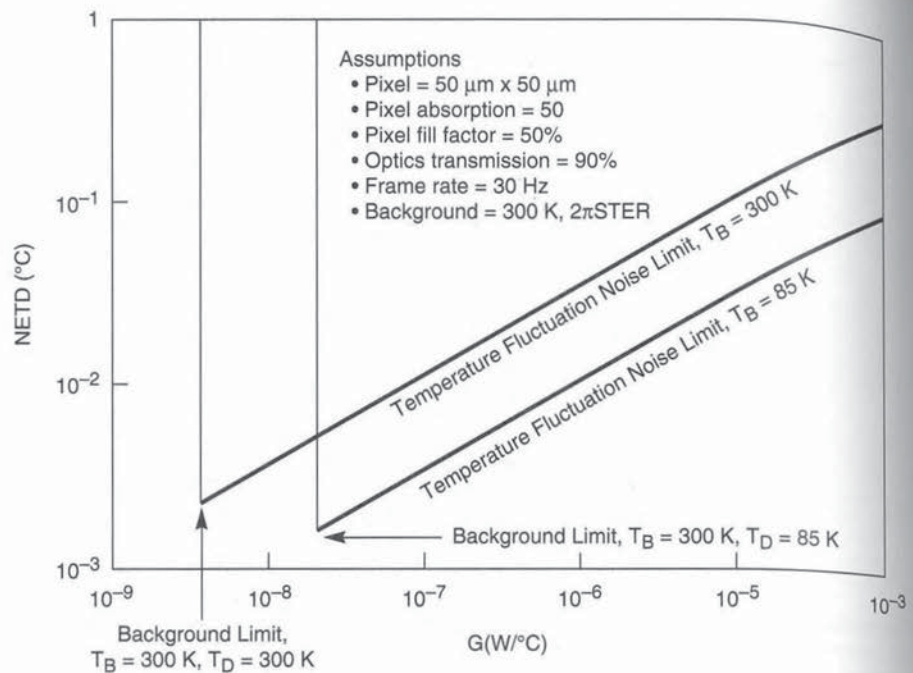


FIGURE 4 Temperature fluctuation noise limit and background fluctuation noise limit of uncooled and cryogenic thermal detector arrays.

Resistive Bolometer Arrays

The development of resistive bolometric arrays has proceeded along two paths: uncooled arrays and cryogenic arrays. Large, uncooled bolometric arrays have been developed at Honeywell by a team lead by R. A. Wood.^{26,27} Silicon microstructure technology is employed to produce the arrays, a process resembling the fabrication of integrated circuits. Twelve arrays are prepared on a 4-inch-diameter silicon wafer. Each monolithic array consists of 240×336 pixels; each pixel is $50 \times 50 \mu\text{m}$. The detecting material is a thin film of vanadium oxide. A Si_3N_4 membrane having a thermal conductance of $1 \times 10^{-7} \text{ WC}^{-1}$ supports the vanadium oxide at each pixel, as shown in Fig. 5. Bipolar transistors implanted in the silicon substrate act as pixel switches for the matrix-addressed array. The response is optimized for the 8–14 μm spectral interval. The thermal response time is adjusted for a 30 Hz frame rate. Each pixel is addressed once per frame by a 5- μs pulse. A thermoelectric stabilizer maintains the array at ambient temperature. Other than a one-shot shutter, the camera has no moving parts.

The measured NETD of the camera with F/1 optics at 300 K is 0.04 K. Given the G value of $1 \times 10^{-7} \text{ WK}^{-1}$ it can be seen from Fig. 4 that the array is within a factor of 4 of the temperature fluctuation noise limit. Furthermore, the pixel thermal isolation is so complete that there is no measurable thermal spreading among the pixels.

Linear resistive bolometric arrays of the high-temperature superconductor YBCO on silicon microstructures have been prepared by Johnson et al.,²⁸ also of Honeywell. A two-dimensional array is under development.²⁹ The monolithic arrays operate at the transition edge from 70 to 90 K. As was true for the uncooled arrays, the superconducting ones employed a silicon nitride membrane to support the thin film and provide thermal isolation. Excess noise at the contacts limited the performance of the 12-element linear

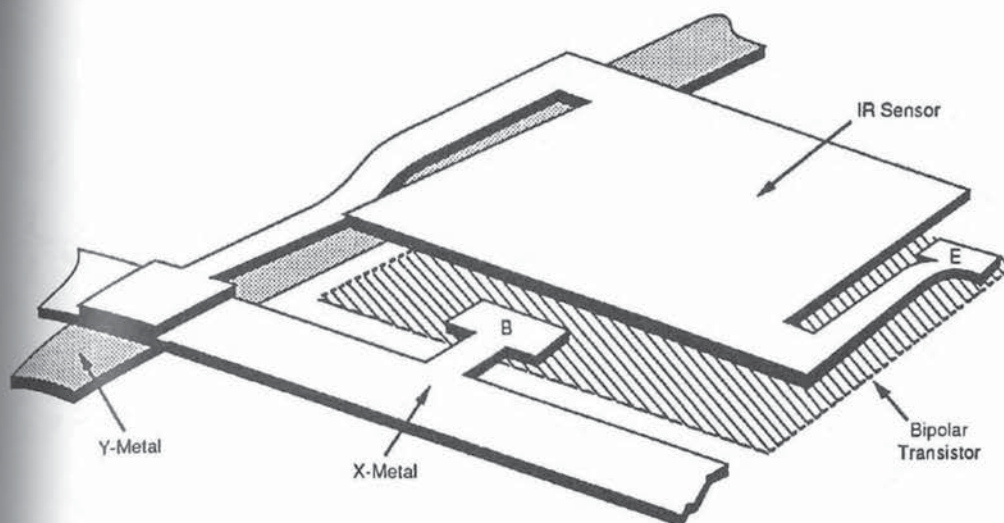


FIGURE 5 Monolithic silicon microbolometer.^{26,27} (© 1992 IEEE.)

array. With no excess noise, the calculated NETD²⁹ of a 240×336 array with $50 \mu\text{m}$ pixels and F/1 optics would be 0.002 K, which is near the 300 K background limit, as shown in Fig. 4.

Pyroelectric Hybrid Arrays

Linear and two-dimensional pyroelectric uncooled arrays have been under development for more than two decades.²⁰⁻³² The arrays employ hybrid construction, in which a bulk pyroelectric ceramic material such as lithium tantalate or lead zirconate is mechanically thinned, etched, and polished, then bump-bonded to a silicon substrate containing readout electronics,³³ as shown in Fig. 6. Reticulation is usually employed to prevent lateral heat conduction through the pyroelectric material. The theoretical system NETD of a two-dimensional uncooled array with F/1 optics is estimated to be 0.1 K.³³ Two-dimensional uncooled arrays operating in the 8–14- μm region having 100×100 pixels, each $100 \times 100 \mu\text{m}$, are available commercially.³⁴ Their NETD (with F/1 speed) is 0.35 K. A two-dimensional pyroelectric monolithic array employing a thin film of lead titanate on a silicon microstructure is under development.³⁵

Ferroelectric bolometer arrays, also known as field-enhanced pyroelectric arrays, have been developed by Texas Instruments.^{36,37} Operation depends upon the temperature dependence of the spontaneous polarization and dielectric permittivity in a ferroelectric ceramic near the Curie temperature. Barium strontium titanate (BST), the selected material, has its composition (barium-to-strontium ratio) adjusted during preparation so the Curie point is 22°C. A thermoelectric stabilizer is employed to hold the BST near 22°C such that the absorbed infrared radiation changes the temperature and thus the dielectric properties. The effect is similar to the pyroelectric effect; however, a voltage is applied to enhance the signal. Construction of this array is naturally similar to that of the pyroelectric array, described above, as shown in Fig. 7. Reticulation of the ceramic is frequently applied to these arrays as well. A radiation chopper is required as both the pyroelectric

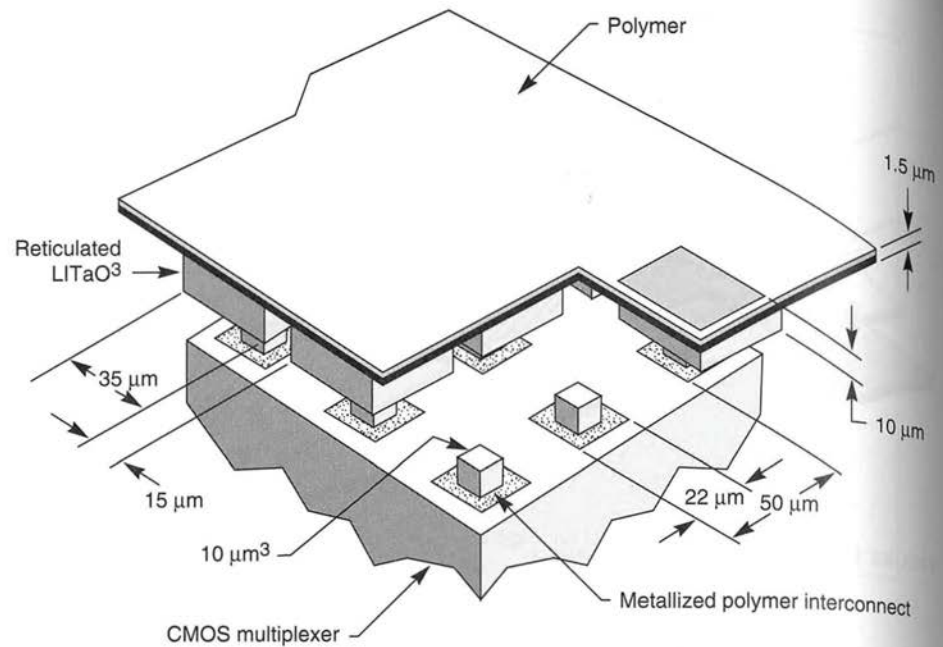


FIGURE 6 Hybrid pyroelectric array structure.³³

and ferroelectric effects depend upon the change in temperature. The Texas Instruments BST array, incorporating 80,000 pixels, each about $50 \times 50 \mu\text{m}$, which are matrix addressed, has an NETD of about 0.05°C (with F/1 optics).

Thermoelectric Arrays

Thermoelectric arrays prepared by silicon micromachining have been described by Choi and Wise.³⁸ Series-connected, thin-film thermocouples, i.e., a thermopile, are prepared on a silicon microstructure, the "hot" junctions (receiving the thermal radiation) on a silicon

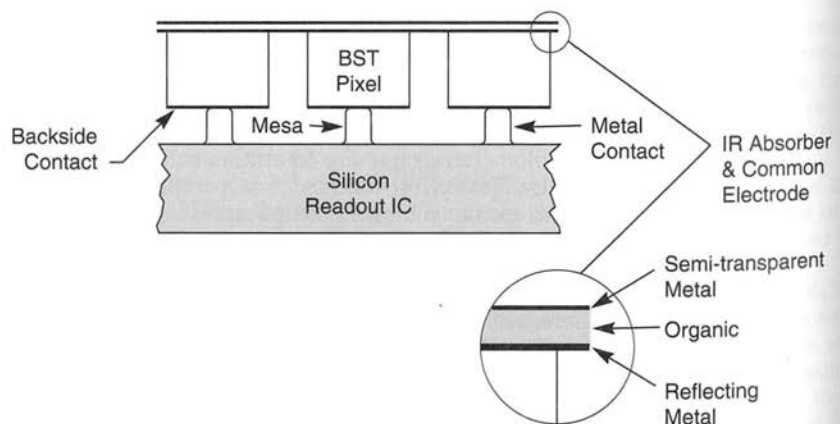


FIGURE 7 Ferroelectric bolometer hybrid array.^{36,37}

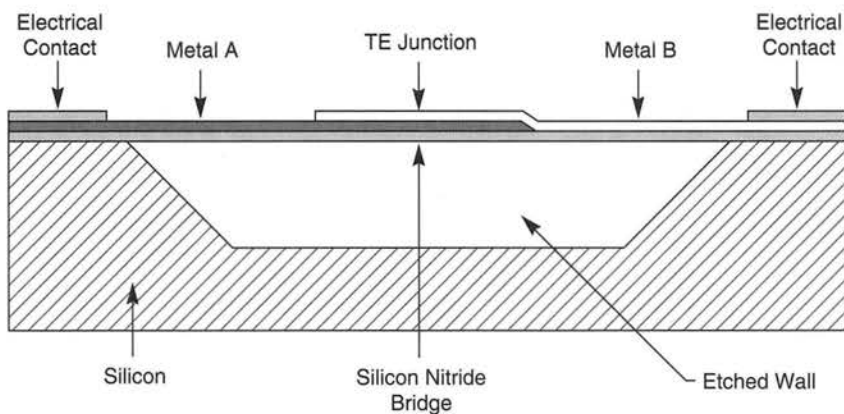


FIGURE 8 Monolithic thermoelectric array.³⁹ (© 1991 Instrument Society of America. Reprinted with permission from the Symposium for Innovation in Measurement Science.)

nitride/silicon dioxide membrane and the "cold" shield junctions on the surrounding silicon substrate. Both 64- and 96-pixel microthermopile linear arrays in silicon microstructures have been prepared by Honeywell,³⁹ each microthermopile consisting of several nickel iron/chromium thermocouples connected in series, as shown in Fig. 7. The "hot" junctions are deposited on silicon nitride membranes, whereas the "cold" junctions are on the silicon substrates. A camera incorporating the linear array has been employed to image moving targets such as automobiles. With an F/0.73 lens, the measured NETD is 0.10°C.

19.4 REFERENCES

1. P. W. Kruse, L. D. McGlauchlin, and R. B. McQuistan, *Elements of Infrared Technology*, John Wiley and Sons, New York, 1962.
2. R. D. Hudson and J. W. Hudson, *Infrared Detectors*, Halsted Press, 1975; and F. R. Arams, *Infrared to Millimeter Wave Detectors*, Artech House, 1973.
3. E. H. Putley, *Optical and Infrared Detectors*, chap. 3, R. J. Keyes (ed.), Springer-Verlag, 1980.
4. R. A. Smith, F. E. Jones, and R. P. Chasmar, *The Detection and Measurement of Infrared Radiation*, Oxford University Press, 1968.
5. R. C. Jones, "The Ultimate Sensitivity of Radiation Detectors," *J. Opt. Soc. Am.* **37**:879 (1974).
6. S. Nudelman, "The Detectivity of Infrared Detectors," *Appl. Optics* **1**:627 (1962).
7. F. J. Low, "Low Temperature Germanium Bolometer," *J. Opt. Soc. Am.* **51**:1300 (1961).
8. S. Zwerdling, R. A. Smith, and J. P. Thierault, "A Fast High Responsivity Bolometer for the Very Far Infrared," *Infrared Physics* **8**:271 (1968).
9. N. Coron, "Infrared Helium Cooled Bolometers in the Presence of Background Radiation: Optimal Parameters and Ultimate Performance," *Infrared Physics* **16**:411 (1976).
10. N. Coron, G. Dambier, and J. Le Blanc, *Infrared Detector Techniques for Space Research*, V. Manno and J. Ring (eds.), Reidel, 1972.
11. N. Coron, G. Dambier, J. Le Blanc, and J. P. Moliac, "High Performance, Far Infrared Bolometer Working Directly in a Helium Bath," *Rev. Sci. Instr.* **46**:492 (1975).
12. W. H. Andrews, R. M. Milton, and W. De Sorbo, "A Fast Superconducting Bolometer," *J. Opt. Soc. Am.* **36**:518 (1946).

13. R. M. Milton, "A Superconducting Bolometer for Infrared Measurements," *Chem. Rev.* **39**:419 (1946).
14. N. J. Fuson, "The Infrared Sensitivity of Superconducting Bolometers," *J. Opt. Soc. Am.* **38**:845 (1948).
15. G. Gallinaro and R. Varone, "Construction and Calibration of a Fast Superconducting Bolometer," *Cryogenics* **15**:292 (1975).
16. K. B. Bhasin and V. O. Heinen (eds.), "Superconductivity Applications for Infrared and Microwave Devices," *Proceedings SPIE* **1292** (1990). (Includes many other references.)
17. W. S. Boyle and K. F. Rodgers, *J. Opt. Soc. Am.* **49**:66 (1959).
18. D. F. Hornig and B. J. O'Keefe, "Design of Fast Thermopiles and the Ultimate Sensitivity of Thermal Detectors," *Rev. Sci. Instr.* **18**:7 (1947).
19. P. B. Felgett, "Dynamic Impedance and the Sensitivity of Radiation Thermocouples," *Proc. Phys. Soc.* **B62**:351 (1949).
20. M. J. E. Golay, "Theoretical Considerations in Heat and Infrared Detection with Particular Reference to the Pneumatic Detector," *Rev. Sci. Instr.* **18**:347 (1947); "Pneumatic Infrared Detector," *ibid.*, **18**:357 (1947); "Theoretical and Practical Sensitivity of the Pneumatic Detector," *ibid.*, **20**:816 (1949).
21. J. R. Hickey and D. B. Daniels, "Modified Optical System for the Golay Detector," *Rev. Sci. Instr.* **40**:732 (1969).
22. M. Chatanier and G. Gauffre, *IEEE Transactions Instr. and Meas.* **IMEE** **179**, 1973.
23. A detector once made by Patterson Moos and cited by R. DeWaard and E. Wormser, "Description and Properties of Various Thermal Detectors," *Proc. IRE* **47**:1508 (1959).
24. E. H. Putley, *Semiconductors and Semimetals*, vol. 5, R. K. Willardson and A. C. Beer (eds.), Academic Press, 1970, chap. 6, "The Pyroelectric Detector"; vol. 12, 1977, chap. 7.
25. Zu-Sheng Wang and Jian-Qi Zhang, "The Mechanism of Type II Pyroelectric Detectors," *Infrared Physics*, vol. 33, no. 6, 1993, 481-486.
26. R. A. Wood, C. J. Han, and P. W. Kruse, "Integrated Uncooled Infrared Detector Imaging Array," *Proc. of the 1992 IEEE Solid State Sensor and Actuator Workshop*, Hilton Head Island, S.C., pp. 132-135.
27. R. A. Wood, "Uncooled Thermal Imaging with Monolithic Silicon Focal Plane Arrays," *Proc. SPIE* **2020**:Infrared Technology XIX (1993).
28. B. R. Johnson, T. Ohnstein, C. J. Han, R. Higashi, P. W. Kruse, R. A. Wood, H. Marsh, and S. B. Dunham, "High- T_c Superconductor Microbolometer Arrays Fabricated by Silicon Micromachining," *IEEE Trans. Appl. Superconductivity* **3**:2856 (1993).
29. B. R. Johnson and P. W. Kruse, "Silicon Microstructure Superconducting Microbolometer Infrared Arrays," *Proc SPIE* **2020**:Infrared Technology XIX (1993).
30. E. H. Putley, "The Pyroelectric Detector," *Semiconductors and Semimetals*, vol. 5, *Infrared Detectors*, R. K. Willardson and A. C. Beer (eds.), Academic Press, New York, 1970.
31. P. A. Manning, D. E. Burgess, and R. Watton, "A Linear Pyroelectric Array IR Sensor," *Proc. SPIE* **590**:2 (1985).
32. R. Watton and M. V. Mansi, "Performance of a Thermal Imager Employing a Hybrid Pyroelectric Detector Array with MOSFET Readout," *Proc. SPIE* **865**:79 (1987).
33. N. Butler and S. Iwasa, "Solid State Pyroelectric Imager," *Proc. SPIE* **1685**:146 (1992).
34. GEC-Marconi Materials Technology Ltd., 9360 Ridgehaven Court, San Diego, CA 92123.
35. B. E. Cole, R. D. Horning, and P. W. Kruse, "PbTiO₃ Deposited by an Alternating Dual-Target Ion-Beam Sputtering Technique," *Ferroelectric Thin Films II*, A. I. Kingon, E. R. Myers and B. Tuttle (eds.), *Materials Research Society Symposium Proceedings*, **243**:185 (1992).
36. C. Hanson, H. Beratan, R. Owen, M. Corbin, and S. McKenney "Uncooled Thermal Imaging at Texas Instruments," *Infrared Detectors: State of the Art*, *Proc. SPIE* **1735**:17 (1992).
37. C. M. Hanson, "Uncooled Ferroelectric Thermal Imaging," *Proc. SPIE* **2020**: Infrared Technology XIX (1993).
38. I. H. Choi and K. D. Wise, "A Silicon-Thermopile-Based Infrared Sensing Array for Use in Automated Manufacturing," *IEEE Trans. on Electron Devices* **ED-33**:72 (1986).
39. M. Listvan, M. Rhodes, and M. L. Wilson, "On-line Thermal Profiling for Industrial Process Control," *Proceedings of the Instrument Society of America, Symposium for Innovation in Measurement Science*, Geneva, N.Y., August 1991.

CHAPTER 22

VISIBLE ARRAY DETECTORS

Timothy J. Tredwell

*Eastman Kodak Company
Sensor Systems Division
Imager Systems Development Laboratory
Rochester, New York*

22.1 GLOSSARY

A	area of the pixel
C_{FD}	total capacitance of the floating diffusion in a CCD output
C_g	gate capacitance per unit area
C_r	readout line capacitance
J_D	total dark current per unit area
J_s	surface generation current
L_e	diffusion length of electrons in silicon
L_p	diffusion length of holes in silicon
N_A	p -type dopant concentration in silicon
N_D	n -type dopant concentration in silicon
N_e	total number of electrons collected in a pixel
$N(0)$	number of photons entering the silicon
$N(x)$	number of photons remaining a distance x below the surface
n_i	intrinsic carrier concentration in silicon
$P(x)$	probability that an electron-hole pair generated a distance x from the surface will be collected before recombination
q	electron charge
s_o	surface recombination velocity
T_{int}	integration time of light in an image sensor
$T(\lambda)$	transmission of light
V_{bi}	built-in voltage for a silicon pn junction
$W(V)$	width of the depletion layer at a given bias voltage in the MOS capacitor or junction-photodiode

22.2 IMAGING DETECTORS

$\alpha(\lambda)$	absorption coefficient of light
ϵ_s	silicon dielectric constant
μ_e	electron mobility in silicon
μ_p	hole mobility in silicon
τ_o	depletion-layer lifetime
τ_p	minority carrier hole lifetime in silicon
Φ_s	electrostatic potential at the silicon-silicon dioxide, also called surface potential

22.2 INTRODUCTION

Since the invention of the image sensor in 1964, solid state image sensors have advanced in resolution, sensitivity, and image quality to the point where they have replaced other methods of converting visible light to electronic signals in nearly all imaging applications. There are two types of image sensors: *area image sensors*, which are used in cameras, and *linear sensors*, which are used in scanning applications. Cameras using area image sensors dominate the camcorder, video and broadcast, machine vision, scientific, and medical fields. Area image sensors for camcorder applications are typically 400,000 picture elements in resolution, 60 dB in dynamic range, and have noise levels of a few tens of electrons. Area image sensors for scientific applications may have resolutions of over six million elements, dynamic ranges exceeding 80 dB, and noise levels approaching the single electron level. Production of area image sensors exceeded eight million devices in 1993. Scanners employing linear solid-state image sensors dominate facsimile, document scanner, digital copier, and film scanner applications. Linear sensors range from 2000-element monochrome arrays with 40 dB of dynamic range used in facsimile applications to 8000 or more element trilinear arrays with 80 dB of dynamic range for high-performance color scanning applications. Production of linear CCD image sensors also exceeded eight million devices in 1993.

The steps involved in image sensing consist of (1) converting the incoming photons to charge at picture element (pixel), and (2) transferring that charge to an output amplifier and converting the charge to a voltage or current signal which can be sensed by circuits external to the sensor. The image-sensing elements will be described first, followed by readout elements. Sensor architectures for area and linear sensors will then be described.

22.3 IMAGE SENSING ELEMENTS

There are four basic types of structures which are used for image sensing: the junction photodiode, the photocapacitor, the pinned (p^+np) photodiode, and the photoconductor.* The first three are generally fabricated in single-crystal silicon as part of the image sensor; the photoconductor is usually fabricated from amorphous silicon deposited over the image sensor. The photoconversion process begins with the absorption

* For some scientific applications in which high quantum efficiency and fill factor are essential, the silicon wafer will be thinned to 10 μm or less in thickness and illuminated from the backside. The frontside contains an area charge coupled device, which is used to collect the photogenerated carriers.

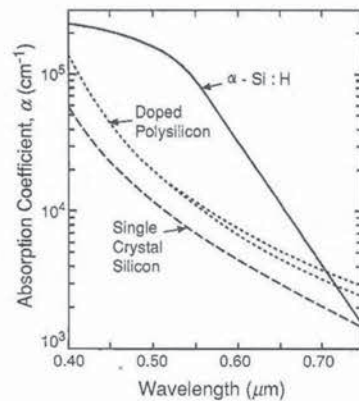


FIGURE 1 Absorption coefficient for light in single-crystal silicon, heavily doped polycrystalline silicon, and hydrogenated amorphous silicon as a function of wavelength.

of a photon in the silicon resulting in the generation of a single electron-hole pair. The absorption of light at a particular wavelength is given by:

$$N(x, \lambda) = N(0)e^{-\alpha(\lambda)x} \quad (1)$$

where $N(x)$ is the number of photons remaining a distance x below the surface, $N(0)$ is the number of photons entering the silicon, and $\alpha(\lambda)$ is the absorption coefficient.¹ The absorption coefficient is shown as a function of wavelength λ for single-crystal silicon, doped polycrystalline silicon, and hydrogenated amorphous silicon in Fig. 1. The absorption depth is defined as the inverse of the absorption coefficient [$d(\lambda) = 1/\alpha(\lambda)$]. In single-crystal silicon the absorption depth is 0.4 μm in the blue (450 nm), 1.5 μm in the green (550 nm), and 3.0 μm in the red (640 nm). In the infrared, the absorption depth increases to 10.5 μm at 800 nm. Beyond 1100 nm, the absorption is virtually zero because the photon energy is less than the 1.1-eV silicon bandgap.

Junction Photodiode

The junction photodiode is one of the most common image-sensing elements. The physical structure and band diagram of the junction photodiode are shown in Fig. 2a for a p -type substrate. The n -type region is formed by ion implantation or diffusion of phosphorous or arsenic to a depth of 2000 to 10,000 Å into the p -type silicon. The n -type dopant region is usually graded, with the highest concentration at the surface. The gradient in n -type dopant concentration results in a gradient in electrostatic potential which accelerates photogenerated carriers (holes in the n -type region) away from the surface. This reduces loss of photogenerated carriers to surface recombination. The photodiode is typically operated with a reverse bias V of 1 to 5 volts. A depletion layer is formed between the n - and p -type regions.* The width of the depletion layer $W(V)$ for an abrupt $n + p$ junction is given by:

$$W(V) = \sqrt{\frac{2\epsilon_s(V + V_{bi})}{qN_A}} \quad (2)$$

* For a detailed review of the device physics of junction diodes and MOS capacitors, see Sze, *Physics of Semiconductor Devices*, Wiley, New York, 1969.

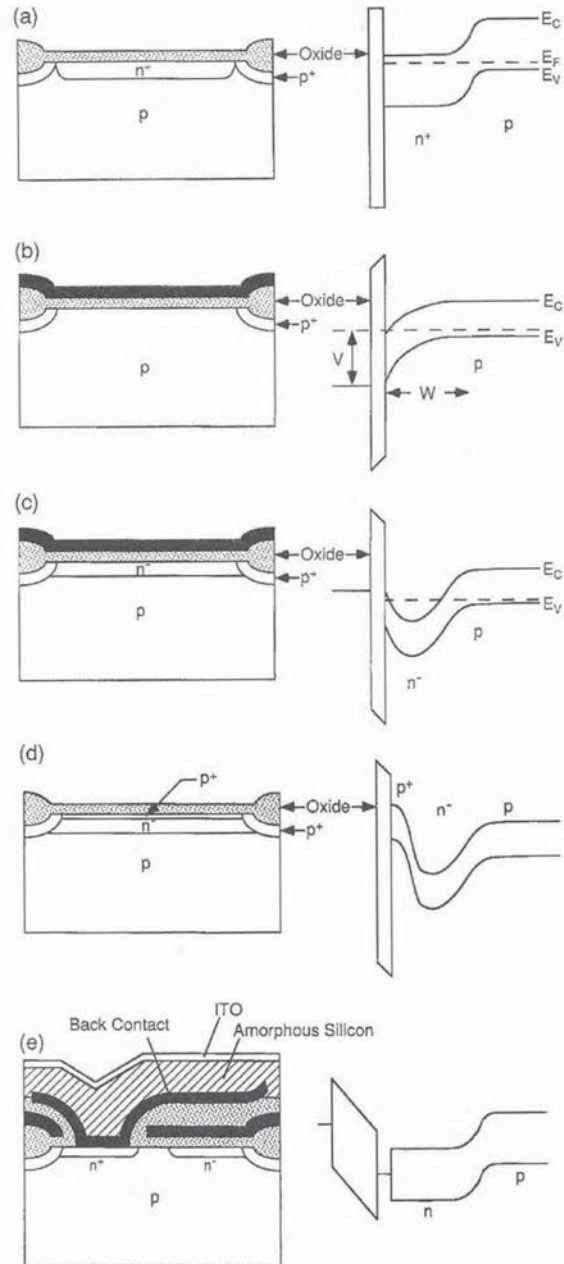


FIGURE 2 Cross-sectional diagrams and band diagrams for (a) junction photodiode; (b) surface-channel MOS capacitor; (c) buried-channel MOS capacitor; (d) pinned, or hole-accumulated, diode; (e) amorphous silicon photoconductor.

where q is the electronic charge, ϵ_s is the silicon dielectric constant, N_A is the p -type dopant concentration, and V_{bi} is the built-in voltage given by

$$V_{bi} = \frac{kT}{q} \ln \frac{N_A}{n_i} \quad (3)$$

where n_i is the intrinsic carrier concentration. For silicon doped at $1 \times 10^{16} \text{ cm}^{-3}$, the depletion width at 5-V reverse bias is $0.8 \mu\text{m}$.

If the diode is illuminated, some of the photons will be absorbed in the n -region, some in the depletion layer, and the remainder in the p -type substrate. The quantum efficiency $\eta(\lambda)$ is the ratio of the charge collected to the number of photons incident on the diode (i.e., 100-percent quantum efficiency refers to one electron-hole pair collected for every incident photon). The quantum efficiency depends on three factors: transmission $T(\lambda)$ of light through the overlying layers into the silicon, absorption of light in the silicon, and the probability $P(x)$ that an electron-hole pair generated a distance x from the surface will be collected before recombination:

$$\eta(\lambda) = T(\lambda) \int_{x=0}^{\infty} (1 - e^{-\alpha(\lambda)x}) P(x) dx \quad (4)$$

The transmission of light through the overlying layers into the silicon can be calculated using standard multilayer interference models.

The collection of the photogenerated charge takes place by two processes: drift and diffusion. Drift is the movement of electrons and holes due to an electric field. Even for small electric fields, transport of carriers by drift will dominate diffusion. This is the case in the depletion region. Outside the depletion region, such as in the p -type substrate, there is no electric field, and carrier transport occurs by diffusion. For example, a photon absorbed in the p -type region will excite an electron from the valence band to the conduction band. The electron will move in a three-dimensional random walk until it recombines or until it encounters the edge of the depletion region, where it is swept across the junction by the electric field. The probability that an electron a distance x' from the junction can diffuse to the junction before recombining is given by:

$$P(x') = e^{-x'/L_e} \quad \text{where } L_e = \sqrt{\frac{KT}{q} \mu_e \tau_e} \quad (5)$$

in which L_e is the diffusion length, μ_e is the electron mobility, and τ_e the electron lifetime (typically $\sim 1 \mu\text{s}$). For a $1\text{-}\mu\text{s}$ lifetime, the electron diffusion length in p -type silicon is $50 \mu\text{m}$. Electrons generated from photons absorbed less than the diffusion length from the junction have a high probability of collection.*† Similarly, photons absorbed in the n -type

* In image sensors, the collection of photogenerated carriers can be complicated by a variety of factors. The doping concentration may not be uniform on either the n - or p -sides. On the n -side, the dopant concentration is designed to decrease from the surface to the junction, building in a potential gradient for holes away from the surface and toward the junction, preventing surface recombination. On the p -side, the dopant concentration may not be uniform owing to the use of epitaxial layers or wells diffused into the silicon. Additionally, the carrier lifetime may not be uniform. Impurity gettering, a process used in many image sensors to remove metallic contaminants during processing, will leave a region of crystalline defects in the silicon starting 20 to $50 \mu\text{m}$ beneath the silicon surface. The defects result in a very short electron lifetime in this region. Finally, diffusion takes place in three dimensions; a carrier absorbed beneath a given pixel in an array will diffuse laterally by the same amount it diffuses vertically. This could cause it to be collected in adjacent pixels.

† See, for example, Lavine et al., "Steady State Photocurrent Collection in Silicon Imaging Devices," *IEEE Transactions on Electron Devices* ED-30:1123-1133 (Sept. 1983).

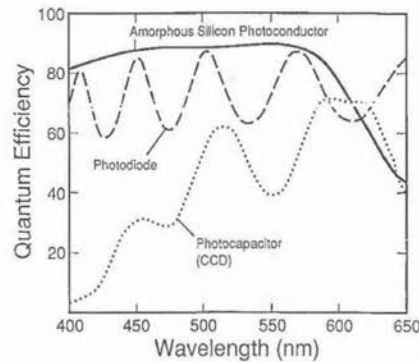


FIGURE 3 Quantum efficiency as a function of wavelength for photocapacitor, photodiode, and amorphous silicon photoconductor.

layer create electron-hole pairs. The holes must travel by diffusion to the junction to be collected. The probability that a hole a distance x' from the junction can diffuse to the junction is given by

$$P(x') = e^{-x'/L_p} \quad \text{where } L_p = \sqrt{\frac{kT}{q} \mu_p \tau_p} \quad (6)$$

in which L_p is the diffusion length, μ_p is the hole mobility, and τ_p is the hole lifetime. For a 1- μ s lifetime, the hole diffusion length in n -type silicon is 30 μ m. Since the n -type region in an $n+p$ junction is less than 1 μ m thick, the holes have no difficulty diffusing through the n -type region to the junction unless the n -type region is so heavily damaged or so heavily doped that the hole lifetime is very short. More typically, loss of quantum efficiency on the n -side results from recombination at the surface.

The quantum efficiency of a junction photodiode is illustrated as a function of wavelength in Fig. 3. In the ultraviolet, the light is absorbed very near to the surface and some of the photogenerated charge can be lost to surface recombination. In the 420- to 700-nm region, most of the light is absorbed close to the junction and is easily collected. The structure in the quantum efficiency illustrated in Fig. 3 results from the multilayer reflections of light in the oxide layer which overlies the photodiodes used in image sensors. The positions of the minima and maxima depend on the thicknesses and indices of refraction of the layers overlying the silicon. Beyond 800 nm, some of the photons are absorbed sufficiently deep in the silicon that the carriers recombine before they can diffuse to the junction; this results in the decrease in quantum efficiency at longer wavelengths.

In most image sensing applications, the junction diode at each picture element is used not only to collect the photogenerated carriers but also to store the carriers until they can be read out. In an imaging array, each photodiode would be reset to a reverse bias V , by a MOS gate. The capacitance of the diode of area A for an abrupt $n+p$ junction is given by $C(V) = \epsilon_s A / [W(V)]$ where $W(V)$ is the depletion width.

When a photogenerated carrier is collected by the junction, it is stored on the junction until it is read out. Storage of a carrier will cause the voltage on the junction to decrease by $q/C(V)$. When the photogenerated charge is removed from the junction during readout, the junction voltage is restored to its original value.

If so much photogenerated charge is stored on the photodiode that the voltage drops to zero before the charge can be read out, additional charge cannot be collected and will diffuse into the p -type substrate. This condition is referred to as saturation. The diffusion of excess charge into neighboring photosites is called *blooming*.

One of the difficulties encountered in using the junction photodiode in image sensor applications is *image lag*. The combination of the capacitance of the photodiode and the channel resistance of the MOS transfer gate used to read out the diode give rise to a time constant for transferring the photogenerated charge from the diode to the readout structure. As a result, not all the charge can be completely drained from the diode during the short reset times typically used in imaging applications. The remaining charge is drained in successive readouts, causing an afterimage. This effect is called image lag.³

MOS Capacitor

The MOS capacitor consists of the silicon substrate (taken to be *p*-type in this section), a thin layer of silicon dioxide (typically 200 to 1000 Å thick), and an electrode (typically polycrystalline silicon doped heavily *n*-type with phosphorous). The physical structure and the band diagrams of the surface channel MOS capacitor are shown in Fig. 2b for a *p*-type substrate. If the gate of the MOS capacitor is biased positive, a depletion layer is created in the *p*-type silicon substrate. The depth of the depletion layer depends on the substrate doping, gate voltage, and oxide thickness. The calculation of the depth of the depletion layer is somewhat more complex than the photodiode* and depends on the electrostatic potential at the surface, called the surface potential Φ_s . For values typical of image sensors ($N_A = 1 \times 10^{15}$, 5-V gate bias, 500-Å oxide thickness) the depletion layer is 2.4 μm deep. On the edges of the photocapacitor (see Fig. 2b) is a heavily doped *p*-type region overlaid by a thick (2000 to 5000-Å) oxide layer. This is called the field or *channel stop* region. Because of the heavier *p*-type doping, the field is not depleted by the voltage on the gate. The channel stops confine the electrons to the channel region.

If the MOS capacitor is illuminated, a fraction of the light will be reflected, a fraction will be absorbed in the polysilicon, and the rest will be transmitted into the silicon substrate. The absorption coefficient of heavily doped polysilicon is shown as a function of wavelength in Fig. 1. The absorption coefficient is $4 \times 10^4 \text{ cm}^{-1}$ at 450 nm and $1.2 \times 10^4 \text{ cm}^{-1}$ at 550 nm. For a 3000-Å-thick polysilicon electrode, less than 30 percent of the blue light and less than 70 percent of the green light is transmitted through the polysilicon. The photons entering the silicon are absorbed, either in the depletion layer of the MOS capacitor or in the undepleted *p*-type silicon beneath the depletion layer. Those photogenerated electrons created in the undepleted *p*-type region move by diffusion until they are captured by the depletion layer or until they recombine. The electrons are held at the silicon-SiO₂ interface, where they remain until they are read out. The quantum efficiency as a function of wavelength is illustrated in Fig. 3. The efficiency is low in the blue owing to absorption in the polysilicon. The structure in the quantum efficiency as a function of wavelength is due to multilayer interference in the polysilicon-oxide-silicon stack.

Charge is stored in the MOS capacitor at the silicon-silicon dioxide interface as a layer of sheet-charge only a few hundred angstroms thick. As additional photogenerated charge is added, the surface potential decreases. If sufficient photogenerated charge is added, the surface potential becomes zero and no additional charge can be stored. This condition is saturation.

The capacitance per unit area on which photogenerated charge is stored is the parallel capacitance of the oxide and the depletion layer:

$$C^{-1} = \left(\frac{t_{\text{ox}}}{\epsilon_{\text{ox}}} + \frac{W(\Phi_s)}{\epsilon_{\text{si}}} \right) \quad (7)$$

* To determine the depletion depth, the surface potential ϕ_s must be calculated. Φ_s depends on the voltage on the gate of the MOS capacitor, the oxide thickness, substrate doping, and weakly on temperature. See Sze, *Physics of Semiconductor Devices*.

where the depletion width is $W(\Phi_s)$.

$$W(\Phi_s) = \sqrt{\frac{2\epsilon_{si}\Phi_s}{qN_A}} \quad (8)$$

in nearly all cases, the oxide capacitance is the dominant term. As a result, the storage capacity of the MOS capacitor is significantly larger than the junction diode in which the charge is stored only on the depletion capacitance.

A variant of the surface-channel photocapacitor is the buried-channel photocapacitor. The structure and band diagram are shown in Fig. 2c. In this device, a lightly-doped n -type region is diffused or implanted into the silicon surface early in the fabrication process. This n -type region is sufficiently lightly doped that it is fully depleted. The n -type dopant in the buried channel results in a band diagram with a potential minimum, or well, for electrons just below the surface.* This well is separated from the surface by a few thousand angstroms in distance and about 1 V in potential. When the buried-channel photocapacitor is illuminated, the electrons collect in the buried channel and do not contact the surface. The primary purpose of the buried channel is to prevent electrons from being trapped by interface states at the silicon-silicon dioxide interface.

The photocapacitor has several advantages over the junction photodiode. These include higher storage capacity per unit area, zero lag readout, and generally lower dark current. The principal disadvantage is the low quantum efficiency in the blue. In some applications, a transparent electrode, such as indium-tin-oxide (ITO) may be substituted to improve the blue response.⁴ ITO has very low absorption over the visible (420 to 750 nm) and can be deposited sufficiently conductive for use as a gate electrode in an image sensor. Materials and processing complexity have prevented incorporation of ITO into commercial image sensors in volume until recently.

Another approach to achieving high quantum efficiency is the thinned backside-illuminated charge-coupled device (CCD). In this approach, the CCD (which is an array of MOS capacitors) is fabricated on the frontside of a silicon wafer. The wafer is then thinned from the backside to 10 μm or less in thickness. The backside is passivated to prevent surface recombination. Photons entering the backside are absorbed in the silicon beneath the MOS capacitors (usually buried channel). The photogenerated carriers diffuse to the capacitors, where they are held until they are read out. This device has quantum efficiency similar to the photodiode in the visible. However, because the silicon is thin, some of the photons at wavelengths beyond 700 nm will not be absorbed and so the quantum efficiency falls off beyond 700 nm. Owing to their extreme complexity in process and their extremely fragile design, backside-illuminated image sensors are limited to special scientific applications, especially astronomy.

Pinned Photodiode

The third type of photosensitive element is the pinned (p^+np) photodiode.⁵ This is sometimes called the *hole accumulation diode*, or HAD. This element combines the best features of the photodiode and photocapacitor, offering the high blue response of the photodiode with the high charge capacity, zero lag, and low dark current of the buried-channel photocapacitor. The pinned photodiode consists of a very shallow (<2000 Å) P^+ layer overlying an n -type buried-channel region. The structure and band

* For a review of the calculation of electrostatic potential and charge capacity of buried-channel MOS capacitors and charge-coupled devices, see B. C. Burkey, G. Lubberts, E. A. Trabka, and T. J. Tredwell, *IEEE Transactions on Electron Devices* **ED-31**(4):423 (April 1984).

diagrams are shown in Fig. 2d. The p^+ surface layer, which contacts the p^+ channel stop region on the sides, holds the electrostatic potential at the surface at 0 V. When the photodiode is illuminated, the photogenerated electrons are held in the n -type buried-channel region just below the surface.

The quantum efficiency of the pinned photodiode is nearly identical to that of the photodiode shown in Fig. 2. Because there is no overlying polysilicon electrode, the blue response is very high, similar to that of the photodiode. Because the buried-channel region can be completely emptied, the pinned photodiode does not have the lag of a normal junction (n^+p or p^+n) photodiode. The pinned photodiode is becoming the most widely used image-sensing element in interline area CCDs used for camcorders and for industrial and medical cameras. The pinned photodiode is also used in some linear image sensors, particularly where low image lag is critical.

Photoconductor

The last type of photosensitive device is the photoconductor. The most common material for the photoconductor is hydrogenated amorphous silicon, although other material systems have been explored. Amorphous silicon photoconductors have been used for two types of image sensors: area image sensors, where it is deposited on top of an area array to improve fill factor (i.e., the proportion of the picture element which is photosensitive), and contact linear sensors, where it is deposited on large ceramic or glass substrates to fabricate very long line or linear arrays.

The structure of an amorphous silicon photoconductor on a CCD image sensor and the corresponding band diagrams are shown in Fig. 2e. The hydrogenated amorphous silicon photoconductor consists of a back electrode, an undoped amorphous silicon layer approximately $1\text{ }\mu\text{m}$ thick, and a transparent top electrode.⁶ Additional doped amorphous silicon or silicon nitride layers may be added to the amorphous silicon front or back surfaces to improve performance. Photons absorbed in the amorphous silicon generate electron-hole pairs. Photogenerated electrons and holes are quickly swept to the back and front electrodes, respectively, because of the high electric field in the photoconductor. When the amorphous silicon is used as part of an area image sensor, the electrons on the back electrode can be transferred into the readout element; when used as part of a contact linear array, the voltage change can be amplified and read out through a multiplexer.

The quantum efficiency of an amorphous silicon photoconductor is shown in Fig. 3. Owing to the wider bandgap of the amorphous silicon, photons of wavelength greater than about 650 nm are not absorbed. The advantage of the amorphous silicon is the high quantum efficiency across the visible wavelengths and the ability to fabricate devices either on top of area CCDs for higher fill factor or to process on glass or ceramic for very large linear sensors. The disadvantages include charge trapping at defects in the amorphous silicon and low carrier mobility, both of which lead to field-dependent nonlinear response and image lag when used in an image sensor. Recent improvements in material and device technology have mitigated many of these disadvantages at the expense of process and device complexity.⁷

Antiblooming in Charge-sensing Elements

Blooming in image sensors occurs when the charge generated in an image-sensing element exceeds its capacity. If no method is provided to remove this excess charge, it will be injected either onto the readout element (CCD or MOS readout line) or into the substrate. If the excess charge is injected onto the readout element, it will usually appear as a bright line in the image. If it is injected into the substrate, the charge can diffuse in a circular pattern and be collected by neighboring elements.

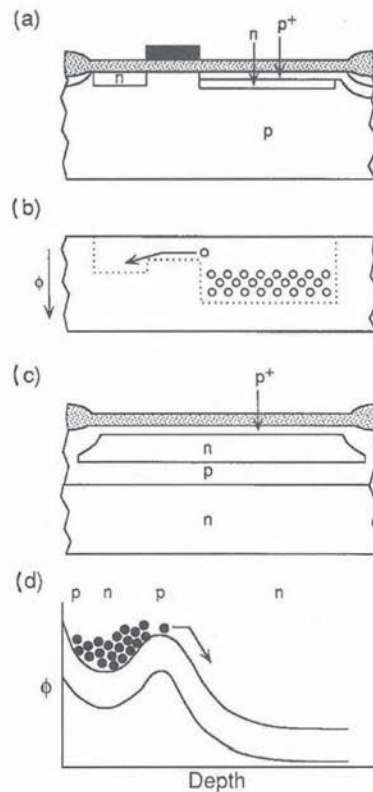


FIGURE 4 Antiblooming methods for image sensors: (a) cross section of lateral antiblooming structure; (b) illustration of electrostatic potential and charge overflow in lateral antiblooming; (c) cross section of vertical antiblooming structure; (d) band diagram and illustration of charge overflow in vertical antiblooming.

There are two basic types of antiblooming circuits: lateral and vertical.^{8,*} In lateral antiblooming, illustrated in Fig. 4a and Fig. 4b, an MOS antiblooming gate and an antiblooming drain are provided adjacent to each image-sensing element. Excess charge on the sensing element overflows the antiblooming gate onto the antiblooming drain. The antiblooming drains of all elements on the array are connected and the current sunk in a bias supply.

In a vertical antiblooming structure, illustrated in Fig. 4c and 4d for a pinned photodiode sensing element, the image-sensing element is fabricated in a shallow, lightly doped *p*-well. A large (10- to 30-V) bias is applied to the *n*-type silicon substrate, causing the *p*-well underneath the photodiode to become completely depleted. Once the charge on the diode exceeds its capacity, the excess charge flows over the saddle-point in the *p*-well,

* Other types of antiblooming are occasionally used in image-sensing arrays. One of these is charge pumping, in which an MOS gate is repeatedly clocked in order to cause excess charge held underneath it to recombine at interface states at the silicon-silicon dioxide interface. In charge pumping, the MOS gate is pulsed sufficiently negative to cause accumulation, resulting in the interface states filling with holes. When the gate is pulsed out of accumulation, excess electrons can recombine with the holes.

under the diode, and into the substrate. The substrate is connected to a bias supply which sinks the blooming current. The vertical antiblooming structure has the advantage of requiring no additional silicon area, so that antiblooming is achieved with no reduction in fill factor. Lateral antiblooming, on the other hand, requires additional silicon area in each pixel, reducing fill factor. The vertical antiblooming has the disadvantage of lower quantum efficiency at wavelengths longer than about 500 nm. Because any light absorbed below the photodiode or photocapacitor is drained into the substrate, the quantum efficiency of photodiodes or photocapacitors with vertical antiblooming is reduced.

Dark Current in Photosensing Elements

Signal in photosensing elements is the result of collection of electron-hole pairs generated by the absorption of light. However, charge is generated at each photosensing element even in the absence of light. This generation is called dark current and is a result of thermal generation of electron-hole pairs. The thermal generation occurs at defects, such as impurities or crystalline defects, in the bulk of the silicon and at surface states at the silicon-silicon dioxide interface.

There are four sources of dark current: (1) diffusion current, which is the thermal generation of carriers in the undepleted n - and p -type regions; (2) depletion layer generation current, which occurs in the depletion layer of a diode or MOS capacitor; (3) surface generation current, which is the generation of electron-hole pairs at interface states at the Si-SiO₂ interface; and (4) leakage, which refers to generation at extended defects such as impurity clusters or stacking faults, particularly in the presence of a large electric field. These sources of dark current are illustrated for a photodiode and a photocapacitor in Fig. 5. The generation of the electron-hole pairs in both diffusion current and depletion-layer generation-recombination current occurs almost exclusively at impurity sites. Impurities with energy levels near midgap, such as gold, copper, and iron, are

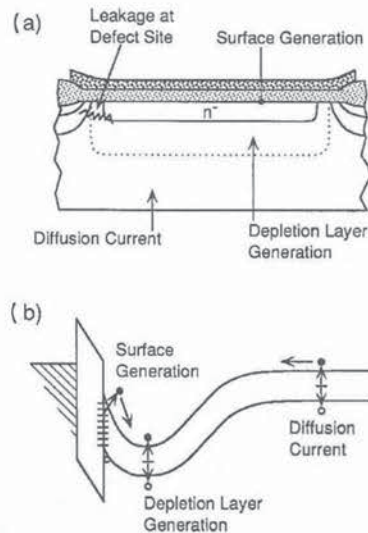


FIGURE 5 (a) Cross section of buried-channel MOS capacitor; (b) band diagram of buried-channel MOS capacitor illustrating mechanisms for dark current generation.

particularly effective in the thermal generation of charge. The depletion-layer generation current is given by:*

$$J_g(V) = \frac{qn_i W(V)}{\tau_o} \quad (9)$$

where $W(V)$ is the width of the depletion layer at a given bias voltage in the MOS capacitor or junction-photodiode, n_i is the intrinsic carrier concentration, and τ_o is the depletion-layer lifetime, for carriers. Typical values of τ_o in clean MOS processes would be 1 to 10 ms and values of J_{gr} would be 30 to 300 pA/cm².

The diffusion current generation current for a p -type region is given by:†

$$J_{diff} = \frac{qn_i^2 L_e}{N_a \tau_e} \quad (10)$$

Since the intrinsic carrier concentration n_i depends on temperature as $e^{-E_g/2kT}$, depletion-layer generation current and diffusion current will have different temperature dependences. Depletion-layer generation current will increase as $e^{-E_g/2kT}$, which corresponds to a doubling in dark current for every 9 to 11°C increase in temperature near room temperature. Diffusion current will increase as $e^{-E_g/kT}$ which corresponds to a doubling every 4.5 to 5.5°C increase in temperature.

The surface generation current J_s occurs almost exclusively at regions where the depletion layer intersects the Si-SiO₂ interface, such as the surface region between the n^- - and p -regions around a photodiode or in the depleted surface under a MOS capacitor. It is given by:

$$J_s = \frac{qn_i s_o}{2} \quad (11)$$

where s_o is the surface recombination velocity. A typical value of J_s for a MOS surface would be 100 pA/cm².‡ The surface generation depends on temperature in the same manner as the depletion-layer current. In very clean MOS processes (i.e., low concentration of metallic impurities in the silicon), the surface current is often the largest component of the overall dark current.

Leakage current occurs at extended defects in the silicon, such as impurity clusters and stacking faults, particularly when these defects are in a depletion layer and so are subject to a high electric field. While there is no single analytical expression for the current generated by an extended defect, they are characterized by very high values of dark current ($\gg 1$ nA/cm²), a very strong dependence on the electric field, and, in regions of

* The expressions here are for the generation current. The total current is the sum of the generation and recombination current and is given by

$$J_{gr}(V) = (qn_i W(V)/\tau_o)(e^{qV/2kT} - 1)$$

However, for diodes under 0.2 V or more of reverse bias ($qV/kT > 8$), such as would be the case in an image sensor, the recombination current is negligible.

† The full expression for the diffusion current from the undepleted p -region in a diode is:

$$J_{diff} = \frac{qn_i^2 L_e}{N_a \tau_e} (e^{qV/kT} - 1)$$

However, for reverse biases more than 0.2 V ($qV/kT > 8$), as would be the case in a photodiode in an image sensor, only the generation term is important.

‡ In calculating the total surface current, the current density J_s is multiplied by the area of depleted surface. For a pn junction diode, this would be the area of depleted surface region around the diode between the n^- - and p -regions (i.e., approximately the junction perimeter times the depletion-layer width); for a MOS capacitor it would be the entire area under the MOS capacitor unless sufficient sheet charge of electrons had formed at the surface to invert the surface. The surface recombination velocity is given by $s_o = N_{st} v_{th} \sqrt{\sigma_n \sigma_p}$ where N_{st} is the density of surface states near midgap, v_{th} is the thermal velocity (2.0×10^7 cm/s) and σ_n and σ_p are the electron and hole cross sections for midgap surface states. See Sze, *Physics of Semiconductor Devices*, for a complete explanation.

high electric field, only a very slight dependence on temperature. In an image sensor, they are visible as "bright spots" in a few isolated pixels against the otherwise low level of background thermal generation.¹⁰

Values for the dark current vary widely because of variation in the amount of impurities in the silicon; the dark current can range from 0.01 nA/cm² in very high quality image sensors to >10 nA/cm² in sensors with significant metallic contamination. The total number of electrons N_e collected in a pixel is:

$$N_e = JAT_{\text{int}}/q \quad (12)$$

where J is the total dark current per unit area, A is the area of the pixel, and T_{int} is the integration time. For a 1/2-in format CCD such as would be used in a camcorder, typical values would be a dark current of 0.5 nA/cm², a pixel area of 100 μm^2 , and an integration time of 1/30 s; the number of thermally generated electrons would be 105 electrons per pixel. Image sensors developed for scientific purposes might have a dark current 5 to 10 times lower at room temperature. These same devices might also be operated below room temperature to reduce the thermal generation to levels below one electron per pixel.

There are two types of noise associated with the charge generated by the dark current: shot noise and pattern noise. The shot noise due to the dark current is the square root of the number of dark electrons in a pixel. Pattern noise is due to pixel-to-pixel variations in the dark current and is often highly correlated between neighboring pixels. A numerical value for the dark pattern noise is typically obtained by using the standard deviation of the pixel values in the dark from a large region of an imager, where the values are obtained by averaging over many frames to eliminate shot noise and other temporal noise sources.*

22.4 READOUT ELEMENTS

The readout element transfers the charge from the image-sensing element (photodiode, photocapacitor, or photoconductor) to the output of the image sensor. In a linear sensor, the readout is only in one direction. In an area sensor, both x and y readout is required. There are two basic types of readout elements: charge-coupled devices, or CCDs, and x - y addressed photodiode arrays (typically called MOS arrays owing to the MOS transistors used in the addressing of the pixels). CCDs are by far the most widely used readout elements owing to their very low noise. Nearly all consumer camcorders, facsimile machines, scanners, copiers, and professional and scientific cameras utilize CCDs. Applications of MOS arrays are largely restricted to those where addressing of an individual pixel or subarray is required.

Charge-Coupled Device (CCD)

CCD Operation. The CCD works by moving packets of charge physically at or near surface of the silicon from the image-sensing element to an output, where the charge packet is converted into a voltage. The CCD is formed by an array of overlapping MOS capacitors.¹¹⁻¹³ There are a number of different types of CCD, including four-phase, three-phase, two-phase, virtual (single-) phase, and ripple-clocked CCDs. The number of phases refers to the number of separately clocked elements within a single stage of the

* A histogram of the dark current values of the pixels is rarely gaussian. In most cases, it exhibits an extended tail at high dark current values due to pixels with crystalline defects. The histogram may also exhibit quantization due to pixels with integral numbers of impurities (i.e., 1, 2, or 3 gold atoms); the quantization may even be used as a signature of the impurity present. See, for example, McColgin et al., "Effects of Deliberate Metal Contamination on CCD Image Sensors," *Materials Research Society Symposium Proceedings*, **262**: 769 (1992) and "Dark Current Quantization in CCD Image Sensors," *Proc. 1992 International Electron Device Meeting*, Washington, D.C., 113-116 (1992).

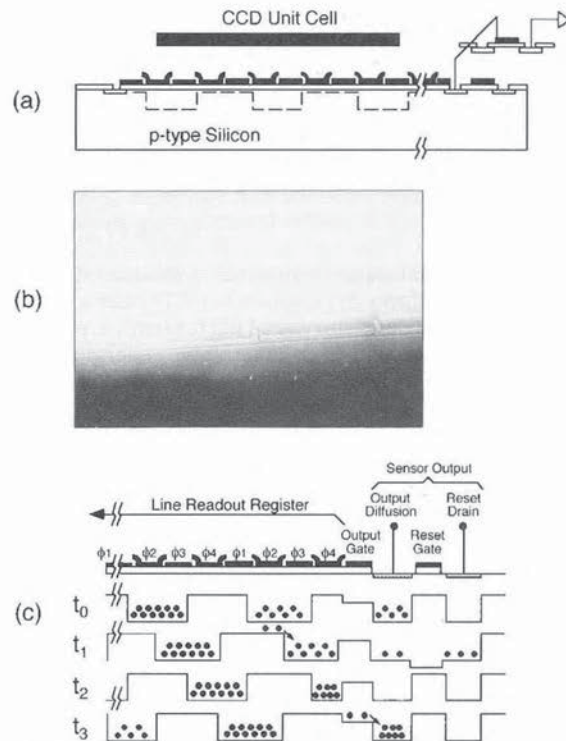


FIGURE 6 (a) Cross-sectional diagram of unit cell of a four-phase CCD; (b) scanning electron micrograph of a unit cell of a four-phase CCD; (c) illustration of charge transfer along a four-phase CCD, showing both transfer along the register and at the sensor output.

CCD. These will be described later. For understanding the principle of charge transfer in a CCD, consider the four-phase CCD illustrated in Fig. 6a.

The four-phase CCD physically comprises a silicon substrate (assumed to be *p*-type for purposes of illustration), a gate oxide 300 to 1000 Å thick, and overlapping polysilicon electrodes 1000 to 5000 Å thick which have been heavily doped with phosphorus to lower their resistance. For a four-phase CCD, two levels of electrode are required. The first level is deposited, then defined photolithographically to form two phases (ϕ_1 and ϕ_3). A thin (500-Å) oxide is grown over the first level of polysilicon to insulate it from the second level. The second level of polysilicon is then deposited, doped with phosphorus, and defined to form the other two phases (ϕ_2 and ϕ_4). An electron micrograph of a four-phase CCD with six-micron long gates is shown in Fig. 6b.

The process of charge transfer in a four-phase CCD is illustrated in Fig. 6c. In order to hold a packet of electrons, two adjacent gates (ϕ_2 and ϕ_3 , for example) would be held at a high positive potential ($\sim +5$ V) while the other two phases would be held at a low potential (~ 0 V). A depletion layer, or well, is formed under ϕ_2 and ϕ_3 , allowing electrons to be held at or below the surface. The other two phases, ϕ_1 and ϕ_4 , serve as potential barriers, keeping the charge packet under ϕ_2 and ϕ_3 . To transfer the electrons through the silicon, the electrode ahead of the charge packet (ϕ_4) is clocked positive and the electrode behind (ϕ_2) is clocked negative. The electrons move along the silicon surface following

the positive potential. This is repeated through all four phases, during which the charge packet is moved forward one pixel.

The CCD may be either surface-channel or buried-channel. In a surface-channel CCD (Figs. 2*b* and 6*a*) the electrons are held at the silicon surface. Surface-channel CCDs are rarely used owing to the trapping of electrons at interface states at the silicon surface. At the silicon-SiO₂ interface there is a density of states of about $1 \times 10^{10} \text{ cm}^{-2} \text{ eV}^{-1}$. These states can trap electrons from one charge packet and reemit the electrons into a later charge packet. This results in transfer efficiency. In the buried-channel CCD (Figs. 2*c* and 6*d*), a lightly doped *n*-type layer is formed in the silicon. This *n*-type layer results in a potential well for electrons just below the surface rather than at the surface (Fig. 2*b*). As a result, the electrons remain separated from the interface and are not trapped by interface states. This results in the ability to transfer charge from one stage to another with very high efficiency. Virtually all CCDs for image-sensing applications utilize buried-channel CCDs. The clock voltages used to drive CCD gates typically swing by 5 to 8 V between high and low levels.

For buried-channel CCDs, transfer rates of up to 20 megapixels per second can usually be achieved without special design considerations. Above this rate, special attention must be paid to optimizing the electric field along the direction of transfer to speed up charge transfer. Transfer rates exceeding 50 MPix/s have been achieved in optimized CCD designs.¹⁴

CCD Output. At the output of the CCD is a circuit to convert the charge packets into a voltage signal. By far the most common type of output circuit is the floating diffusion with source follower amplifier. The floating diffusion output is shown schematically in Fig. 7*a* and a photograph of the end of a CCD shift register with the first stage of the amplifier is shown in Fig. 7*b*. The floating diffusion output consists of an output gate (OG), a floating diffusion, a reset gate (RG), and a reset drain. The floating diffusion is an *n*⁺-type region in between the output gate and the reset gate. The floating diffusion is connected to the gate of a source-follower amplifier. The output gate is held at a fixed dc voltage, creating a barrier potential over which the packet of electrons can be transferred onto the floating diffusion when the last phase of the CCD register is clocked to its low ($\sim 0 \text{ V}$) voltage.

The sequence of events in the CCD output is shown in Fig. 6*c*. As the charge packet is transferred along the CCD, it arrives at the last phase (ϕ_4) before the output gate (time t_2 in Fig. 6*c*). When ϕ_4 is clocked low (time t_3), the packet of electrons is transferred over the output gate onto the floating diffusion (time t_0). The voltage of the floating diffusion changes by an amount

$$V = Nq/C \quad (13)$$

where N is the number of electrons and C is the total capacitance of the floating diffusion itself, the interconnect to the source-follower amplifier and the input capacitance of the amplifier. In typical CCDs, this capacitance would be in the 10-fF to 50-fF range. Because this capacitance is so small, there is a large change in voltage for a small change in charge. The charge-to-voltage conversion is an important parameter in CCD design; for a 10-fF capacitance the charge-to-voltage conversion is 16 microvolts per electron. After the change in voltage has been sensed by the amplifier, the charge packet must be removed from the floating diffusion before the next packet arrives. This is achieved by clocking the reset gate positive (time t_1 in Fig. 6*c*), allowing the charge packet to flow from the floating diffusion to the reset drain. The reset drain is held at a constant positive voltage, typically $\sim 10 \text{ V}$. The reset drain is then turned off and the floating diffusion is prepared to accept the next packet of electrons.

The change in voltage from the floating diffusion is typically buffered on-chip by a source-follower (Fig. 7*c*). A two-stage source follower is most often used. The first stage utilizes very small-dimensioned FET transistors in order to minimize the input capacitance. The second uses much larger FET transistors in order to achieve sufficient drive current to

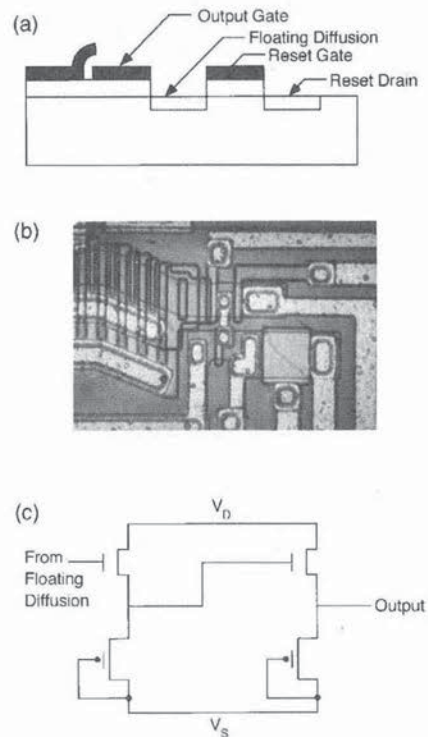


FIGURE 7 (a) Cross section of the floating diffusion output for a CCD; (b) photomicrograph of a floating diffusion output including the first stage of the source-follower amplifier; (c) two-stage source-follower amplifier with on-chip loads.

overcome off-chip capacitances such as package and lead capacitances on the circuit board. The source-follower amplifier is typically designed to have a bandwidth on the order of ten times the CCD pixel rate. For high-pixel-rate applications (>10 MPix/s), three-stage source-follower amplifiers are employed.

For special purpose applications, such as CCD signal processing, nondestructive readout is required. Floating gate output amplifiers are used in these applications. These amplifiers are similar to the floating diffusion except that the floating diffusion is replaced by a MOS gate which is connected to the MOS amplifier. Other outputs, such as buried-channel JFET structures, have seen limited use in very low noise applications.¹⁵

Types of CCDs. In addition to the four-phase CCD described here, there are a number of other types of CCD shift registers, including three-phase, two-phase, and virtual phase. The different types are illustrated in Fig. 8. The four-phase CCD (Fig. 8a) was described previously. The three-phase (Fig. 8b) consists of three different layers of polysilicon electrodes. The charge is normally held under one of the three; during transfer, the gate ahead of the charge packet is clocked positive and the gate holding the charge is clocked negative in order to transfer the charge one gate ahead. The three-phase CCD has the advantage of a shorter unit cell than the four-phase but at the expense of additional processing complexity (i.e., a third polysilicon layer).

In the two-phase CCD, each phase has a barrier and a well region. The barrier is

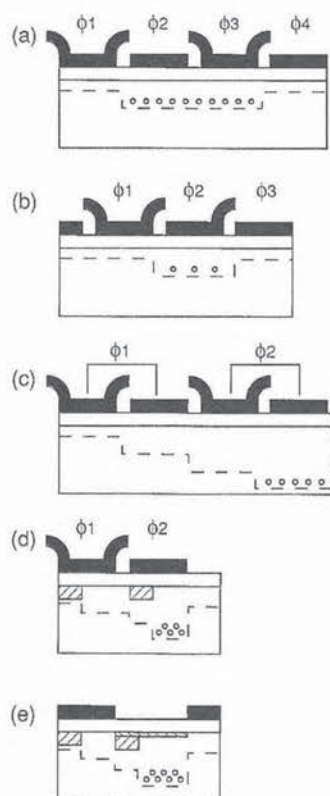


FIGURE 8 Types of CCD registers: (a) four-phase; (b) three-phase; (c) pseudo-two-phase; (d) true two-phase; (e) virtual phase. In the pseudo-two-phase CCD, each phase consists of two polysilicon gates, one of which is offset in potential from the other by an implanted threshold adjustment. In the true two-phase CCD, each phase consists of a single polysilicon gate in which the implanted threshold adjust region is formed underneath a portion of the gate.

formed by doping the barrier region slightly less *n*-type than the well region, making the electrostatic potential in the barrier region a few volts lower than the well region for the same gate voltage. Electrons will flow over the barrier region and be held in the well region. There are two methods of fabricating a two-phase CCD. In the first method (Fig. 8c), two separate gates are used for each phase; one gate receives a threshold adjustment implantation in order to create the barrier region. The two gates, however, are connected and thus driven at the same voltage. In the other method (Fig. 8d), the barrier and well regions are created under a single polysilicon gate. The two-phase CCD is very commonly used for three reasons. First, it requires only two clocks (ϕ_1 and ϕ_2), simplifying drive circuitry. Second, the clocks are complementary. This reduces clock feedthrough. Third, the two-phase has a higher horizontal density, especially the two-phase with barrier and well regions under the same gate.

The virtual phase CCD^{16,17} (Fig. 8e) consists of one gate with both barrier and well

regions within it and a second region, the virtual phase, in which a shallow, high-dose ion implantation is used to create a heavily doped surface region which pins the surface potential at 0 V. The virtual phase has both barrier and well regions within it and acts the same as a polysilicon gate held at a constant voltage. Charge is first transferred from the clocked phase into the virtual phase, then the charge is transferred from the virtual phase into the next clocked phase. The virtual phase CCD has a number of advantages, including higher quantum efficiency than two-polysilicon-level area CCD sensors and simpler clocking. The disadvantages of the virtual phase include the need for larger voltage swings on the single-clock and high-clock feedthrough into the output due to the lack of complementary clocks.

CCD Characteristics. The four major performance parameters of a CCD shift register and output amplifier are charge-handling capacity, charge-transfer efficiency, charge-to-voltage conversion ratio, and noise. The charge capacity is the number of electrons which can be held and transferred in the CCD shift register. As charge is added to a shift register, a point is reached at which excess charge cannot be held; the excess charge either overflows into adjacent pixels or overflows into the bulk beneath the CCD or, in the case of a buried-channel CCD, overflows the barrier to the Si-SiO₂ surface. The charge capacity is a function of the device design, device layout, and CCD process. Figure 9a shows the electrostatic potential and charge distribution in a buried-channel CCD at three levels of charge: approximately one-quarter of saturation, at saturation, and beyond saturation. Below saturation, the electrons fill the center of the buried channel in the region of largest potential, separated from the surface by approximately 0.2 μm in distance and about 500 mV in potential. As additional charge is added, the electron distribution spreads towards the surface and the potential barrier to the surface drops. Beyond saturation, the electrons contact the surface directly, resulting in charge-transfer inefficiency and blooming to neighboring pixels. The charge capacity of most CCD shift registers is of the order 1×10^{12} electrons/cm² of area in which the charge is held. In most CCD cells, the charge is held in only a fraction of the total cell area both along the CCD register and across the register. Typically, area CCD image sensors are designed with CCD charge capacities in the range of 50,000 to 200,000 electrons. Linear CCD image sensors, because of the larger amount of silicon area available for the CCD shift register, often are designed for 100,000 to 1,000,000 electrons.

The second major performance parameter for CCD shift registers is charge transfer efficiency. The transfer of charge from one stage to the next is neither instantaneous nor complete, limiting both transfer rate and the total number of stages in the CCD. There are two intrinsic mechanisms governing charge transfer: drift and diffusion. Drift is the movement of charge in the presence of an electric field. There are two origins of the electric field seen by an electron during charge transfer: the self-induced field resulting from the other electrons under the gate and the externally induced, or fringing, field. During the early stages of charge transfer, the self-induced field is large and is the dominant factor. After the charge concentration under the gate has decreased to a low value, the remainder of the charge transfer will be governed either by diffusion or by drift due to externally induced, or fringing, fields.

For a surface-channel CCD, the self-induced fields can be estimated from the formula:

$$E = -\frac{q}{C} \frac{dN}{dx} \quad (14)$$

where C is the gate capacitance per unit area and $N(x)$ is the density of electrons as a function of distance x from the edge of the electrode. In the early stages of charge transfer, both N and dN/dx are large. As the transfer proceeds, both N and dN/dx become small and the charge transfer is governed by fringing fields and diffusion.

Fringing fields are due to the two-dimensional nature of the electrostatic potential. If

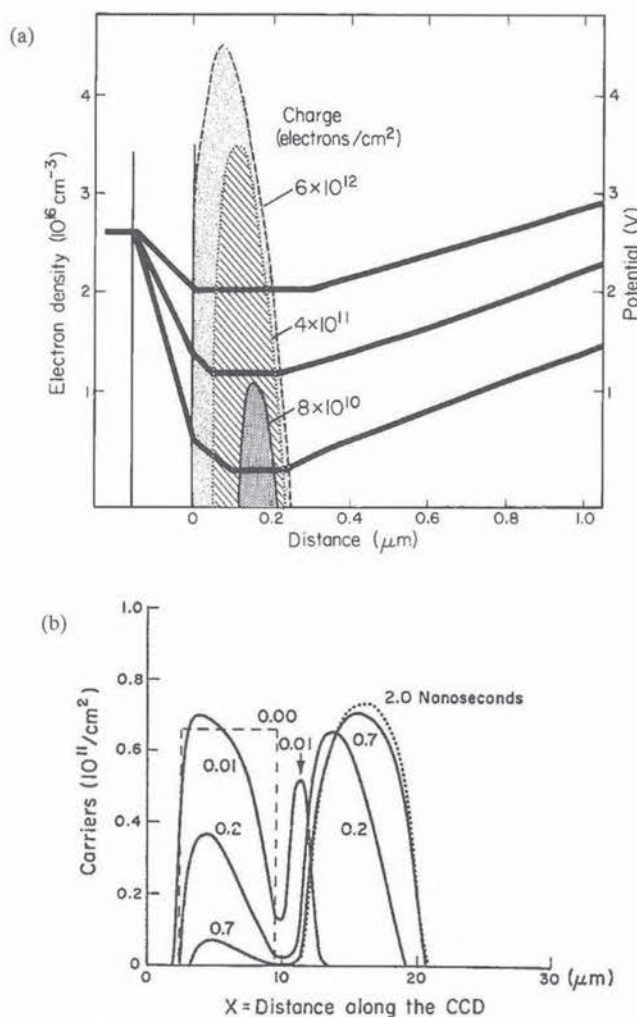


FIGURE 9 (a) Charge distribution and electrostatic potential as a function of distance from the surface for a buried-channel CCD for three levels of charge density; (b) charge density as a function of distance along a CCD for various times following the beginning of charge transfer from one 8- μm stage to another.

the charge is being transferred from gate 1 to gate 2, the potential will change smoothly between the two gates. The effect of the potential from one gate will typically extend 1 to 3 μm into a neighboring gate. The charge within the range of the fringing field will move by drift to the neighboring gate. Charge out of the range of the fringing fields will move by diffusion.

Figure 9b shows an example of charge transfer calculated for charge transfer from one 8- μm -long CCD gate into an adjacent gate. The charge density is shown as a function of distance along the CCD at several times after the start of charge transfer. At the start of the transfer, all the charge is under the first gate. At 0.01 ns into the transfer, the charge has moved from the edge of the first gate into the second gate, as a result of self-induced

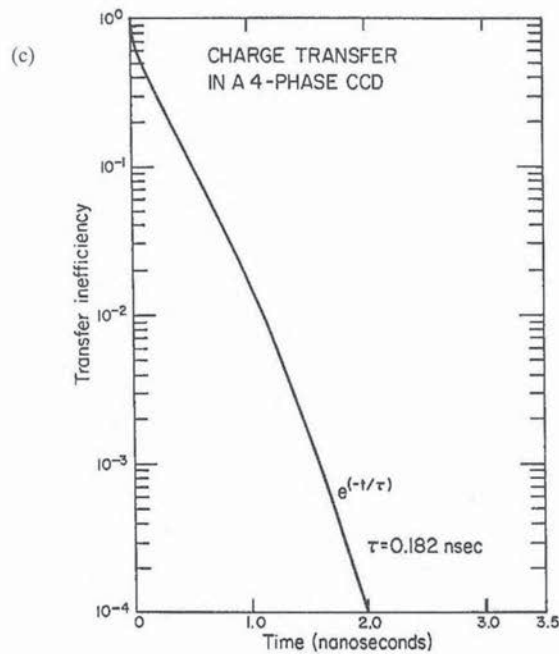


FIGURE 9 (Continued) (c) charge transfer inefficiency as a function of time from the start of charge transfer for the same example.

drift. By 0.2 ns, approximately half the charge has been transferred. By 0.7 ns, over 90 percent of the charge has moved into the second gate, leaving a residual in the first. At this point, the self-induced drift is sufficiently small that drift due to fringing fields and diffusion are the dominant mechanisms. Figure 9c shows the charge transfer inefficiency as a function of time for this example. Two slopes are evident in the transfer inefficiency. In the first 0.5 ns, the charge is transferred rapidly owing to the self-induced drift. For times longer than 0.7 ns, the transfer is due to fringing fields in this example.

For CCDs with longer gates or lower fringing fields than the example above, the final ~10% of the charge must transfer by diffusion. Charge transfer by diffusion follows an exponential time dependence.

$$N(t) = N(0)e^{-t/\tau_{\text{diff}}} \quad (15)$$

where the time constant τ_{diff} for diffusion is

$$\tau_{\text{diff}} = \frac{4L_g^2}{\pi^2 D} \quad (16)$$

where L_g is one gate length and $D = KT\mu_e/q$.

For electrons at room temperature, $D = 25.8 \text{ cm}^2/\text{s}$. For an 8- μm long gate, the diffusion time constant is ~10 ns. To achieve a transfer inefficiency below 2×10^{-5} per transfer, 11 time-constants, or 110 ns in this example, are required. For this reason CCDs are typically designed with gate lengths shorter than ~8 μm and are also designed to build in fringing fields.

In the simplest case for very low levels of transfer inefficiency, the total transfer inefficiency in a CCD register is the product of the number of stages N in the register and

the transfer inefficiency per stage. Each stage will require two or more transfers. Virtual phase and two-phase require two transfers per stage, three-phase requires three transfers, and four-phase requires four transfers. The inefficiency per stage, however, will likely depend in a complicated manner on the amount of charge in the charge packet, the amount of charge in preceding charge packets,* the voltages, and frequency of operation.

In addition to the intrinsic sources of transfer inefficiency, there are a variety of extrinsic sources. These include surface and bulk traps and potential wells and barriers. The traps and the potential obstacles hold back an amount of charge from a charge packet. The charge is reemitted to later charge packets. The inefficiency due to traps depends on whether the traps have been filled by preceding charge packets and the emission time constants of the traps, and so is not modeled in a simple manner.

The intrinsic sources of noise in CCD shift registers include dark current and output amplifier noise. In addition, other sources of noise not intrinsic to the CCD itself include noise due to clock feedthrough from the CCD clocks to the output signal and noise in external electronics. The generation of dark current in CCD shift registers is the same as described at the end of Sec. 22.3 for photosensing elements. Associated with the dark current are both shot noise and pattern noise. The magnitude of the pattern noise in a CCD shift register is reduced over that of a single element since the charge packet averages the dark current over many pixels as it is transferred to the output.

The noise associated with the CCD output consists of the Johnson or thermal noise, the $1/f$ noise of the output amplifier, and the kTC noise associated with resetting the floating diffusion. The kTC noise is due to uncertainty in the amount of charge remaining on the floating diffusion following reset owing to thermal fluctuations in the reset gate. The rms noise σ_n in the number of electrons caused by kTC noise is given by

$$\sigma_n = \sqrt{\frac{kT}{q} C_{FD}} \quad (17)$$

where C_{FD} is one total floating diffusion capacitance. For a 10-fF floating diffusion capacitance, the rms noise is ~ 40 electrons.

The kTC noise may be eliminated entirely by use of a signal-processing technique called correlated double sampling (CDS). In correlated double sampling, the output level is sampled before the charge packet is transferred onto the floating diffusion and again after transfer of the charge packet (times t_2 and t_0 respectively in Fig. 6c). The two values are subtracted either by an analog circuit or by digital subtraction. Any uncertainty in the voltage level of the floating diffusion following reset is subtracted and thus the kTC noise eliminated. The output amplifier low-frequency noise, called $1/f$ noise because of the inverse frequency ($1/f$) shape of the noise power spectrum, is also reduced but not eliminated by correlated double sampling. The total noise of a CCD output amplifier is in the range of 7 to 40 rms electrons per pixel depending on the amplifier design and the operating speed. Values of a single rms electron or less have been obtained for very slow pixel rates under cooled conditions.¹⁸

MOS Readout

The other major category of readout structures is the MOS readout. The individual light-sensing elements (photodiodes, phototransistors, or photoconductors) at each pixel are connected to a readout line by means of a transfer gate. Each pixel along the readout line is addressed separately by addressing circuitry. When a particular pixel is addressed, the transfer gate is turned on and the charge transferred from the pixel to the readout

*The charge in the preceding packets will affect the filling of both bulk and interface traps. See, for example, Sequin and Thompsett, pp. 70-108.

line. An amplifier at the end of the readout line senses the change in voltage or current resulting from the charge transfer.

Typically, the pixels would be addressed serially along the line. The first pixel would be addressed, causing the charge from the image-sensing element to be transferred onto the readout line. The voltage change or current would be sensed, the readout line reset to its original voltage if necessary, and the next pixel addressed. This is different from a CCD. In the CCD, charge from all pixels is transferred into the CCD register simultaneously. Individual pixels or groups of pixels cannot easily be addressed in a random fashion by the CCD, but this random addressing can be accomplished readily by the MOS readout.

There are several types of MOS readout devices. These include the CID,¹⁹ the AMI,²⁰ and the CMD²¹ in addition to the normal MOS array. The CID has no readout line. Each pixel consists of two overlapping gates, one controlled by a row address and the other by column address.¹¹ When neither row nor column of a particular pixel is being addressed, the photogenerated charge is held under both gates and can be transferred between them. When a row of a pixel is addressed, the charge transfers onto the column gate. Then both row and column are addressed, the charge is injected into the substrate, and the current is sensed. The CID is not widely used in visible imaging applications because the charge conversion sensitivity is very poor and noise very high compared to the CCD or to the other MOS architectures.

In the amplified MOS imager (AMI), the image-sensing element at each pixel consists of a phototransistor rather than a simple photodiode. The photogenerated charge is stored on the gate of the MOS transistor.¹⁹ When a particular pixel is addressed, the photogenerated charge modulates the transistor current. This current amplification at each pixel helps to overcome many of the noise and speed limitations of conventional MOS arrays.

MOS readout differs in an important way from CCD readout. In MOS readout, the charge is transferred from a single pixel onto a readout line and the change in voltage or current in the readout line is sensed. In a CCD, the charge packets are kept intact while being transferred physically to a low-capacitance output. The lower sensitivity of a simple MOS array can be illustrated as follows. The change in voltage on the readout line is given by $V = Nq/C$ where N is the number of electrons, q the electron charge, and C the readout line capacitance. Because the readout line covers the full length of the array, its capacitance is in the picofarad range (typically 2 to 10 pF depending on design and process). This compares to the 10-femtofarad capacitance for the CCD output. As a result, the voltage swings on the readout line are very small (16 nV/electron for a 10-pF readout line capacitance). This leads to a high sensitivity to clock noise due to capacitive feedthrough of the row and column address clocks onto the readout line. The feedthrough may be many times larger than the signal in most MOS sensors. Once the charge has been transferred onto the readout line, it is sensed either by a current-sensitive amplifier or by a voltage-sensitive amplifier, followed by a reset of the readout line to its original voltage.

CCD readout has the advantages of very high sensitivity and low noise. However, CCD readouts are limited in charge-handling capacity, while MOS readouts are capable of carrying very large amounts of charge and so are not as limited on the high end of the dynamic range. However, because the MOS readout line has much higher capacitance than the CCD, the sensitivity is lower and the noise is higher. Another difference is in the readout architecture. The CCD readout is essentially a serial readout device; it is not suited to random readout or partial-array readout. The MOS array, however, can be addressed in a manner similar to a memory, making it well-suited to pixel or partial-array addressing.

22.5 SENSOR ARCHITECTURES

Solid-state image sensors are classified into two basic groupings: linear and area. Linear sensors include single-line arrays, multilinear arrays for color scanning, and time delay and

integrate (TDI) arrays for low-light-level scanning. Area sensor architectures include the frame transfer CCD, the interline transfer CCD, and various forms of MOS x - y addressed arrays.

Linear Image Sensor Arrays

Linear sensors are used almost exclusively in scanning systems for scanning of documents, film, and three-dimensional still objects. There are two basic classes of scanning systems: contact scanners and reduction scanners. These are illustrated in Figs. 10a and 10b. In reduction scanners (Fig. 10a), the sensor is smaller than the document to be scanned; lenses are used to image the document onto the sensor. In contact scanners (Fig. 10b), the sensor is the same width as the item to be scanned, usually a document. Relay optics is used between the sensor and the document. Selfoc lenses (Fig. 10c and 10d) and roof-mirror-lens arrays are the two types of relay optics used most frequently.

There are three basic architectures for linear sensing arrays: MOS line arrays, CCD linear and multilinear sensors, and time-delay and integrate (TDI) sensors. These architectures are illustrated in Fig. 11. The MOS array is used most often in contact scanning applications where material or processing problems make CCD arrays impractical. These applications include arrays fabricated from polysilicon or amorphous silicon on nonsilicon substrates, arrays covering large distances, or arrays requiring special processing (such as logarithmic amplification) at each pixel. The CCD linear and multilinear arrays are used most often in reduction scanning where wide dynamic range and small pixel size is required. However, contact arrays are also often realized by butting multiple CCD arrays end-to-end. TDI arrays are used in very low light level scanning applications where integration over many lines is required to achieve adequate signal-to-noise.

The MOS linear array (Fig. 11a) consists of individual photosensing elements, an

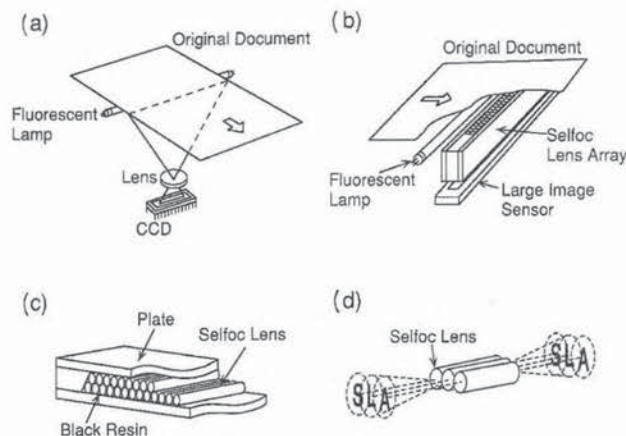


FIGURE 10 (a) Reduction document scanner using linear image sensor, in which the image on the page is reduced by a lens onto the line array; (b) contact document scanner in which the length of the image sensor is the same as the document width and one-to-one relay optics is used to transfer the image from the document to the array; (c) diagram of a selfoc lens array used as transfer optics between document and array in a contact scanner; (d) operation of a selfoc lens array.

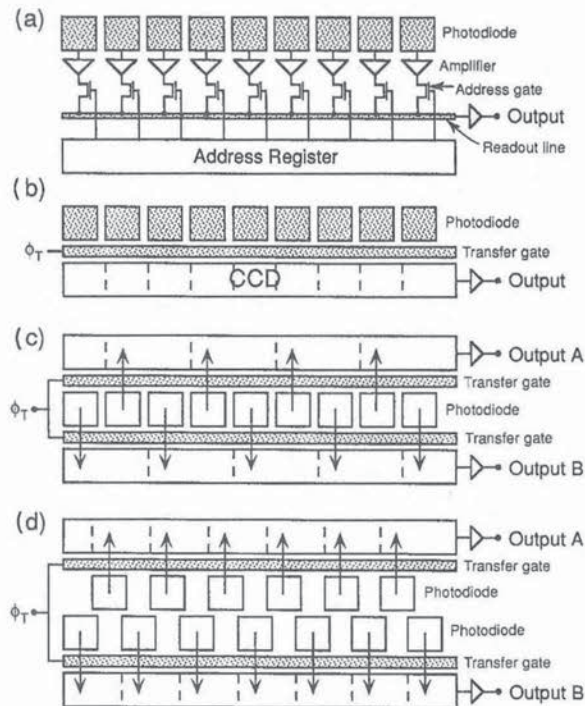


FIGURE 11 Architectures for linear image sensors: (a) MOS line array consisting of photodiodes, preamplifier, MOS switches addressed by an address register, and a readout line with amplifier; (b) linear CCD image sensor consisting of photodiodes, transfer gate, and CCD readout; (c) linear CCD image sensor with two CCD output registers, one for the odd diodes and the other for the even diodes, for higher horizontal pitch; (d) staggered linear CCD image sensor with two rows of photodiodes offset by one-half pixel to increase horizontal sampling.

amplifier, an address switch, and a readout line and amplifier. The photosensing element is usually a photodiode, although photoconductors are used in contact scanning arrays fabricated from amorphous silicon. Since the charge generated in the diode is generally very small (in the hundreds to thousands of electrons), a simple amplifier is usually placed at each pixel to drive the high-capacitance readout line. The use of amplification at each pixel can allow some signal-processing functions, such as logarithmic amplification, clipping, triggering and latching, etc., to be performed at the pixel. A MOS switch placed after the amplifier allows each pixel to be addressed in sequence; the switch is driven by an address register. At the end of the readout line is an amplifier which may buffer and/or amplify the signal. The MOS array has the advantage of process simplicity and the ability to perform signal processing at each pixel; it has the disadvantage of low signal level (because of the large readout line capacitance) and pattern noise introduced by feedthrough from the switching transistor.

The linear CCD image sensor is the most often used architecture for scanning applications owing to its low noise, high sensitivity, wide dynamic range, and small pixel pitch. Figure 11b shows the simplest type of linear CCD, in which a single row of photodiodes is connected to a single CCD register via a transfer gate. In operation, the

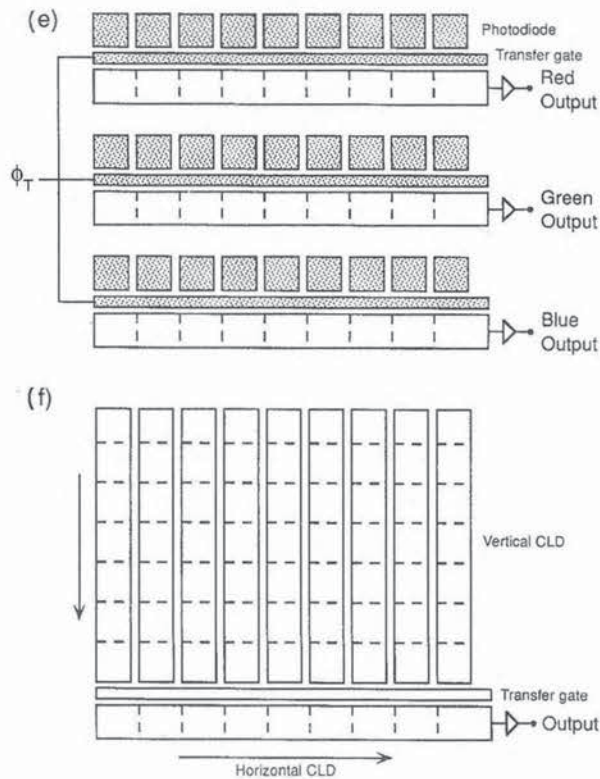


FIGURE 11 (Continued) (e) trilinear CCD in which three CCDs are fabricated on the same silicon die, each with its own color filter; (f) time-delay and integrate (TDI) array, in which the charge in the vertical registers is clocked in phase with the motion of the scene or document being imaged in order to increase signal-to-noise.

signal is integrated on the image-sensing element (generally a photodiode) for a line time. The horizontal CCD is then stopped, the transfer gate opened, and the charge transferred from all the photodiodes simultaneously to the CCD. The transfer time is typically a few microseconds. The transfer gate is then closed, integration resumed for the next line, and the CCD clocked to read out the charge packets. Many arrays also feature antiblooming for situations where the light level may not be controlled, as well as electronic shuttering, which allows an integration time on the photodiodes to be less than the readout time of the CCD.

For linear-sensing applications requiring a higher pixel density, a double-sided readout is often used (Fig. 11c). In this architecture, a CCD array is placed on either side of the line of photodiodes and charge transfer from the diodes alternates between the top and bottom CCDs. This architecture uses lower horizontal clock rates and a higher pixel pitch, since the diode pitch can usually be made smaller than the CCD pitch. The charge packets from the two arrays may be multiplexed into one output if desired. The disadvantage of this architecture is slight differences between even and odd pixels, due to slight differences in the two outputs (or slight differences due to multiplexing the two registers).

Another architecture which is used to further decrease the sampling pitch is the

staggered linear array (Fig. 11d). In the staggered array, two rows of photodiodes are offset by a half pixel. The two rows are read out by CCD arrays. The first array is delayed (usually in a digital line store) and then combined with the second to form a double-density scan.

Multilinear arrays (Fig. 11d) have recently been developed for color scanning applications. In this architecture, several (usually three) linear arrays are combined on the same silicon die separated by a distance equivalent to an integral number of scan lines. Color filters (either integral or in close proximity) are aligned over the arrays. External digital line delays are used to realign the three arrays. Separate electronic shuttering may be provided for each array in order to adjust for differences in intensity in each of the bands.

The third major class of line arrays is time-delay and integrate, or TDI, arrays.^{22,23} The TDI architecture is shown in Fig. 11f. TDI arrays are used when inadequate signal-to-noise ratio from a single-line array requires averaging over multiple lines. Applications for TDI arrays include high-speed document scanners and space-based imaging systems. Instead of a single row of photodiodes, the TDI array utilizes CCD stages in the vertical dimension which are clocked synchronously with the movement of the document to be scanned. The signal level in a TDI array increases linearly with the number of stages. The noise level, however, increases most as the square root of the number of stages.

Area Image Sensor Arrays

There are three major classes of area image sensor architectures: MOS diode arrays, frame-transfer CCDs, and interline-transfer CCDs. Within each there are a number of variations. CCDs have come to dominate the majority of applications owing to their higher sensitivity. However, MOS arrays are still used for specialized applications where addressability or high readout rate is important.

Historically, the physical dimensions of the active imaging areas of CCD arrays for consumer and commercial applications are specified by the size of the vidicon tube which it replaces. The common format sizes include $\frac{1}{4}$ -in, $\frac{1}{3}$ -in, $\frac{1}{2}$ -in, and 1-in. In most cases, the aspect ratio is 4:3, reflecting the television standard. Table 1 lists the formats and the corresponding dimensions of the imaging area of the array.

TABLE 1 Image Area Dimensions and Pixel Dimensions for Various Format Image Sensors

The format name is given in inches based on historic image tube formats. The pixel dimensions are based on a 484×768 pixel array.

Optical format	1-in	$\frac{2}{3}$ -in	$\frac{1}{2}$ -in	$\frac{1}{3}$ -in	$\frac{1}{4}$ -in
Active area (4:3 aspect ratio)					
Height (mm)	9.6	6.6	4.8	3.3	2.4
Width (mm)	12.8	8.8	6.4	4.4	3.2
Diagonal (mm)	16	11	8	5.5	4
Pixel dimensions ($484 \text{ lines} \times 768 \text{ pixels}$)					
Height (μm)	19.8	13.6	9.9	6.8	4.9
Width (μm)	16.7	11.5	8.3	5.7	4.2

The standards for the number of vertical lines in arrays for consumer and professional video are usually based on the corresponding television standards, including NTSC, PAL, and the various HDTV standards. NTSC has 484 active lines, PAL has 575, the Japanese HDTV standard has 1035, and the European HDTV standard has 1150. In all cases the lines are interlaced; i.e., odd lines are read out in one field and even lines in the next. Historically, for sensors for NTSC television, the number of pixels horizontally in a line has been associated with multiples of the color subcarrier frequency; common horizontal pixel counts include 384, 576, and 768. Sensors for the Japanese HDTV standard typically have 1920 horizontal pixels. The pixels are rectangular rather than square. Table 1 lists approximate pixel dimensions in micrometers for a few common formats for a $484(V) \times 768(H)$ pixel image sensor.

For industrial, scientific, graphics electronic photography, digital television, and multimedia applications, however, the nonsquare pixels and interlaced readout of sensors based on television standards are significant disadvantages. Image sensors designed for these industrial and commercial applications typically have square pixels and progressive scan readout. In interlaced NTSC readout, for example, the first field consists of lines 1, 3, 5 $\dots$ 483 and is read out in the first 1/60 second of a frame. The second field consists of lines 2, 4, 6 $\dots$ 484 and is read out in the second 1/60 second of a frame. The resulting temporal and spatial displacement between the two fields is undesirable for these applications. In addition, digital compression of interlaced scan moving images, especially with motion estimation, is difficult and also introduces artifacts.

In progressive scan readout each line is read out sequentially. There is no even or odd field, only a single frame. As a result there are no temporal and spatial sampling displacement differences. However, for a given resolution and frame rate, the readout rate of a progressive scan image sensor is double that of an interlaced scan. In addition, for these applications, the number of pixels is often based on powers of 2: such as 512×512 or 1024×1024 —facilitating memory mapping and image processing.

MOS Area Array Image Sensors. The architecture of MOS area arrays is illustrated in Fig. 12.²⁴ It consists of the imaging array, vertical and horizontal address registers, and output amplifiers. The pixel of a MOS array consists of an image-sensing element (photodiode, photocapacitor, phototransistor, or photoconductor), a row-address gate, and a vertical readout line. The row-address gate is bussed horizontally across the array and is driven from a row-address register on the side(s) of the array. At the start of a line, a single row is addressed, causing the charge from all the photodiodes in a row to be transferred onto the vertical readout line. Horizontal address gates are placed at the bottom of the vertical readout line. Buffer amplifiers to drive the horizontal readout line may also be placed at the end of the vertical readout line. The horizontal address register then serially addresses each vertical readout line, sequentially turning on the horizontal address gates. After the horizontal addressing is completed, the readout lines may be reset and precharged and the next row addressed.

There is a wide range of variations on this basic architecture. An example is the charge modulation device, or CMD,^{25,26} array in which a phototransistor is placed at each pixel site in order to achieve amplification and to achieve high currents to drive the capacitance of the vertical and horizontal readout lines. Other examples include arrays with sophisticated charge collection circuits at the end of each vertical readout line.

Frame Transfer CCD Image Sensors. CCD area arrays fall into two categories: frame transfer and interline transfer. The simplest form of frame transfer CCD is the full-frame CCD, shown in Fig. 13a. A photograph of a few pixels of a frame transfer CCD is shown in Fig. 14a. The array consists of a single image area composed of vertical CCDs and a single horizontal register with an output amplifier at its end. In this architecture, the pixel consists of a single stage of a vertical CCD. This type of device requires an external shutter. When the shutter is opened, the entire surface of the sensor is exposed and the

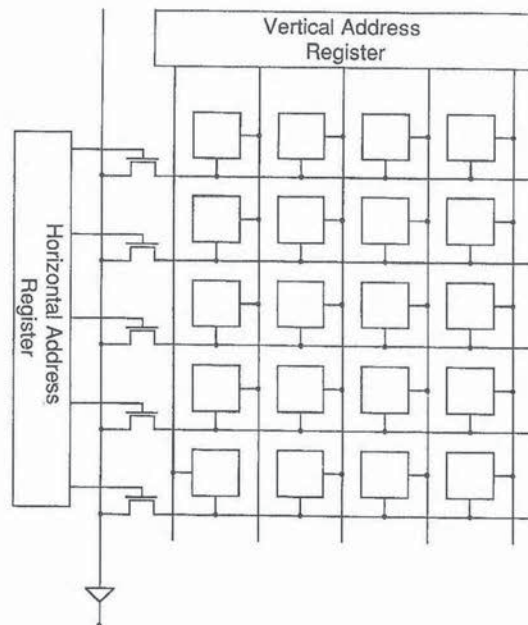


FIGURE 12 MOS photodiode array, consisting of vertical and horizontal address registers and readout line.

charge is collected in the CCD potential wells at each pixel. After the shutter is closed, the sensor is read out a row at a time by clocking a row of the vertical register into the horizontal register, then clocking the horizontal register to read out the row through the output amplifier. For higher readout rates dual horizontal CCDs are used in parallel. The full-frame CCD has the advantage of progressive scan readout high fill factor, very low noise, and wide dynamic range. However, it requires an external shutter. It is most often used in still electronic photography, scientific, industrial, and graphics applications.

For motion imaging applications, a shutter is not practical. In order to overcome the need for a shutter, frame transfer CCDs incorporate a storage area in addition to imaging area. For interlaced video applications, this storage area is sufficiently large to hold a field (242 lines in NTSC television). The image area consists of vertical CCDs. For interlaced NTSC video, there are 242 pixels vertically in the image area. Interlace is achieved by changing the gates under which integration is performed in the even and odd fields. For example, for a four-phase CCD, integration would be performed under phases 1 and 2 in one field and 3 and 4 in another, effectively shifting the sampling area by half a pixel in each field. The storage area consists also of 242 pixels vertically. The device operation is as follows. The image area integrates for a field time and the photogenerated charge held in the vertical CCDs. The vertical CCDs are then clocked in order to rapidly transfer the charge from the image area into the storage area. Because the sensor is still under illumination, this transfer time must be much shorter than the integration time. This transfer typically requires 0.2 to 0.5 ms. The storage area is then read out a row at a time by transferring a row into the horizontal register and clocking the horizontal register. While this readout is occurring, the image area is integrating the next field. The most significant disadvantage of frame transfer CCDs is the image smear caused by illumination during the transfer from the image to the storage area. This smear can be on the order of 3 percent.

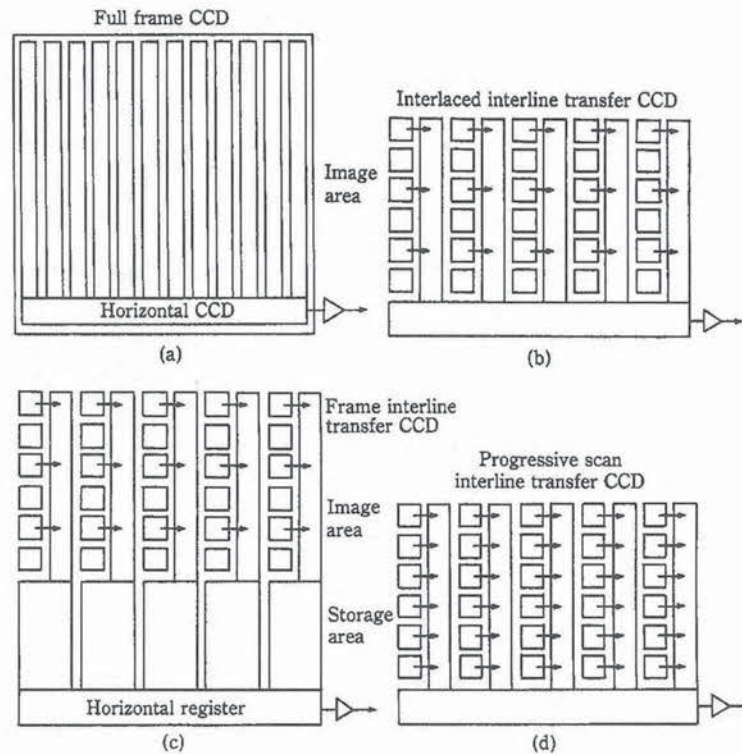
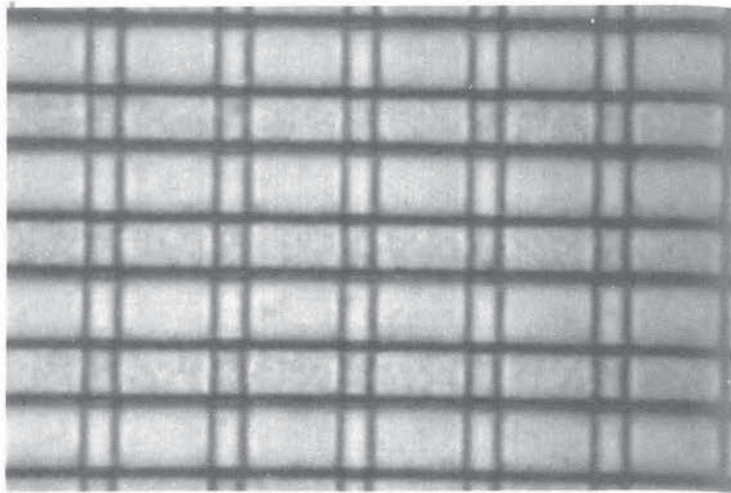


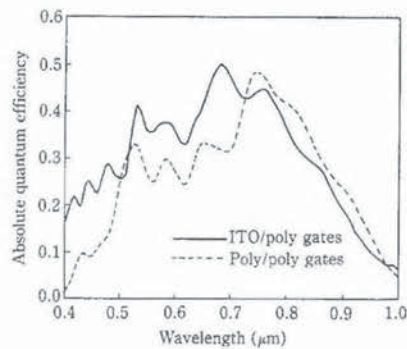
FIGURE 13 (a) Architecture of full-frame CCD; (b) architecture of interlaced interline transfer CCD; (c) architecture of frame interline transfer CCD, in which a storage area is provided to reduce smear during readout; (d) progressive scan interline transfer CCD, in which every photodiode is read out into the vertical CCD simultaneously.

In both frame transfer and full-frame devices, the light must pass through the polysilicon electrodes before being absorbed in the silicon. Owing to the high absorption of short wavelengths in the polysilicon, the quantum efficiency in the blue is only about 20 percent and the quantum efficiency in the green is about 50 percent. Figure 14b shows the quantum efficiency for a full-frame image sensor with polysilicon gates. Three approaches have been used to improve the efficiency: the virtual phase CCD, transparent electrodes, and backside illumination. In the virtual phase CCD (see "Types of CCDs" in Sec. 22.4), the second polysilicon electrode is replaced with a very shallow highly doped p^+ layer, very similar to the pinned photodiode (Fig. 2d). Since there is no electrode over this phase to absorb the light, higher quantum efficiency is achieved, particularly at wavelengths less than 500 nm. Backside illumination provides nearly unity quantum efficiency but requires that the sensor be thinned to less than 10 μm . Owing to its cost, backside thinning is employed only in sensors for very specialized scientific or aerospace applications. Indium-tin-oxide, or ITO, is the most commonly used transparent electrode.²⁷ Usually it is substituted for the second level of polysilicon. Figure 14b also shows the quantum efficiency of a full-frame device in which ITO has been substituted for one of the polysilicon gates.

Interline Transfer CCD Image Sensors. The interline transfer CCD is fundamentally



(a)



(b)

FIGURE 14 (a) Photograph of a few pixels in the image area of a full-frame CCD; (b) quantum efficiency as a function of wavelength for a full-frame CCD with polysilicon and with transparent indium-tin-oxide electrodes.

different from the frame-transfer CCD in that, in addition to the vertical CCD, the pixel also contains a separate image-sensing element (photodiode, pinned photodiode, photocapacitor, or photoconductor) and a transfer region between the photodiode and the vertical CCD. Interline CCDs with photodiodes for sensing elements are considered first. Figure 13b illustrates the architecture of an interline transfer CCD and Fig. 15 illustrates a pixel of the CCD. The CCD and the transfer region between the diode and CCD are covered with a light shield (typically aluminum, although metal silicides are also used). The light shield prevents any light from entering the vertical CCD registers, allowing them to be read out while the sensor is illuminated. When the sensor is illuminated, the photogenerated charge is held on the photodiode. During the vertical retrace interval at the end of a field, the photogenerated charge is transferred into the vertical CCD by clocking the CCD gate over the transfer region. Once the charge has been transferred from the diodes into the vertical CCDs, the diodes resume integrating and the vertical

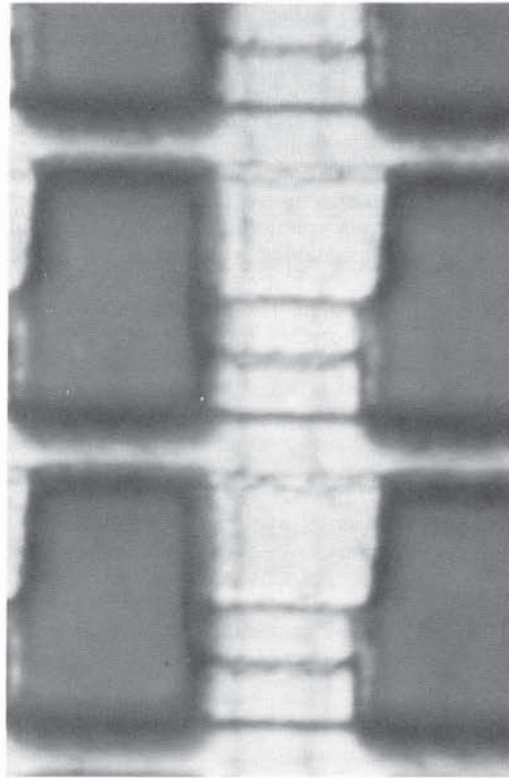
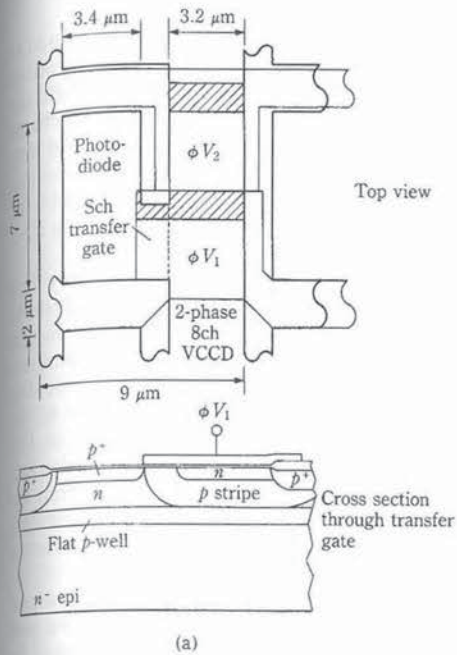


FIGURE 15 (a) Diagram of a pixel of an interline transfer CCD; (b) photograph of a pixel from an interline transfer CCD.

CCDs are clocked in order to transfer a row at a time from the image area into the horizontal CCD.

For consumer and most commercial applications, interlaced interline CCDs are used to be consistent with television standards. In interlaced interline CCDs there is one vertical CCD stage for every two photodiodes. During the retrace time before the first field the charge from the odd rows of photodiodes is transferred into the vertical CCDs; the charge is then transferred out a row at a time into the horizontal CCD. During the retrace time before the second field the charge from the even rows of photodiodes is transferred into the vertical CCDs; once again, the charge is transferred out a row at a time into the horizontal CCD. In NTSC television the fields are approximately 1/60 second long; in PAL the field time is 1/50 second.

Because the vertical CCDs in an interline CCD are covered by a light shield, very little stray light is absorbed in the CCD. However, due to light scattering under the lightshield and lateral diffusion of photogenerated electrons, some charge can reach the vertical registers as they are read out during a field. This results in smear. For consumer applications the level of smear is not noticeable. However, for especially demanding applications such as television broadcast cameras, a field storage area is added to the bottom of the image area. This architecture is called the frame interline transfer, or FIT CCD²⁸ (Fig. 13c). Following transfer of the photogenerated charge from the diodes into the vertical CCDs, the vertical CCDs are clocked to rapidly shift the charge from the image area into the storage area. This transfer typically takes less than 0.5 ms, reducing smear 30-fold. One line at a time is transferred from the storage area to the horizontal register and the horizontal register readout.

For still electronic photography, scientific, computer-related, graphics and professional applications, interlaced video is not desirable. For these applications a progressive scan architecture is utilized. In a progressive scan interline CCD there is a full CCD stage for every photodiode.²⁸ Following integration, the photogenerated charge from all the photodiodes is transferred into the vertical CCDs. The photodiodes resume integration and the charge packets from the vertical registers are transferred into the horizontal a row at a time. The progressive scan interline CCD requires twice the vertical CCD density and is therefore more complicated to fabricate. However, progressive scan readout provides many advantages in image quality for both motion and still imaging and is emerging as a likely standard for future digital HDTV systems as well as multimedia.

Nearly all interline CCDs used in camcorder, broadcast camera, or commercial applications utilize vertical antiblooming in order to prevent blooming when the sensor is illuminated beyond saturation. A cross-section diagram of an interline CCD with vertical antiblooming is shown in Fig. 15b. The device illustrated uses a pinned photodiode photosensing element. The CCD is built on an *n*-type silicon substrate. A *p*-well is formed about 2 μm deep in the *n*-type silicon. An *n*-type buried-channel is then formed followed by a *p*+ surface layer. The *n*-type substrate is reverse biased with respect to the *p*-well. When the photodiode is illuminated above saturation, the excess electrons spill out of the *n*-type buried channel and into the substrate. Owing to the vertical overflow, however, photogenerated carriers from photons absorbed below the *p*-well are drawn into the substrate and are not collected by the diode. Thus the quantum efficiency of these devices falls rapidly at wavelengths beyond 550 nm. Figure 16 shows the quantum efficiency of an interline CCD with vertical antiblooming as a function of wavelength. The internal quantum efficiency of the photodiode is nearly 100 percent to about 550 nm, after which it decreases. However, since the photodiode only occupies about 20 percent of the pixel, and this aperture is typically reduced further by the light shield in order to eliminate optical scattering into the vertical CCD, the actual efficiency of the device is only about 15 percent. The photoresponse is linear at low signal levels but becomes nonlinear at charge levels near saturation.²⁹

Two major approaches are used to improve the fill factor (and therefore the quantum efficiency) of interline CCDs. These are the use of microlens arrays to focus the light incident on a pixel onto the photodiode and vertical integration of image sensors with

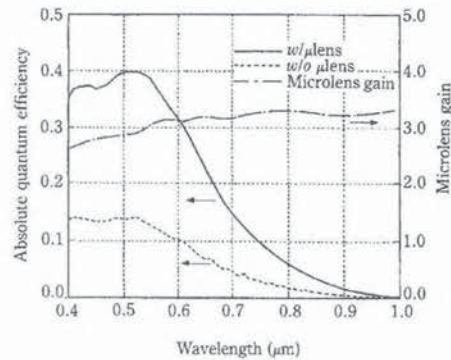


FIGURE 16 Quantum efficiency as a function of wavelength for an interline transfer CCD with and without microlens array.

amorphous silicon photoconductors to achieve a high fill factor. Figure 17 shows a microlens array on top of an interline CCD.³⁰ The lenses are formed by coating a spacer layer on the wafer of CCD devices followed by a lens-forming layer. The lens forming layer is patterned and then reflowed to form the lens arrays. The quantum efficiency of an interline CCD with and without a microlens array is shown in Fig. 18. A 3-fold improvement in quantum efficiency is achieved because light from nearly the entire pixel area is focused onto the photodiode.

The second approach to improving fill factor in interline CCD devices is to vertically integrate the device by use of an amorphous silicon photoconductor.⁷ The structure of the photoconductor and its band diagram are illustrated in Fig. 2e. The amorphous silicon is

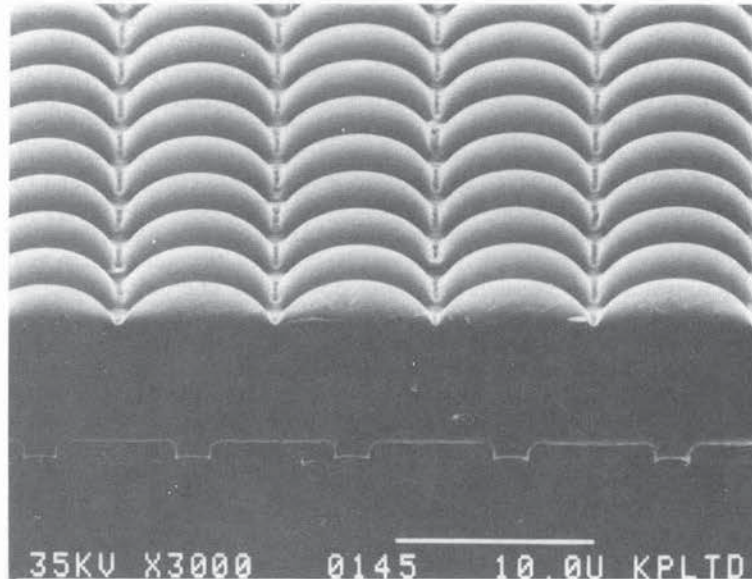


FIGURE 17 Scanning electron micrograph of microlens array on top of an interline transfer CCD.

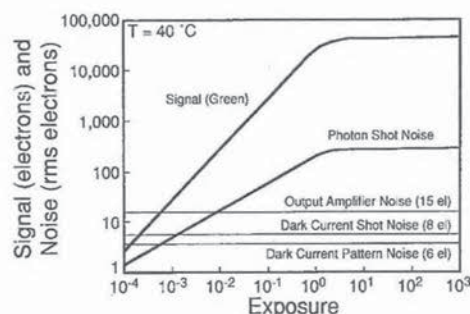


FIGURE 18 Quantum efficiency of interline CCD with and without microlens array.

about $1\text{ }\mu\text{m}$ thick; owing to its high absorption coefficient in the visible, it can achieve nearly 100 percent internal quantum efficiency. The back contact of the photoconductor contacts the diode in the interline CCD. Light absorbed in the amorphous silicon generates electron-hole pairs. The amorphous silicon is biased such as to create a high field which sweeps out the electrons to the back contact, where they can be stored on the diode until they are transferred into the CCD. Because the photoconductor is fabricated on top of the CCD pixel, it can have nearly 100 percent fill factor. In addition, because of the wide bandgap of the amorphous silicon, its dark current (due to thermal generation of carriers) is often lower than the single-crystal silicon. However, the amorphous silicon photoconductors suffer difficulties related to charge trapping. Owing to impurities and dangling or strained bonds, there is a high density of traps in amorphous silicon. These can trap charge carriers and reemit them at a later time, causing image lag. Because the trapping and detrapping is field-dependent, it can also result in nonlinear response.

CCD Performance. In all CCDs, both interline and frame transfer, the signal output is linear with the illumination except at levels approaching saturation. This linear response is in marked contrast to image tubes, which exhibit highly nonlinear response. For interline CCDs, the total charge capacity ranges from 100,000 electrons for larger cells (such as the $13.6\text{ (V)} \times 11.6\text{ (H)}\text{-}\mu\text{m}$ cell typical for a $\frac{2}{3}$ -in format) to 20,000 electrons or less for smaller cells (such as the $6.8 \times 5.8\text{-}\mu\text{m}$ cell in a $\frac{1}{3}$ -in format 484×768 -pixel CCD).

The principal noise sources in both interline and full-frame image sensors are dark current pattern and shot noise and output amplifier noise. To illustrate the contributions, modern CCDs have dark current levels at 40°C of less than 1 nA/cm^2 . For a $\frac{2}{3}$ -in format sensor, this would correspond to about 320 electrons per pixel dark level. The corresponding shot noise would be 18 rms electrons and the pattern noise would typically be at a similar level. However, because of the nonrandom nature of pattern noise, its appearance is considerably more noticeable. The output amplifier noise would also be at about the 15-rms electron level. Thus, the overall noise for this example would be in the 30-electron range and the dynamic range would be over 2000 for a charge capacity of 90,000 electrons. For scientific applications where the sensor can be cooled and the readout performed at a lower frequency, noise levels less than 5 electrons can be achieved and even subelectron noise has been reported.¹⁸

Color Imaging

Silicon based CCDs are monochrome in nature. That is, they have no natural ability to determine the varying amounts of red, green, and blue (RGB) illumination presented to the photodetectors. There are three techniques to extract color information.

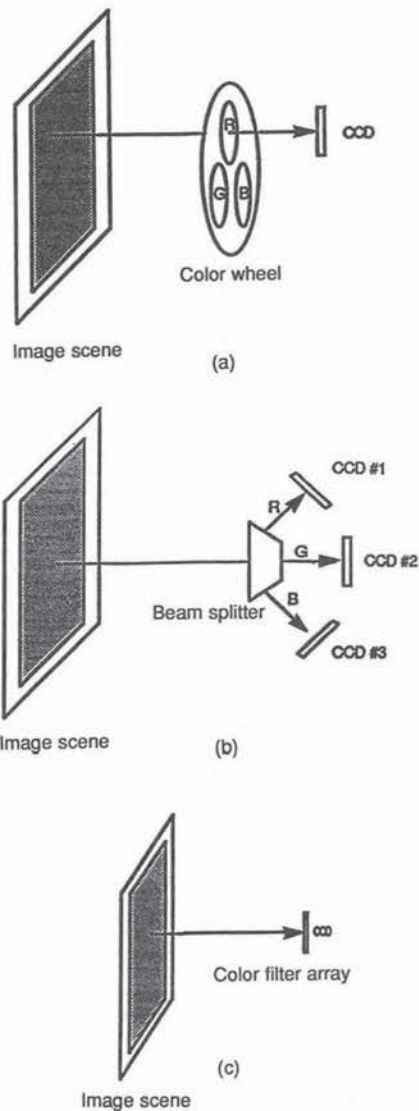


FIGURE 19 Methods of color separation in cameras with area image sensors: (a) color sequential using a color wheel; (b) a prism with dichroic beam splitters and three image sensors; (c) single-chip image sensor with integral color filter array.

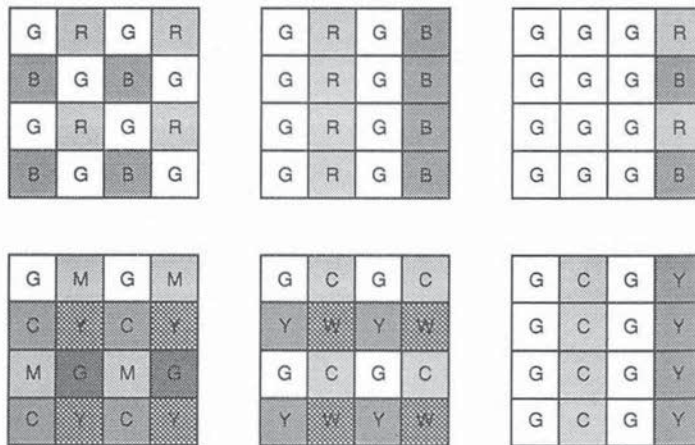
1. Color Sequential (Fig. 19)—A color image can be created using a CCD by taking three successive exposures while switching in optical filters having the desired RGB transmittances. This approach is normally used only to provide still images of stationary scenes. The resulting image is then reconstructed off-chip. The advantage to this technique is that resolution of each color can remain that of the CCD itself. The disadvantage is that three exposures are required, reducing frame times by more than a factor of three. Color

misregistration can also occur due to subject or camera motion. The filter switching assembly also adds to the mechanical complexity of the system.

2. Three-chip Color (Fig. 19)—Three-chip color systems use an optical system to split the scene into three separate color images. A dichroic prism beam splitter is normally used to provide RGB images. Color images can then be detected by synchronizing the outputs of the three CCDs. The disadvantage to such a system is that the optical complexity is very high and registration between sensors is difficult.

3. Integral color Filter Arrays (CFA) (Fig. 19)—Instead of performing the color filtering off-chip, filters of the appropriate characteristics can be fabricated above individual photosites.^{31,32} This approach can be performed during device fabrication using dyed (e.g., cyan, magenta, yellow) photoresists in various patterns. The major problem with this approach is that each pixel is sensitive to only one color. Off-chip processing is required to "fill in" the missing color information between pixels.^{33,34}

In order to minimize size, weight, and cost, most consumer color camcorders use a CCD sensor with an integral CFA. The photosites are covered with individual color filters—for example, a red, green, and blue striped filter, or a green, magenta, cyan, and yellow mosaic filter. Some popular CFA patterns are shown in Fig. 20. Because each photosite can sense only one color, the color sampling is not coincident. For example, a blue pixel might be seeing a white line, while nearby red and green pixels are seeing a dark line in the scene. As a result, high-frequency luminance edges can be aliased into bright color bands. These color bands depend not only on the color filter pattern used, but also on the optical prefilter and CFA interpolation algorithm. The color bands are caused by aliasing, which is a property of any sampled system. Aliasing occurs when the frequency of the input signal is greater than the Nyquist limit of one-half the sampling frequency. If the input frequency is well below the Nyquist limit, there are many samples per cycle. This allows the input to be reconstructed perfectly, if a proper reconstruction filter is used. When the input frequency is greater than the Nyquist limit, there are less than two samples per cycle. The sample values now define a new curve, which has a frequency lower than the input frequency. In effect, the high frequency takes on the alias of a lower frequency.



Note: Color R = Red, G = Green, B = Blue
Y = Yellow, M = Magenta, C = Cyan, W = White

FIGURE 20 Common color filter array patterns.

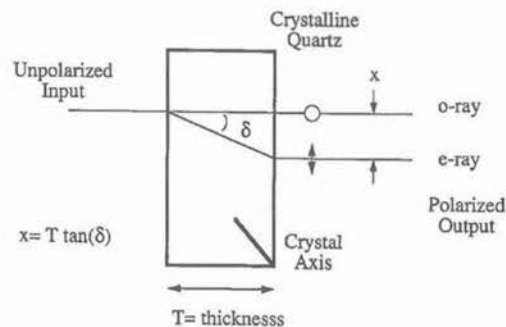


FIGURE 21 Birefringent blur filter used to reduce aliasing in single-chip color image sensors.

Aliasing is a particular problem with color sensors, since the sampling phase is different for the different color photosites. Therefore, the aliased signal has different phases for different colors. This creates the color bands.

The color aliasing is reduced by using an optical anti-aliasing or "blur" filter, positioned in front of the color CCD sensor.³⁵ Blur filters are typically made of birefringent quartz, with the crystal axis oriented at a 45° angle, as shown in Fig. 21. In this orientation, the birefringent quartz exhibits the double refraction effect. An unpolarized input ray emerges as two polarized output rays. The output ray separation is proportional to the filter's thickness. A 1.5-mm-thick plate will give a separation of about 9 μm. Figure 21 shows a simple "two-spot" filter. More complex filters use three or more pieces of quartz cemented in a stack.

22.6 REFERENCES

1. G. Lubberts and B. C. Burkey, "Optical and Electrical Properties of Heavily Phosphorous-doped Epitaxial Silicon Layers," *J. Appl. Phys.* **55**(3):760-763 (Feb. 1, 1984).
2. O. S. Heavens, *Optical Properties of Thin Solid Films*, Dover, New York, 1965.
3. N. Teranishi et al., "An Interline CCD Image Sensor with Reduced Image Lag," *IEEE Transactions on Electron Devices*, **ED-31**(12):1829-1833 (Dec. 1984).
4. S. Kosman et al., "A Large Area 1.3 Megapixel Full-Frame CCD Image Sensor with a Lateral Overflow Drain and a Transparent Gate Electrode," *Proc. IEEE International Electron Device Meeting*, Washington, D.C. (Dec. 1990).
5. B. Burkey et al., "The Pinned Photodiode for an Interline Transfer CCD Image Sensor," *Proc. 1984 International Electron Device Meeting*, Washington, D.C. (Dec. 1986), p. 28.
6. M. Sasaki, H. Ihara, and Y. Matsunaga, "A $\frac{2}{3}$ -in 400 k-Pixel Sticking-Free Stact CCD Image Sensor," *IEEE Transactions on Electron Devices*, **ED-28**(11) (Nov. 1993).
7. H. Yamashita et al., "A $\frac{2}{3}$ -inch 2-megapixel stack CCD Imager," *Proceedings of the 1994 International Solid State Circuits Conference*, p. 224.
8. E. G. Stevens et al., "A 1-Megapixel, Progressive-Scan Image Sensor with Antiblooming Control and Lag-Free Operation," *IEEE Transactions on Electron Devices*, **ED-38**(5) (May 1991).
9. D. K. Schroder, "The Concept of Generation and Recombination Lifetimes in Semiconductors," *IEEE Transactions on Electron Devices* **ED-29**(8) (August 1982), p. 17.
10. J. Van der Spiegel and G. J. Declerck, *Solid State Electronics* **27**:147 (1984).
11. Carlo, H. Sequin and Michael F. Tompsett, *Charge Transfer Devices Advances in Electronics and Electron Physics*, suppl. 8, Academic Press, Inc., New York, 1975.

12. M. J. Howes and D. V. Morgan, (eds.), *Charge Coupled Devices and Systems*, Wiley, Chicago, 1979.
13. D. F. Barbe et al., *Charge-Coupled Devices*, Springer-Verlag, New York, 1980.
14. M. Azuma et al., "A Fixed Pattern Noise Free $\frac{2}{3}$ ", 1.3 M-Pixel CCD Image Sensor for HDTV Camera System," *Proc. 1991 IEEE International Solid-State Circuits Conference*, San Francisco, 212-213 (Feb. 1991).
15. Y. Matsunaga and S. Ohsawa, " $\frac{1}{3}$ Inch Interline Transfer CCD Image Sensor with Negative Feedback 94 dB Dynamic Range Charge Detector," *Proc. 1991 IEEE International Solid-State Circuits Conference*, San Francisco, 210-211 (Feb. 1991).
16. J. Janesick, "Open Pinned-Phase CCD Technology," *SPIE Proceedings*, vol. 1159, San Diego (Nov. 1989).
17. J. Hynecek, "Virtual Phase Technology: A New Approach to Fabrication of Large-Area CCDs," *IEEE Transactions on Electron Devices* **ED-28**(5):483-489 (May 1981).
18. J. Janesick et al., "Fano-Noise-Limited CCDs," *Optical and Optoelectric Applied Science and Engineering Symposium: X-Ray Instrumentation in Astronomy*, San Diego (Aug. 14-19, 1988).
19. J. Carbone et al., "New Low Noise Random Access, Radiation Resistant and Large Format Charge Injection Device (CID) Imagers," *Proc. of SPIE Conference 1900* (Jan. 31-Feb. 4, 1993).
20. M. Sugawara et al., "An Amplified MOS Imager Suited for Image Processing," *Proc. 1994 IEEE International Solid-State Circuits Conference*, San Francisco, 228-229 (Feb. 1994).
21. M. Ogata et al., "A Small Pixel CMD Image Sensor," *IEEE Transactions on Electron Devices*, **ED-38**(5):1005-1010 (May 1991).
22. D. H. McCann, M. H. White, A. P. Turley, and R. A. Frosch, "Time Delay and Integration Detectors Using Charge Transfer Devices," U.S. Patent Number 4,280,141, July 1981.
23. Thompson et al., "Time-Delay-and-Integration Charge-Coupled Devices Using Tin Oxide Gate Technology," *IEEE Transactions on Electron Devices* **ED-25**(2):132-134 (Feb. 1978).
24. Ohba et al., "MOS Area Sensor: Part 2: Low-noise MOS Area Sensor with Antiblooming Photodiodes," *IEEE Trans. Electron Devices* **ED-27**(8):1682-1687 (Aug. 1980).
25. T. Nakamura et al., "A New MOS Image Sensor Operating in a Non-Destructive Readout Mode," *Proc. 1986 IEEE International Electron Device Meeting*, Washington, D.C. (Dec. 1986), 353-356.
26. Nomoto et al., "A $\frac{2}{3}$ Inch 2M-Pixel CMD Image Sensor with Multi-Scanning Functions," *Proc. 1993 IEEE International Solid-State Circuits Conference*, San Francisco (Feb. 1993), 196-197.
27. McCann et al., "Buried Channel CCD Imaging Arrays with Tin Oxide Transparent Gates," *IEEE International Solid State Circuits Conference*, pp. 30, 31, 261, 262 (Feb. 1978).
28. K. Harada, "A $\frac{2}{3}$ -Inch, 2M-Pixel Frame Interline Transfer CCD HDTV Image Sensor," *Proc. 1992 IEEE International Solid-State Circuits Conference*, San Francisco, 170-171 (Feb. 1992).
29. E. Stevens, "Photoresponse Nonlinearity of Solid State Image Sensors with Antiblooming Protection," *IEEE Transactions on Electron Devices*, **ED-38**(2):229-302 (Feb. 1991).
30. A. Weiss et al., *J. Electrochem. Soc.* **133**:110C (1986).
31. P. Dillion et al., "Fabrication and Performance of Color Filter Arrays for Solid-State Imagers," *IEEE Trans. Electron Devices* **ED-25**:97-101 (1978).
32. P. Dillion, D. Lewis, and F. Kaspar, "Color Imaging System Using a Single CCD Array," *IEEE Trans. Electron Devices*, **ED-25**:102-107 (Feb. 1978).
33. K. A. Parulski, et al., "A Digital Color CCD Imaging System Using Custom VLSI Circuits," *IEEE Trans. Consumer Electronics* **35**(3):382-389 (Aug. 1989).
34. K. A. Parulski, "Color Filters and Processing Alternatives for One-Chip Cameras," *IEEE Trans. Electron Devices* **ED-32**(8):1381-1389.
35. J. E. Greivenkamp, "Color Dependent Optical Prefilter for the Suppression of Aliasing Artifacts," *Appl. Opt.* **29**(5):676-684, (Feb. 1990).